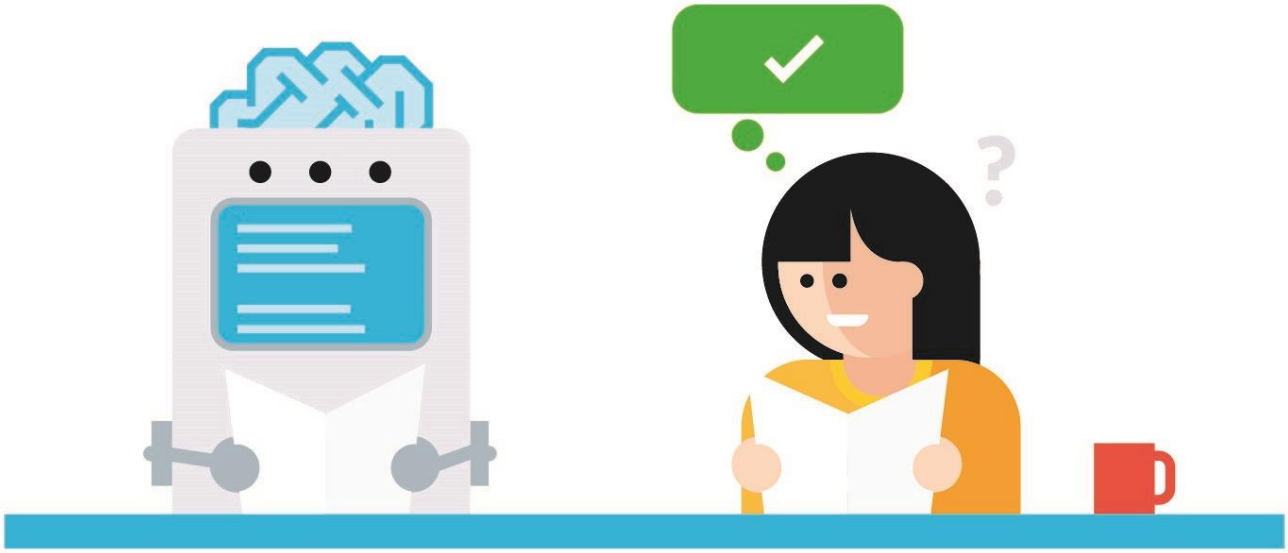


ChatGPT, ou quand l'intelligence artificielle se met à écrire comme nous – à quoi faut-il réfléchir ?

Table des matières

Lancé le 30 novembre 2022, ChatGPT a d'emblée stupéfié des millions d'utilisatrices et d'utilisateurs. Cet article présente les éléments principaux à connaître pour comprendre le phénomène, ainsi qu'une série de questions à garder à l'esprit à l'avenir, que cela soit sur le plan politique, professionnel ou personnel.

La Fondation pour l'évaluation des choix technologiques TA-SWISS a pour mandat d'examiner les effets de nouvelles technologies sur l'ensemble de la société. Dans cette optique, elle encourage le dialogue entre disciplines et l'échange de points de vue. L'apparition de ChatGPT et d'autres applications similaires concerne de nombreux secteurs d'activité, interroge des disciplines variées, et pourrait induire des changements conséquents pour une grande partie de la population. C'est pourquoi TA-SWISS examine l'opportunité d'une étude TA à ce sujet et continue d'observer de près les développements dans ce domaine.

De quoi s'agit-il ?

ChatGPT a été lancé par l'entreprise californienne OpenAI, spécialisée dans le développement d'intelligences artificielles. Comme son nom l'indique, il s'agit d'un agent conversationnel (*chatbot*), conçu à l'aide d'un modèle informatique capable d'analyser et de générer du texte en langage humain appelé *Generative Pre-trained Transformer 3* (GPT-3). En d'autres termes, ChatGPT est une intelligence artificielle avec laquelle il est possible de « converser » sur un site web. En l'espace de quelques secondes, l'application produit des réponses aux questions les plus variées, des blagues, des résumés de textes, ou encore des mails, des essais ou des poèmes. Elle peut aussi fournir des lignes de code informatique et signaler des fautes dans un code. Si la qualité des résultats s'avère inégale, un constat s'impose : les réponses sont souvent impressionnantes, et les textes très correctement formulés et structurés. À tel point qu'ils sont parfois difficiles à distinguer de ceux d'un être humain.

