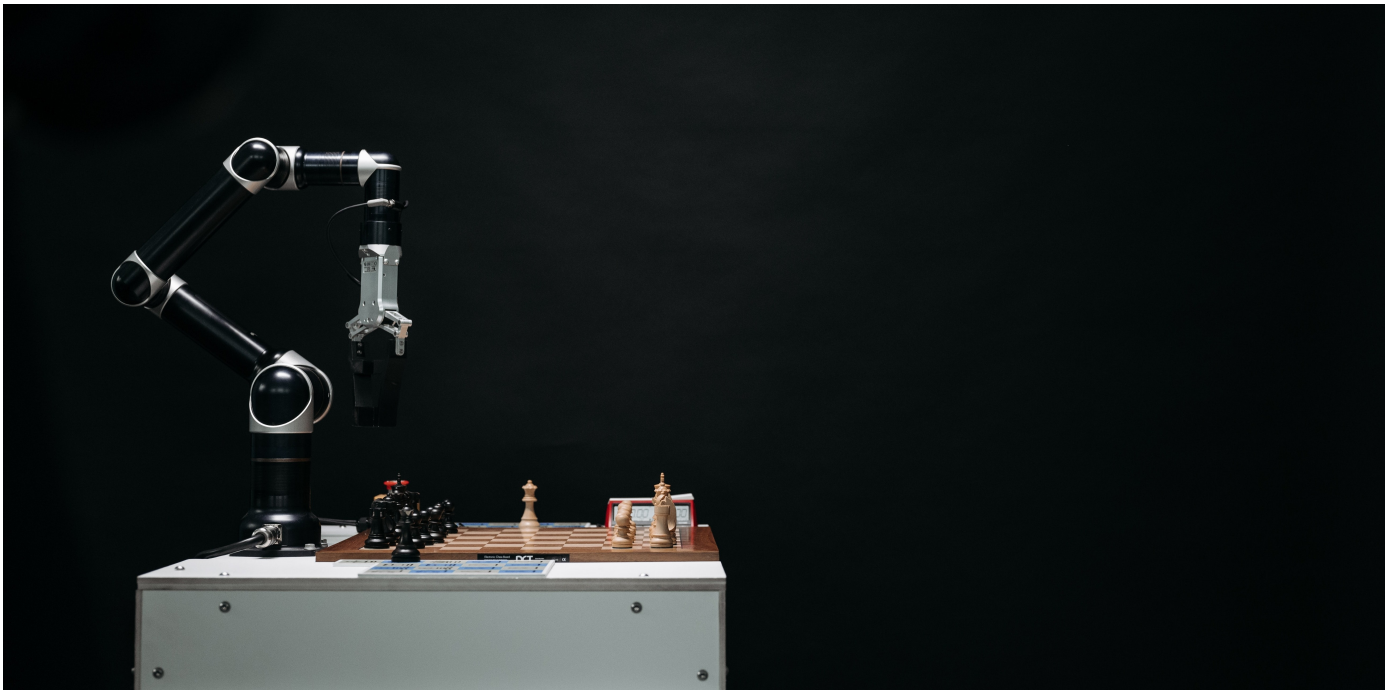
[Dossier IA générative] - De l'entraînement à la pratique : l'IA générative et ses usages

Rédigé par Martin Biéri - 26 avril 2023

Comme nous l'avons vu dans les articles précédents, les IA génératives soulèvent des questions et des enjeux sur les modèles sous-jacents et leurs modalités d'entraînement. Mais qu'en est-il « en sortie » ? Quels sont les enjeux liés à l'usage d'un système d'IA tel que ChatGPT ? Les risques semblent à la hauteur des bénéfices annoncés par les promoteurs de l'outil.