Bien qu'il occupe régulièrement la une des journaux, ChatGPT n'est pas la première application d'intelligence artificielle produisant des textes entiers. Il s'inscrit en réalité dans une continuité des développements de la recherche en intelligence artificielle (voir ci-dessous). Cependant, son apparition a marqué le moment où cette possibilité est devenue accessible à un large public. Et depuis, les développements s'accélèrent. À l'heure où nous écrivons ces lignes, Microsoft teste l'intégration d'une application suivant le même principe que ChatGPT dans son moteur de recherche Bing,[1] ainsi que dans sa suite de programmes Microsoft 365 (Word, Excel, PowerPoint, Teams, etc.).[2] Google[3] et Baidu[4] en Chine ont aussi déclaré ajouter prochainement une solution similaire à leurs moteurs de recherche. Alibaba travaille également sur un projet semblable à ChatGPT.[5] Par ailleurs, OpenAI propose désormais un abonnement payant à ChatGPT, avec des services plus étendus que la version gratuite, et a mis à disposition une interface de programmation (*API*) qui permet aux entreprises intéressées d'intégrer ChatGPT à leurs propres produits.[6] En parallèle, Open AI a sorti le modèle GPT-4 en mars 2023, une version encore plus performante que GPT-3.[7] ChatGPT et ses semblables ont donc de l'avenir devant eux, d'où la nécessité de réfléchir à leurs effets sur la société.

Études de TA-SWISS en lien avec ce sujet

- Étude « Quand les algorithmes décident à notre place : opportunités et risques de l'intelligence artificielle » (2020)
- Étude « Numérisation et démocratie : citoyens et institutions face à la numérisation de la démocratie en Suisse » (2021)
- Rapport « Les robots à la lumière de l'évaluation des choix technologiques » (2022)
- Étude en cours « Deepfakes et réalités manipulées » (publication prévue en 2024)
- Étude en cours « Culture et numérisation » (publication prévue en 2024)

[1] Voir le blog de Microsoft, « Reinventing search with a new AI-powered Microsoft Bing and Edge, your copilot for the web » (février 2023, https://blogs.microsoft.com/blog/2023/02/07/reinventing-search-with-a-new-ai-powered-microsoft-bing-and-edge-your-copilot-for-the-web/).

[2] Voir le blog de Microsoft, « Introducing Microsoft 365 Copilot » (mars 2023, https://blogs.microsoft.com/blog/2023/03/16/introducing-microsoft-365-copilot-your-copilot-for-work/)

[3] Voir le blog de Google, « An important next step on our AI journey » (février 2023, https://blog.google/technology/ai/bard-google-ai-search-updates/).

[4] Swissinfo, « Le chinois Baidu va lancer son propre robot face à ChatGPT » (février 2023, https://www.swissinfo.ch/fre/toute-l-actu-en-bref/le-chinois-baidu-va-lancer-son-propre-robot-face-%C3%A0-chatgpt/48264576).

[5] Voir CNBC, « Chinese tech giant Alibaba working on a ChatGPT rival; shares jump » (février 2023, https://www.cnbc.com/2023/02/08/chinese-tech-giant-alibaba-working-on-a-chatgpt-rival-shares-jump.html).

[6] Voir le blog de OpenAI, « Introducing ChatGPT and Whisper APIs » (février 2023, https://openai.com/blog/introducing-chatgpt-and-whisper-apis).

[7] Voir le blog de OpenAI, « GPT-4 » (mars 2023, https://openai.com/research/gpt-4).

Comment fonctionne ChatGPT ?

Pour l'instant, en mars 2023, tout le monde peut ouvrir un compte gratuitement sur le site de ChatGPT, à condition d'indiquer des coordonnées détaillées. Il s'agit ensuite d'y entrer une question ou une requête (*prompt*), avec la possibilité de préciser la longueur, la structure et le ton du texte à produire, ou les thèmes à aborder. Pour exécuter ces requêtes, ChatGPT a été « entraîné » à l'aide de milliards de textes disponibles sur Internet jusqu'à fin 2021. Cela signifie que l'ensemble de ces textes a été analysé mathématiquement pour en saisir les régularités linguistiques. En effet, ChatGPT fonctionne sur la base de statistiques concernant l'ordre des mots dans ce corpus et calcule la probabilité qu'un mot en suive un autre. Ainsi, ses réponses reflètent un agencement de mots récurrent dans les textes qui l'ont nourri, et ont donc une forte composante aléatoire : raison pour laquelle ChatGPT produit une autre réponse si on lui repose la même question.

Afin d'impliquer des êtres humains dans le développement du modèle, OpenAI a fait appel à l'apprentissage par renforcement à partir de la rétroaction humaine (*Reinforcement Learning from Human Feedback*). L'entreprise a engagé du personnel pour évaluer une grande quantité de réponses, en validant les résultats plausibles et en écartant les résultats non acceptables, afin d'améliorer l'algorithme de ChatGPT. Par ailleurs, OpenAI s'est efforcée d'effectuer un calibrage éthique, notamment pour éviter les propos violents, ouvertement sexistes ou racistes, mais aussi les jugements de valeur ou les instructions utiles à des actes illégaux ou dangereux (voir ci-dessous). Depuis le lancement de l'application, les requêtes des utilisateurs et utilisatrices y sont analysées afin de continuer son entraînement.