Sommaire :

Des précédents parlants et des enjeux économiques pour les grandes plateformes

Il existe un certain nombre de précédents avant l'arrivée de ChatGPT qui n'ont pas forcément reçu un accueil aussi glorieux. A commencer par l'annonce et le déploiement peut-être trop précipités d'outils, à l'instar du chatbot Tay, lancé par Microsoft en 2016. Il avait été retiré de la plateforme quelques heures après sa mise en ligne, **suite à des propos problématiques**, notamment homophobes ou antisémites (qui avaient été eux-mêmes générés par des internautes cherchant – au mieux – à tester les limites de l'IA). Côté Meta, deux lancements ont reçu des accueils mitigés au cours de l'année 2022 : **la tentative Blenderbot** de l'été, puis une seconde **pour la recherche, Galactica**, qui avait reçu un accueil « **au vitriol** » selon Yann LeCun (responsable scientifique de l'IA chez Meta). Dans le monde académique, dès 2020, un **papier de recherche critique de ces « grands modèles de langage »** provoque un scandale : certaines de ses autrices, Timnit Gebru et Margaret Mitchell, sont licenciées par Google à la suite de sa rédaction. En 2022, c'est un ingénieur de Google qui fait la une, Blake Lemoine, qui a vu dans cet IA une conscience – **ses déclarations entraînent également son renvoi**. Autant dire que le sujet des usages des modèles de langage n'est ni nouveau, ni apaisé…

Ce qui est nouveau, c'est que la relative stabilité de ChatGPT, comparée aux précédentes tentatives, a déclenché une « course à l'armement » entre les grandes plateformes, pour ne pas perdre du terrain. Les IA génératives d'images (Dall-e, Midjourney, StableDiffusion…) avaient également déjà séduit les utilisateurs au cours de l'année 2022. Pour les entreprises leaders de la tech, il ne faut donc pas louper le train de l'IA générative de texte, au risque de voir le marché être déjà conquis par un outil concurrent. Ainsi, après **Bard de Google**, il y a donc **Ernie, lancé par le chinois Baidu**, puis **LLaMA par Meta** (qui a d'ailleurs **fuité sur internet** !), qui n'ont pas encore été mis à l'épreuve du grand public… Les annonces d'implémentation se multiplient également : par Microsoft avec l'outil d'OpenAI dans son outil Teams (pour générer des comptes-rendus de réunion en visioconférence), dans Snapchat, etc. De fait, seules les grandes entreprises semblent aujourd'hui en capacité de tirer profit des IA génératives à court terme, et les acteurs qui ont déjà massivement investi dans le cloud **pourraient avoir une longueur d'avance** compte tenu du **coût estimé des requêtes**. Par ailleurs, le coût d'entrée après

la mise en place d'une potentielle régulation (comme celle sur l'IA en cours de discussion au Parlement Européen) pourrait être plus élevé. Le New York Times revient par ailleurs sur cette course, et notamment le retour en force de Microsoft, qui a fait le choix de parier avec OpenAI sur les larges modèles de langage plutôt que de **continuer avec son assistant vocal Cortana**, au contraire de ses concurrents.

Dans cette surenchère, il y a nécessairement une certaine prise de risque, comme en témoigne **un article du Washington Post** (qui cite le New York Times), dans lequel il est fait référence à une levée de certaines règles de contrôle par Google afin d'accélérer le lancement de ses propres outils, suite à l'engouement provoqué par ChatGPT : « *Google, qui a contribué à la mise au point d'une partie de la technologie permettant le développement de ChatGPT, a récemment publié un "code rouge" concernant le lancement de produits d'IA et a proposé une "voie express" pour raccourcir le processus d'évaluation et d'atténuation des dommages potentiels* » (traduction LINC). Cette prise de risque n'a pas été toujours payante par ailleurs : l'accueil de Bard a été mitigé, notamment pour des erreurs factuelles lors de la présentation de l'outil. Concernant la protection des données et des libertés, quels seraient ces « dommages potentiels » en dehors de la création du modèle et de son alimentation ? Il semble qu'ils soient de plusieurs ordres, concernant à la fois la sécurité des données par la facilitation – voire l'industrialisation – d'attaques, ou la réputation en ligne, par l'utilisation de données pour générer du contenu malveillant.

Des enjeux de générations malveillantes, textuelles notamment, mais pas que

Le fonctionnement de ChatGPT montre qu'il n'est pas à proprement parler un moteur de recherche, mais un générateur de texte : s'il peut être précis dans ses réponses, notamment sur des sujets pointus, il commet des erreurs factuelles. L'utilisation d'un système d'IA par le média en ligne CNET pour la production d'articles journalistiques en est l'illustration : ce type d'IA génératives a été utilisé dans le but de simplifier l'écriture des articles de la rubrique finance. Cependant, comme le rapporte **un article de Gizmodo**, des erreurs (notamment mathématiques) se sont glissées dans les articles générés automatiquement – des erreurs qui sont différentes de celles qu'une personne humaine ferait.