Plus largement, ChatGPT s'insère dans le contexte de **l'intelligence artificielle dite « générative »**, en essor depuis quelques années déjà. Il s'agit de modèles produisant des images, des textes ou des vidéos à partir de quelques mots-clés seulement. Ces contenus « de synthèse » sont eux aussi le fruit d'une analyse purement statistique d'une multitude de données, sur la base d'un « apprentissage profond » (*deep learning*).[8] En ce qui concerne la génération de textes, toutes ces innovations découlent du développement de « **modèles de langage** ». Ces modèles informatiques répertorient les schémas de séquences de mots dans les textes avec lesquels ils sont entraînés, en vue de reproduire ces schémas lors de la génération de nouveaux textes.[9] Afin de garantir une cohérence d'une phrase à l'autre, ces modèles saisissent des relations de dépendance entre les mots d'un texte entier, et selon différents contextes (grâce à une technique nommée *self-attention*, développée chez Google en 2017).[10]

ChatGPT, lui, repose sur **GPT-3**. Créé par OpenAI, il s'agit aujourd'hui du plus vaste modèle de langage au monde : son entraînement repose sur 175 milliards de paramètres d'analyse et 570 gigabytes de textes.[11] En parallèle, Google a également développé ses modèles de langage (BERT, puis LaMDA), tout comme Meta (XLM-R, et plus récemment LLaMa[12]). Il existait d'ailleurs d'autres générateurs de textes avant ChatGPT. Par exemple, Jasper Chat ressemble à ChatGPT et a été mis sur pied en partenariat avec OpenAI.[13] De même, la plateforme Github Copilot, à laquelle OpenAI a également participé, permet de générer du code à la demande.[14]

---

[8] Voir par exemple les applications Dall-E (elle aussi lancée par OpenAI) et Stable Diffusion pour la génération d'images, ou Synthesia pour la génération de vidéos.

[9] J. Goldstein, G. Sastry, M. Musser, R. DiResta et al., *arXiv* (2023, https://arxiv.org/abs/2301.04246), pp. 15-19.

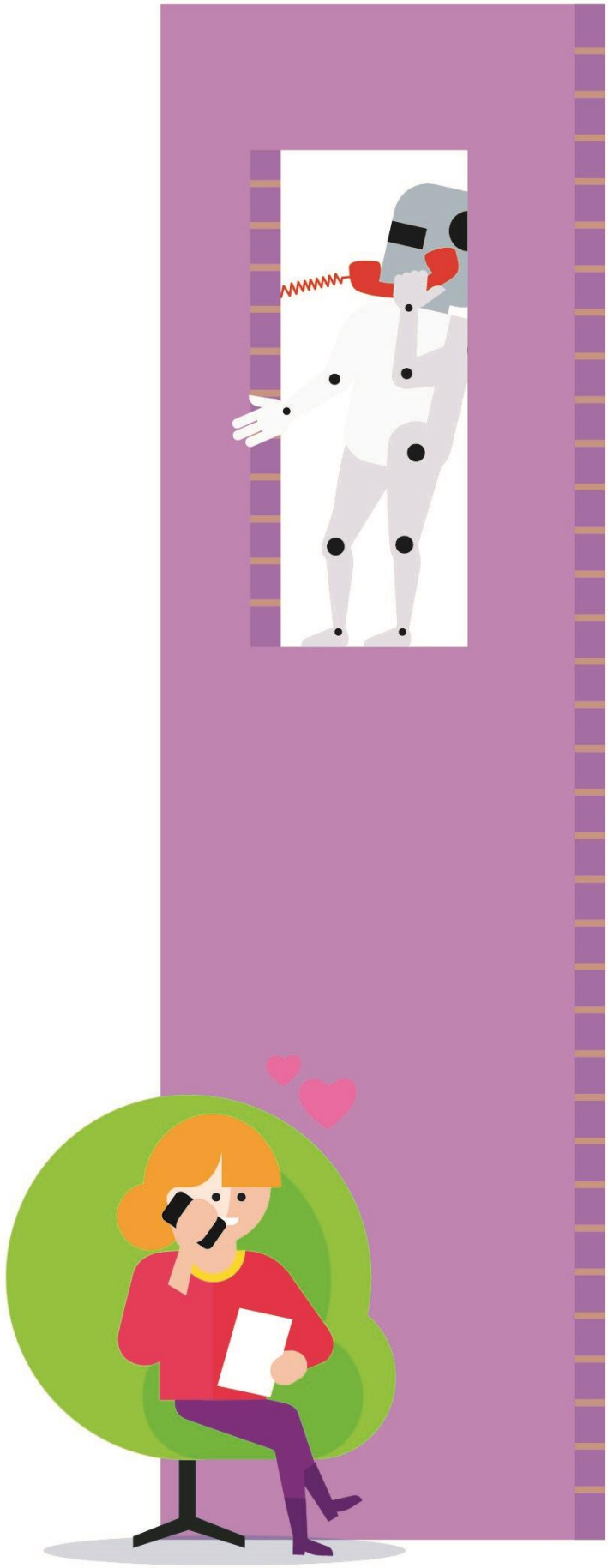
[10] A. Vaswani, N. Shazeer, N. Parmar et al., « Attention Is All You Need », *Advances in neural information processing systems* (2017, https://arxiv.org/pdf/1706.03762.pdf).

[11] J. Goldstein, G. Sastry, M. Musser et al., « Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations », *arXiv* (2023, https://arxiv.org/abs/2301.04246). p. 2.

[12] Voir le blog de Meta AI, « Introducing LLaMA: A foundational, 65-billion-parameter large language model » (février 2023, https://ai.facebook.com/blog/large-language-model-llama-meta-ai/).

[13] Voir Jasper Chat, et pour d'autres exemples, C. Agar, « The Best AI Novel Writing Software For 2022 », *TheTechReviewer* (2022, https://thetechreviewer.com/best-ai-novel-writing-software/).

[14] Voir https://github.com/features/copilot.



Une mine d'opportunités, mais aussi de questions économiques en suspens

À en croire de multiples témoignages circulant dans les médias, ChatGPT est d'ores et déjà un outil qui connaît un grand engouement. Que cela soit pour obtenir des réponses ou des idées pour un projet, ou pour planifier les étapes à suivre pour l'exécution d'une tâche, ChatGPT fournit des ressources touchant à des domaines d'activités très variés. Ainsi, l'application est souvent perçue comme une promesse de gains de temps et de productivité, pour la rédaction de texte et de code, ou la recherche d'informations.[15] En même temps, de nouvelles compétences semblent émerger pour tirer le meilleur d'applications comme ChatGPT, notamment en leur posant les questions les plus pertinentes (« *prompt engineering* »).[16]

Pour ces mêmes raisons, ChatGPT et ses successeurs risquent bien d'altérer les domaines impliquant la production d'écrit – en langage naturel ou informatique. Parmi les secteurs les plus visiblement concernés figurent le journalisme, la communication et le marketing, mais aussi la programmation informatique, le droit ou la recherche scientifique, et même certains formats littéraires. Par ailleurs, si une application comme ChatGPT était insérée dans un moteur de recherche, les visites de sites Internet pourraient décroître et entraîner des pertes de revenus pour leurs propriétaires.[17] On retrouve ici un défi récurrent des progrès de l'intelligence artificielle : quels emplois pourraient être transformés, voire remplacés par ces technologies ? En même temps, quelles nouvelles activités pourraient-elles créer ? Et comment s'adapter à ces nouveaux outils ?

Dans un autre registre, il convient de relever que la mise en place et l'utilisation de tels modèles de langage est extrêmement **coûteuse**. Ces modèles requièrent une vaste quantité de données pour leur entraînement et leur adaptation au fil du temps, ce qui nécessite une énorme puissance de calcul sur leurs serveurs. Il est vrai que ces modèles peuvent être mis en libre accès par la suite (*open source*). D'autres personnes peuvent alors les adapter, en y rajoutant leurs données ou en complétant certaines fonctions (selon la méthode du *transfer learning*). C'est le cas pour GPT-2, mais pas encore GPT-3. Néanmoins, les coûts de base et la possession des données donnent un avantage massif, et donc une large influence, aux géants de la tech.[18] Comment la grande concurrence de ces entreprises sur ces produits se répercutera-t-elle sur l'économie mondiale ? Comment la Suisse peut-elle se positionner dans ce domaine ?

Sur le plan de la durabilité, les gros systèmes comme GPT requièrent une **grande consommation d'énergie** pour exécuter tous les calculs de leur entraînement, puis les requêtes des utilisatrices et utilisateurs.[19] Les recherches actuelles tentent certes d'élaborer des méthodes de calcul et des matériaux moins gourmands en énergie. Néanmoins, si les serveurs de ces applications ne tournent pas avec une énergie durable, leur empreinte carbone risque d'être massive. Par ailleurs, des **conditions de travail précaires** ont été dénoncées dans une partie de la chaîne de production des systèmes d'intelligence artificielle. En particulier, leur encadrement humain repose sur des actions répétitives visant à contrôler le maximum de résultats en un minimum de temps, y compris des contenus parfois très violents. Cette tâche est souvent déléguée à des sous-traitants exerçant une grande pression sur leur personnel pour

---

[15] Voir par exemple M. Chui, R. Roberts et L. Yee, « Generative AI is here: How tools like ChatGPT could change your business », *McKinsey* (2022, https://www.mckinsey.com/capabilities/quantumblack/our-insights/generative-ai-is-here-how-tools-like-chatgpt-could-change-your-business).