Parfois, ce sont plutôt des inventions complètes, mais qui sont cohérentes et plausibles : **la véracité n'est pas forcément son but premier**, ou tout du moins, elle peut entrer en conflit avec d'autres grandes règles mises au point par les concepteurs – c'est notamment **l'hypothèse de Scott Alexander dans son blog Astral Codex Ten**. En simplifiant, elle serait au nombre de trois : donner une réponse satisfaisante, dire la vérité – bien que cet aspect soit complexe – et ne pas offenser, et l'arbitrage de l'outil se fait parfois au détriment d'une pour satisfaire une autre. Au point où commence à émerger des descriptions moins élogieuses pour ChatGPT, comme celle de « bullshit generator », c'est-à-dire « générateur de baratin ». Pour autant, cela permet également de générer du texte « à la manière de » (de personnes célèbres, artistes, politiques, etc.) et ainsi, des créations plutôt intéressantes ! Elles le sont un peu moins lorsqu'elles sont utilisées pour créer des messages à la manière de l'administration fiscale ou d'une banque **afin de créer un mail d'hameçonnage**. Il n'aura pas fallu attendre longtemps pour voir des utilisations dans **des arnaques à la cryptomonnaie** ou encore **une fausse interface ChatGPT**, amenant au téléchargement d'un cheval de Troie.

Des tests ont également été faits autour de la connaissance technique de ChatGPT (voir nos articles d'exploration) pour savoir à quel point le chatbot peut devenir un assistant de code. Les performances de l'outil semblent plutôt bonnes, et donc suffisantes pour générer du code malveillant, permettant de passer à l'étape supérieur en termes d'attaque. De son côté, OpenAI cherche à dissuader toute tentative de ce genre d'utilisation de son outil, en faisant en sorte que certaines requêtes ne soient pas prises en compte ou refusées par ChatGPT. Pour autant, en tournant différemment ces requêtes, il reste possible d'accéder à la production de ce code, en découpant la demande en plusieurs morceaux, ou en passant **par du « jeu de rôle »**, qui permet dans certains cas de contourner les restrictions de l'outil. L'entreprise OpenAI indique être consciente de ces failles, et annonce « patcher » régulièrement son chatbot afin d'endiguer les mauvaises utilisations.

Réputation en ligne, accès à l'information et ChatGPT échaudé

La question des deepfakes se pose depuis plusieurs années (relire ici **l'interview de Nicolas Obin sur la synthèse vocale en 2019**), mais elle pourrait prendre une autre dimension avec les possibilités offertes par ces systèmes. Ils permettent, dans une certaine mesure, de pouvoir « industrialiser » des contenus qui pourraient porter atteinte à la réputation des personnes. En tant que « générateur », ChatGPT pourrait être un outil pratique dans le but de créer de la désinformation, puisqu'il est capable de mettre en forme un argumentaire cohérent et plausible, sans forcément s'astreindre à la véracité comme nous l'avons vu plus haut (voir **l'article du New York Times sur le sujet**).

Un autre écueil lié à l'accès à l'information et qui est propre aux chatbots de manière générale (et en particulier aux assistants vocaux, comme nous le pointions **dans notre Livre blanc paru en 2020**) est celui du moteur de réponse remplaçant le moteur de recherche. La sélection de la réponse se fait maintenant par l'outil, et non plus par l'utilisateur, qui ne peut plus naviguer entre plusieurs liens pour trouver la réponse souhaitée, ou alors confronter différents points de vue. A ceci s'ajoute l'absence de sources dans la réponse du chatbot, ne permettant pas de vérifier soi-même l'information. Le fait que ChatGPT puisse lui-même inventer des sources obscurcit également son usage. A sa décharge, ChatGPT n'était toutefois pas présenté comme un moteur de recherche : mais d'autres IA génératives ont été plus ambitieuses. En effet, Google a annoncé son IA Bard, qui dispose d'informations plus fraîches, car lié à son moteur de recherche, là où les informations sur lesquelles est entraîné ChatGPT s'arrêtent en 2021, date limite de la collecte de ses données d'entraînement. Mais Microsoft n'a pas tardé à réagir avec une démonstration de ChatGPT et de Bing, son propre moteur de recherche. Cette jonction n'a pas été la démonstration la plus convaincante, avec quelques ratés : le chatbot, **testé et titillé par des journalistes**, s'est « fâché », multipliant les « hallucinations ».