[16] Voir par exemple D. Holzer, « KI schafft neue Jobs: Was macht eigentlich ein Prompt Engineer? », *Br24* (2023, https://www.br.de/nachrichten/netzwelt/ki-schafft-neue-jobs-was-macht-eigentlich-ein-prompt-engineer,TX4P23Z).

[17] Voir A. Seydtaghia, « Le match des moteurs de recherche est relancé grâce à l'intelligence artificielle », *Le Temps* (2023, https://www.letemps.ch/economie/match-moteurs-recherche-relance-grace-lintelligence-artificielle).

[18] Voir R. Fulterer « KI-Forschung nach dem Vorbild Cern: Die ehemalige Spass-Firma Hugging Face fordert Meta und Google heraus », NZZ (2022, https://www.nzz.ch/technologie/huggingface-diese-nach-einem-emoji-benannte-firma-fordert-meta-und-google-heraus-ld.1682435).

[19] Voir D. Patterson, J. Gonzalez, Q. Le et al., « Carbon Emissions and Large Neural Network Training », *arXiv* (2022, https://arxiv.org/ftp/arxiv/papers/2104/2104.10350.pdf) ou L. Anthony, B. Kanding et R. Selvan, « Carbontracker: Tracking and Predicting the Carbon Footprint of Training Deep Learning Models » *arXiv* (2020, https://arxiv.org/abs/2007.03051).

de très bas salaires, dans des pays fortement touchés par la pauvreté.[20] D'après une récente enquête du *Time* au Kenya, le cas de ChatGPT semble confirmer la règle.[21]

En revanche, l'utilisation de ChatGPT à des fins commerciales peut aller à l'encontre de droits de **propriété intellectuelle** rattachés aux textes analysés par ChatGPT, notamment s'il reprend des formulations protégées, mais sans l'indiquer. Comment identifier les données qui ont contribué à former une réponse et déterminer si une infraction a véritablement eu lieu ? Et de manière générale, faut-il rémunérer l'utilisation de ces données originelles par ChatGPT, et selon quel modèle (quelle part reviendrait aux propriétaires des données utilisées, ou à OpenAI ?)[22]

De l'école à l'université, un tournant pour la formation

Suite au lancement de ChatGPT, c'est également les institutions de formation qui se sont retrouvées en ébullition. ChatGPT se prête facilement à la rédaction de passages de travaux écrits ou à la recherche d'idées pour faire un devoir, et semble capable de répondre à des questions d'examen. En l'état actuel, cela signifie de nouvelles perspectives de triche et de plagiat disponibles en quelques clics. Selon diverses expériences relayées par les médias, les résultats de ChatGPT ne sont pas toujours éblouissants, mais souvent suffisants pour passer un test, voire obtenir une très bonne note.[23] S'en remettre à l'application pourrait également entraver l'acquisition des connaissances et compétences visées par un exercice, puisque la machine nous dispenserait de l'effort à fournir. Cela pourrait écarter les processus de l'apprentissage impliquant notre compréhension et conscience. Par ailleurs, la rapidité des réponses pourrait dépasser le temps nécessaire à notre traitement des informations.[24] Cependant, de nombreux enseignants et enseignantes y voient également des opportunités pour leur cursus, par exemple en évaluant de manière critique les réponses de ChatGPT.[25]

En plus de remettre en question les systèmes d'évaluation actuels, ChatGPT suscite une large réflexion sur les **méthodes d'apprentissage**, mais aussi de production de savoir. Dans quelle mesure et de quelle manière faut-il intégrer ce type d'intelligence artificielle dans l'enseignement ? Comment en accompagner l'usage, sachant qu'une interdiction totale semble difficile à mettre en œuvre ? Quelles sont les compétences liées à l'intelligence artificielle à développer dans les formations ?[26] Où faut-il mettre des limites pour préserver les compétences humaines ? Et quelle légitimité pourrait-elle avoir dans la recherche scientifique, par exemple en tant qu'outil de recherche, de synthèse ou de rédaction ?[27]

Parmi les mesures envisagées, il est question de développer des applications permettant de **détecter des textes fabriqués par une intelligence artificielle**.[28] Un chercheur d'OpenAI a notamment men-

---

[20] Voir Antonio Casalli, *En attendant les robots. Enquête sur le travail du clic* (2019, Le Seuil).

[21] B. Perrigo, « Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic », *Time* (2023, https://time.com/6247678/openai-chatgpt-kenya-workers/).

[22] I. Wildhaber, « ChatGPT aus rechtlicher Perspektive: Was Unternehmen beachten sollten », Universität St. Gallen (2023, https://www.unisg.ch/de/videodetail/news/chatgpt-aus-rechtlicher-perspektive-was-unternehmen-beachten-sollten/).

[23] Voir P. Kaelin, « Luzerner Hochschulen wappnen sich gegen ChatGPT », *zentralplus* (2023, https://www.zentralplus.ch/beruf-bildung/luzerner-hochschulen-wappnen-sich-gegen-chatgpt-2512228/).

[24] Voir M. de Vevey, « ChatGPT : à quoi ressemblera l'enseignement dans dix ans ?, *uniscope* (2023, https://wp.unil.ch/uniscope/chatgpt-a-quoi-ressemblera-lenseignement-dans-dix-ans/).

[25] Service écoles-médias du canton de Genève, « Risques et opportunités de ChatGPT pour l'enseignement : premiers éléments d'analyse », (2023, https://edu.ge.ch/sem/ressources/risques-et-opportunites-de-chatgpt-pour-lenseignement-premiers-elements-danalyse-3574).

[26] À ce sujet, voir l'étude de TA-SWISS sur l'intelligence artificielle, pp. 171-8 et 301-2.

[27] Voir C. Stokel-Walker et R. Van Noorden, « What ChatGPT and generative AI mean for science », *Nature* (2023, https://www.nature.com/articles/d41586-023-00340-6).

[28] Voir les applications *DetectGPT* (E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning et C. Finn, « DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature », Université de Cornell (2023, https://arxiv.org/pdf/2301.11305.pdf) ou *GPTZero* (C. Cassidy, « College student claims app can detect essays

tionné le projet d'élaborer une marque invisible dans les métadonnées des textes de ChatGPT : une sorte de « filigrane » certifiant qu'un texte a été produit par ChatGPT.[29] Cependant, diverses questions demeurent : qui détiendrait l'accès à ces outils, sachant qu'ils peuvent révéler des informations sur les textes de tout un chacun ? Quel serait leur degré de fiabilité ? Et que faire face aux applications concurrentes servant à contourner ces outils et qui ne manqueront pas d'apparaître ?

Dans une optique d'accès à la formation, il convient également de mentionner que ce type d'applications pourrait réduire certains obstacles à l'apprentissage, par exemple en reformulant des contenus en langage simplifié.[30] Elles pourraient aussi apporter une aide en cas de difficultés à s'exprimer par écrit, que cela soit de manière générale ou dans une langue étrangère. Reste la question du prix de ces services, qui pourrait réduire ces opportunités.

Limites de ChatGPT et risques à surveiller

Malgré le consensus sur le progrès technique qu'elle incarne, l'application comporte des risques considérables à surveiller. Tout d'abord, les modèles de langage statistiques opèrent avec des probabilités, sans raisonnement ou interprétation. Il est donc essentiel de souligner que les notions de vrai ou faux n'existent pas dans ces systèmes, et qu'ils ne peuvent « comprendre » les contenus qu'ils traitent. Ainsi, ChatGPT fait beaucoup d'**erreurs factuelles**. De même, lorsqu'il s'agit d'indiquer des sources, il n'est pas rare que ChatGPT invente des liens ou des articles de toutes pièces.[31] Seule une personne connaissant déjà bien un domaine est à même d'évaluer la réponse à une question sur ce sujet. Or, les réponses sont souvent formulées avec assurance, et de manière convaincante sur le plan formel. Cela peut donner l'impression à l'utilisateur ou l'utilisatrice d'avoir affaire à une haute compétence, voire à une personne qui réfléchit en face. Le risque que des individus se fient à de fausses indications est donc tangible, et peut s'avérer lourd de conséquences, par exemple en cas de questions médicales, politiques ou légales. Le remplacement d'un jugement humain par un calcul statistique de mots pose également un problème fondamental d'éthique.[32]

Par ailleurs, les applications génératrices de textes peuvent se faire orienter de manière abusive. Par exemple, elles pourraient être employées pour alimenter des campagnes d'influence ou de **désinformation** de masse, en générant rapidement des phrases adaptées à chaque contexte, notamment sur les réseaux sociaux ou les forums.[33] En outre, ChatGPT pourrait contribuer à fabriquer des lignes de code servant à une **cyberattaque**.[34]

written by chatbot ChatGPT », *Guardian* (2023, https://www.theguardian.com/technology/2023/jan/12/college-student-claims-app-can-detect-essays-written-by-chatbot-chatgpt)).