Se créent également les besoins d'une justification dans l'autre sens, comme en témoignent les répercussions dans le travail de certains « créateurs » de contenus (comme des journalistes **qui doivent montrer patte blanche**), et justifier que leur travail est bien « fait maison », dans un « **test de Turing inversé** ». Et ce, dans un mouvement où l'IA semble satisfaire certains éditeurs pour la publication d'articles « simples », à l'instar de Buzzfeed ou encore CNET, précédemment cité, jusqu'au **groupe de presse Axel Springer** ou même à des médias installés comme le Financial Times (qui utilise déjà Midjourney **pour illustrer certains articles**). Des magazines se retrouvent même submergés par **les soumissions d'articles ou de contenus générés par une IA**, certaines personnes y voyant une manière de faire facilement de l'argent. Enfin, il y aussi les « producteurs » de contenus, moins protégés par le droit d'auteur (voir **l'article 2 de ce dossier**), notamment via les blogs ou les forums. Ceux-là auront plus de mal à prouver que leur contenu a été utilisé pour entraîner le chatbot, ou à obtenir une quelconque rétribution pour leur travail, qu'elle soit pécuniaire ou en termes de visibilité : **la réponse des IA ne génère pas de visite sur le site source**, donc pas d'affichage ou abonnement possible (quand les sources ne sont pas inventées !).

Bien évidemment, cela concerne tout autant l'image que le son : pour le premier, l'exploration est déjà bien lancée (avec des outils comme Dall-e, Midjourney ou StableDiffusion). La génération d'images est également encadrée pour éviter les débordements, mais des détournements sont déjà présents, touchant notamment les femmes pour générer **des contenus à caractère pornographique**. Comme expliqué précédemment, le phénomène n'est pas si nouveau que ça, mais les outils pour le faire sont de plus en plus accessibles et donc ce genre de détournements de plus en plus industrialisé (ou industrialisable). Pour le domaine de la voix, si la synthèse vocale n'est pas encore aussi accessible que la génération de texte et d'image, plusieurs actualités montrent que l'on s'en approche : tout d'abord, une publication de Microsoft qui démontre **les capacités de son outil Vall-e** – sans qu'il soit possible de faire des tests –, mais aussi ce détournement illustré par le média Motherboard, « ***How I Broke Into a Bank Account With an AI-Generated Voice*** » - « Comment je me suis introduit dans un compte bancaire avec une IA de génération vocale ». Enfin, les premières arnaques (scam) générées grâce à un système d'IA ont vu le jour, **comme le rapporte le Washington Post** : des familles reçoivent des appels de proches, dont la voix est imitée leur demandant de l'aide financière urgente (mais dont les premiers essais réussis **datent de 2019**). Ce dernier point soulève notamment la question de la « preuve » de la reconnaissance du locuteur, comme nous l'évoquions déjà dans **l'article LINC Protection des témoins : casser la voix et l'image**.

La détection des contenus générés par une IA comme solution ? Sur internet, personne ne sait que vous êtes un chat(bot)

Le fait de pouvoir savoir si un contenu a été généré par une IA générative pourrait être une première solution à certains des enjeux listés dans cet article. Pour des objectifs de transparence d'abord (pour les médias l'utilisant notamment), mais également pour savoir quand il est probable qu'il y ait des erreurs dans un texte, pour éviter les tentatives de manipulation, ou encore pour indiquer qu'une image n'est pas réelle.

Il y aurait deux moyens techniques pour cela : le premier est une « empreinte » (***watermarking***), similaire à un filigrane, permettant de savoir avec certitude que le contenu a été généré. Cela serait donc à la charge du concepteur de l'IA de laisser cette empreinte. Ce procédé semble plus simple pour de l'IA générative d'image : il pourrait y avoir un marquage, de manière visible ou invisible pour les personnes (qui serait détectable automatiquement dans ce cas). Ainsi, il existe déjà des initiatives de label ou signalétique. Par exemple, **AI Label propose trois pictogrammes** : *No AI used* (pas d'utilisation d'IA), *Assisted by AI* (assisté par une IA) ou *Made with AI* (conçu avec une IA).