[29] Voir S. Aaronson, « My AI Safety Lecture for UT Effective Altruism » (2022, notes de conférence, https://scottaaronson.blog/?p=6823).

[30] Voir B. Blume, *Deutsches Schulportal der Robert Bosch Stiftung*, « ChatGPT: Das Ende vom Lernen wie wir es kennen » (2023, https://deutsches-schulportal.de/kolumnen/chatgpt-das-ende-vom-lernen-wie-wir-es-kennen/).

[31] À noter qu'une solution incorporée dans un moteur de recherche pourrait indiquer des sources. Voir A. Glaese, N. McAleese, M. Trebacz et al., « Improving alignment of dialogue agents via targeted human judgements », *DeepMind* (2022, https://arxiv.org/pdf/2209.14375.pdf), où il est question d'une application de Google appelée Sparrow, ou le prototype WebGPT d'OpenAI (https://openai.com/blog/webgpt/).

[32] Concernant la nécessité d'un contrôle de décisions automatisées par un être humain, voir l'étude de TA-SWISS sur l'intelligence artificielle, pp. 136-7 et 293.

[33] J. Goldstein, G. Sastry, M. Musser et al., « Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations », *arXiv* (2023, https://arxiv.org/abs/2301.04246).

[34] Voir l'expérience réalisée par des spécialistes en cybersécurité de l'entreprise Check Point Software (S. Ben-Moshe, G. Gekker, G. Cohen, « OpenAI: AI That Can Save the Day or HACK it Away », *cp<r>* (2022, https://research.checkpoint.com/2022/opwnai-ai-that-can-save-the-day-or-hack-it-away/).

Les **données d'entraînement** de l'application constituent elles aussi une source d'inquiétude, récurrente dans le domaine de l'intelligence artificielle. Ces données proviennent de textes humains, véhiculant les opinions et attitudes de celles et ceux qui les ont écrits. Ainsi, elles contiennent également des biais et stéréotypes, qui se retrouvent ensuite dans les réponses du programme et renforcent les discriminations reflétées dans ces données. Le plus fréquemment, ce sont les femmes et les minorités qui sont concernées en première ligne.[35] Malgré les mesures de précaution d'OpenAI, de nombreuses assertions discriminatoires de ChatGPT ont été répertoriées sur les réseaux sociaux.[36] Sur le plan technique, il s'avère difficile d'éviter ou de corriger les biais de données constituant un système d'intelligence artificielle. Par ailleurs, l'identification même de ces biais relève d'une démarche sociale, qui repose sur des jugements normatifs et ne se réduit donc pas à des paramètres techniques.

Dans ce souci d'équité, une **représentation** adéquate de toutes les langues et cultures semble elle aussi peu probable, car les grands modèles de langage actuels ont principalement été développés à partir de l'anglais. De même, leur entraînement sur la base de contenus présents en ligne opère une sélection, puisque certains points de vue sont très peu représentés sur Internet, tandis que d'autres y sont amplifiés, par exemple sur les réseaux sociaux.[37] En revanche, l'utilisation de ces programmes pourrait conduire à une **standardisation** des opinions, des informations et du savoir, car ils fonctionnent de manière purement statistique. Tandis que les contenus plus rares auront tendance à être écartés, les positions les plus fréquentes seront favorisées. Cet effet risque d'être difficile à repérer, ainsi qu'à surveiller, car les utilisatrices et utilisateurs n'ont pas accès au processus de production d'une réponse. En revanche, on peut s'attendre à ce que les futures applications ingèrent de plus en plus de textes artificiels – ce qui pourrait renforcer cette dynamique de standardisation.[38]

[35] Voir E. M. Bender, T. Gebru, A. McMillan-Major, et S. Shmitchell: « On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? », *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, *Association for Computing Machinery* (2021, https://dl.acm.org/doi/abs/10.1145/3442188.3445922).

[36] Voir par exemple I. Vock, « ChatGPT proves that AI still has a racism problem », *New Statesman* (2023, https://www.newstatesman.com/quickfire/2022/12/chatgpt-shows-ai-racism-problem).

[37] « On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? », op cit.

[38] Voir K. Creel et D. Hellman, « The Algorithmic Leviathan: Arbitrariness, Fairness, and Opportunity in Algorithmic Decision Making Systems », *Virginia Public Law and Legal Theory* (2021, https://ssrn.com/abstract=3786377).