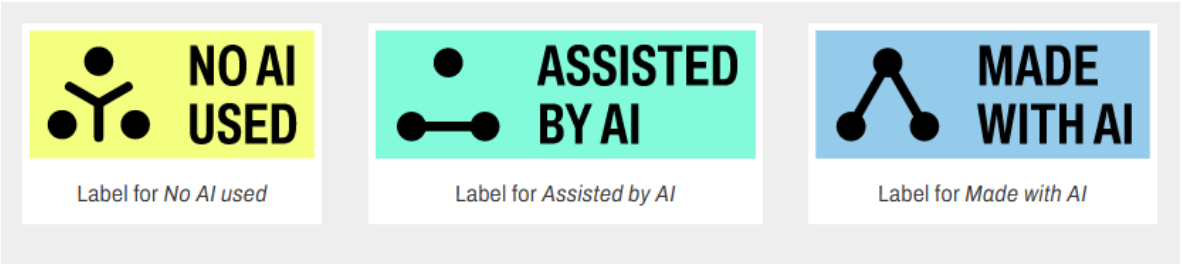


*Figure 1 : Pictogrammes proposés par **AI Label***

Cela pourrait être aussi le cas dans la génération de son ou de voix, avec l'ajout d'un signal audio, indétectable à l'oreille, mais qui pourrait être automatiquement reconnu par une machine. Pour le texte, cela semble plus compliqué : ce pourrait être l'insertion d'un biais (l'utilisation récurrente et le choix de certains mots, créant un « schéma » détectable par analyse automatique – voir, par exemple, **Kirchenbauer et al., 2023**).

La deuxième option est la détection ex-post, notamment en entrainant un algorithme à faire la différence entre des textes écrits par une IA et des textes rédigés par des humains. C'est notamment ce que propose OpenAI, **qui a lancé AI Text Classifie**r. Depuis le lancement de ChatGPT, d'autres outils ont été lancés en parallèle, comme **DetectGPT** ou encore **GPTZero**. Mais plusieurs problématiques se posent, à commencer par le taux de détection. Par exemple, l'outil d'OpenAI affiche une précision assez faible : 26% des textes générés par l'IA sont correctement labellisés comme tels (pour 9% de faux positifs, c'est-à-dire d'écrits humains attribués à une IA). D'autres outils sont mis en place par des acteurs tiers, **pour repérer les spams générés par ChatGPT**.

Certains points restent cependant complexes : d'abord, il faudra suivre les différentes itérations du modèle de ChatGPT pour répercuter les possibles changements et mettre à jour le système de détection ; ensuite, la taille du texte joue également un rôle pour pouvoir faire l'analyse. Les courts extraits offrent forcément moins de garantie. AI Text Classifier comme GPTZero annoncent d'ailleurs que leur outil n'est que pour une première analyse, afin de « marquer » les textes qui pourraient être rédigés par une IA. Qu'en sera-t-il pour les textes qui auront été modifiés partiellement ? Est-ce que le filigrane pourra tenir ? Dernier écueil : la question de la langue. GPTZero indique par exemple que ses données d'entraînement sont en langue anglaise… Limite à laquelle il faut ajouter aussi que la maîtrise d'une langue est également un facteur discriminant pour les étrangers : une moins bonne maîtrise de la langue, ou un langage en tout cas plus « scolaire » pourrait également amener à être classé comme étant une IA par rapport au texte que l'on a soumis. La limite de « à la manière de » se pose également : prendre un style journalistique (avec des formulations marquées comme « ce que l'on sait de… »). Si le but est de se rapprocher d'un style humain, les marqueurs de différenciation seront plus compliqués à déterminer.

Pour les images, l'ex-post pourrait se fonder sur ce qui est déjà existant en matière d'analyse de retouches d'image : par les méta-données, **l'autocorrélation ou taux de compression des pixels de l'image**, ou également par apprentissage – comme pour les textes – entre de « vraies » images et des images générées. Cela étant, il pourrait être ajouté du « bruit » sur l'image, afin de tromper l'analyse ou en tout cas de la rendre inopérante : impossible alors de savoir si l'image est réelle ou non. Tout ceci avec un certain paradoxe : les systèmes d'IA sont plus performants que les personnes dans la détection des contenus générés par un autre système d'IA.