Face à tous ces risques se pose la question de savoir quels sont les usages que nous souhaitons encourager, et quelle forme leur donner. Comment ces nouvelles applications s'intègrent-elles dans les **normes** existantes en matière d'intelligence artificielle ? Faut-il imaginer des normes supplémentaires pour encadrer le développement ou l'utilisation de modèles de langage ? Quelles sont les options envisageables en pratique, au vu de l'omniprésence d'Internet et de son caractère international ? Quelles exigences éthiques faut-il appliquer à ces modèles, et qui décide lesquelles de leurs réponses sont acceptables ? Comment garantir une transparence sur ces critères et décisions ?

Par ailleurs, qui est responsable de problèmes découlant de l'utilisation de ces outils ? Comment sensibiliser le public aux limites et à la faillibilité de ces chatbots ? Que fait-on des données personnelles des utilisateurs et utilisatrices, et quelles sont les mesures de protection des données à prendre ? Que se passe-t-il lorsqu'une personne écrit des informations sensibles dans l'application ? Ces nombreux défis montrent l'importance d'un large débat public.

Des questionnements sur l'avenir de nos textes

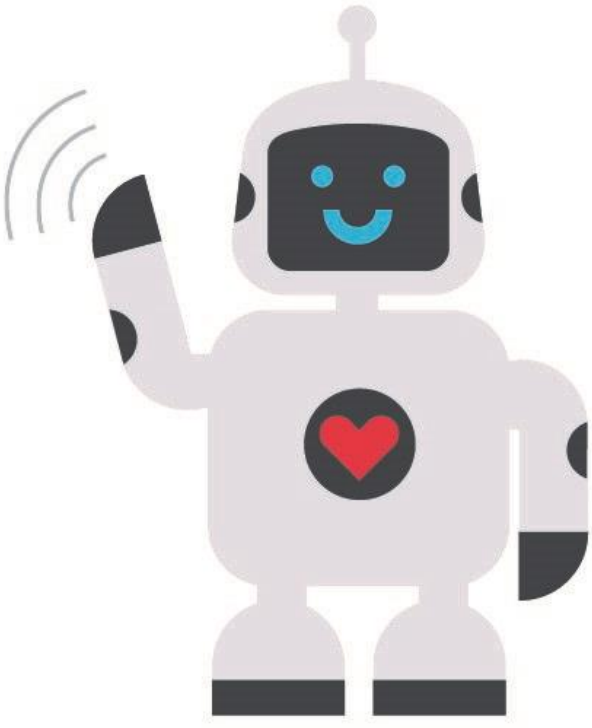
Si les problèmes d'éthique apparaissent clairement, l'émergence d'intelligences artificielles capables de produire des textes ressemblant à ceux d'un être humain nous interroge également sur nos pratiques d'écriture et de lecture - et donc sur notre rapport au monde. Quand un calcul d'une suite de mots probable dans des textes sélectionnés pour un modèle informatique peut-il nous suffire ou non, et pourquoi ? Y a-t-il des caractéristiques propres aux textes d'une intelligence artificielle, par rapport aux nôtres, et que révèlent-elles ? Comment l'utilisation de ces outils va-t-elle influencer la production de texte et notre réflexion durant le processus d'écriture ?

Aujourd'hui déjà, de nombreux textes en ligne proviennent d'une intelligence artificielle, sans que nous ne le remarquions forcément. Si des applications comme ChatGPT se généralisent, leur nombre augmentera encore. Cela soulève la question de savoir quand nous devrions être au courant de l'origine artificielle d'un texte. Les contenus générés par une machine devraient-ils être déclaré comme tels ? Par ailleurs, il convient de s'interroger sur les effets d'une augmentation de textes produits par une intelligence artificielle sur nos langues, et comme le langage façonne la pensée, sur nos visions du monde. Comment appréhender l'influence d'une « pensée » statistique et axée sur des probabilités ? Faudrait-il développer des modèles se basant sur des écrits humains uniquement ?

Face à ces questions également, une vaste réflexion est nécessaire pour déterminer le dosage adéquat d'interactions entre textes humains et artificiels. Comme l'ont suggéré ces quelques pages, une approche interdisciplinaire et multisectorielle est incontournable pour appréhender les effets de ces technologies et les réactions à adopter. La Fondation TA-SWISS continue donc d'observer de près les développements dans ce domaine. Vos commentaires sont les bienvenus.

Nous remercions vivement le Dr Bruno Baeriswyl, le Prof. Antoine Bosselut, le Dr. Olivier Glassey, le Prof. Lorenz Hilty, la Dre Anna Jobin et le Prof. Reinhard Riedl pour leurs précieux commentaires lors de l'élaboration de ce texte.

Laetitia Ramelet,
cheffe de projet chez TA-SWISS,
mars 2023



TA-SWISS
Stiftung für Technologiefolgen-Abschätzung
Brunngasse 36
3011 Bern

www.ta-swiss.ch

mitglied der
akademien der wissenschaften schweiz