Transparence et opacité : un retour aux sources pour reprendre la main ?

Les acteurs de l'écosystème ne sont bien évidemment pas aveugles à ces potentiels dommages, revers de la médaille des possibilités offertes par les systèmes d'IA génératives. A titre d'illustration, l'initiative Paterneship on AI, organisation à but non lucratif et regroupant différents types d'acteurs (industries, médias, société civile, recherche – dont OpenAI, Amazon, Apple, BBC, Berkeley, etc.), a proposé un cadre (framework) pour « **développer, créer et partager les médias synthétiques** » (donc, les contenus audiovisuels générés par les systèmes d'IA). La transparence apparaît donc comme une deuxième solution aux enjeux listés.

Il existe aussi d'autres pistes de réflexion dans la déconstruction des modèles, pour mieux les comprendre, notamment dans le cas des IA génératives d'image, à l'instar de **StableAttribution**, qui permet d'estimer quelles sont les images qui ont pu alimenter (et donc influencer) l'IA pour qu'elle produise cette image en particulier. Ou encore ***Have I been trained***, l'équivalent de ***Have I been pwned*** mais pour les images utilisées dans les modèles d'entraînement, et pouvoir activer son opposition (bien que…).

Car, l'une des principales problématiques de ces systèmes d'IA reste leur opacité, notamment les choix dans leurs règles de fonctionnement. Pour ChatGPT, ce sont ceux de ses concepteurs pour l'instant, mais l'intervention d'un des fondateurs d'OpenAI – ou un rachat par un riche magnat – pourrait entraîner un certain revirement dans la manière d'appréhender les choses. C'est une problématique classique du web 2.0, **à travers une standardisation qui n'est pas forcément discutée**, mais « imposée » par les grands acteurs.
L'opacité concerne aussi les conditions de labellisation des données, mais également la sous-traitance et les transferts des données. Dans le premier cas, **c'est notamment une révélation du Time** qui a permis de comprendre que cette labellisation avait été sous-traitée au Kenya, soulevant les problématiques classiques de « travail du clic » et de protection psychologique des personnes chargées de trier et décrire les contenus violents, pédopornographiques, etc. Une actualité qui se heurte avec le travail de modération pour Meta opéré dans le pays, dont les conditions sont dénoncées **par un ancien modérateur de la firme américaine**.

Une régulation nécessaire ?

Si les mesures techniques s'avèrent toutes limitées, la solution pourrait provenir d'une approche régulatoire encadrant plus précisément les obligations des fournisseurs d'IA génératives pour clarifier le statut des IA génératives et/ou des chatbots. C'est bien l'objectif poursuivi par les négociateurs du règlement IA qui s'interrogent sur l'inclusion des IA à finalité générale (« general purpose AI ») et des IA génératives dans les usages à haut risque. En France, le rapport du Comité Pilote d'Ethique du Numérique **sur les agents conversationnels** a émis 13 préconisations, 10 principes de conception et 11 questions de recherche à suivre pour ce type de systèmes. Parmi ceux-ci, « Affirmer le statut des agents conversationnels », « Réduire la projection de qualités morales sur un agent conversationnel » ou « Encadrer techniquement les deadbots » sont des préconisations dont la mise en œuvre permettrait sans doute de répondre aux enjeux précédemment cités.

Reste la question de l'effectivité de principes dégagés au sein des sociétés européennes utilisatrices vis-à-vis de systèmes conçus aux Etats-Unis ou en Chine…

La grande variété de ce que comprend les termes « systèmes d'IA » et de ce qu'ils permettent de faire entraînent de fait une complexité dans l'appréhension de tous les enjeux qu'ils soulèvent et des risques associés (**certains essaient déjà d'en mesurer l'impact dans différents domaines de la société**). Les derniers ajouts concernant les systèmes d'IA génératives **dans les discussions autour de l'IA Act** concrétisent les problématiques de la régulation : allant de la limitation des risques à l'adaptation des usages, notamment selon les domaines… Notamment lorsque les entreprises sont particulièrement actives **pour protéger leurs intérêts**.

Illustration - Pavel Danilyuk (Pexels)

*Article rédigé par **Martin Biéri**, Chargé d'études prospectives*

VOIR PLUS D'ARTICLES DE L'AUTEUR