Qu’est-ce que l’intelligence artificielle?

Un guide

Habilo Médias

Meta

GUIDE

QU'EST-CE QUE L'INTELLIGENCE ARTIFICIELLE?

Vous avez sûrement déjà entendu parler de l'intelligence artificielle (IA). Vous avez peut-être lu des articles vantant ses prouesses impressionnantes, comme sa capacité à créer des images et des vidéos en quelques minutes ou de tenir des conversations qui paraissent totalement humaines. L'IA est souvent qualifiée de véritable « révolution » pour les personnes en situation de handicap[1], car elle permet d'automatiser des tâches très chronophages et fastidieuses. Il a également été constaté que les agents conversationnels d'IA contribuent à réduire le sentiment de solitude chez les utilisateurs[2]. Mais qu'est-ce que l'IA exactement, et à quoi faut-il être vigilant? Quels en sont les bénéfices, et quels en sont les dangers?

Ce guide offre un aperçu de ce qu'est l'IA, en particulier l'IA *générative*, et présente deux exemples d'outils d'IA que vous êtes susceptible de rencontrer. Il aborde ensuite les principaux enjeux éthiques et sociaux liés à l'IA générative.

Qu'est-ce que l'IA?

L'**IA** (intelligence artificielle) utilise des *algorithmes informatiques* pour accomplir des tâches avec peu ou pas d'intervention humaine.

Un **algorithme** est une série d'instructions ou d'étapes pour réaliser une tâche. Contrairement aux algorithmes traditionnels, les algorithmes d'IA ne sont pas programmés manuellement, mais entraînés. Cela signifie qu'ils apprennent à partir d'un ensemble de données, comme une collection de millions d'images ou de textes. Ils identifient des motifs ou des corrélations dans ces données et s'en servent pour résoudre le problème pour lequel ils ont été conçus.

« Il n'est pas nécessaire de fournir une liste détaillée d'instructions… Vous donnez à la machine des données, un objectif, et vous lui indiquez quand elle est sur la bonne voie — puis vous la laissez trouver la meilleure façon d'atteindre le but. » — Hannah Fry, Hello World

Les algorithmes d'IA sont beaucoup plus puissants et flexibles que ceux créés par des humains, mais ils sont aussi plus complexes à analyser et à comprendre. On peut savoir quelles données sont utilisées et voir les résultats qu'ils produisent, mais le processus entre les deux reste souvent flou. C'est pourquoi l'IA est parfois qualifiée de « boîte noire ». De plus, les algorithmes d'IA les plus avancés peuvent évoluer et s'adapter avec le temps, si bien que même leurs concepteurs

1 Aquino, S. (2024) AI could be a game changer for people with disabilities. *MIT Technology Review*. https://www.technologyreview.com/2024/08/23/1096607/ai-people-with-disabilities-accessibility

2 De Freitas, J., Uguralp, A. K., Uguralp, Z. O., & Stefano, P. (2024). AI Companions Reduce Loneliness. *arXiv preprint arXiv:2407.19096.*

et utilisateurs ne savent pas toujours précisément comment ils fonctionnent.

Même si l'IA n'est pas programmée de manière traditionnelle, l'intervention humaine reste indispensable dans son processus d'apprentissage. Les humains font des retours en notant la qualité des réponses, en ajoutant des légendes ou des annotations aux données, ou en s'assurant qu'elle ne génère pas de contenu inapproprié, violent ou explicite[3]. L'IA « semble si humaine parce qu'elle a été entraînée par une IA qui imitait des humains, qui eux-mêmes évaluaient une IA imitant des humains, en prétendant être une version améliorée d'une IA formée à partir d'écrits humains[4]. »

Qu'est-ce que l'IA générative?

L'IA générative désigne les systèmes d'intelligence artificielle capables de créer des contenus comme des images, des vidéos, des voix ou du texte. Elle fonctionne en *codant* d'abord de nombreux exemples du type de contenu à générer, puis en les *décodant* pour générer du contenu inédit[5].

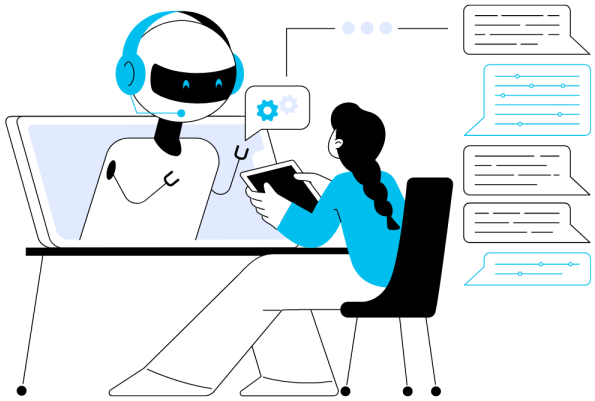
Du point de vue de l'utilisateur, utiliser l'IA générative commence par soumettre ce qu'on appelle une *requête* : c'est-à-dire une description de ce que vous souhaitez que l'IA crée (texte, image, vidéo, musique, etc.). Cela peut être aussi simple qu'une demande basique (« une pomme »), ou bien inclure des directives plus précises (« une pomme dans le style pointilliste de Cézanne »). On peut également imposer des contraintes à l'IA ou lui demander d'adopter un rôle spécifique.

Examinons les deux exemples les plus courants d'IA générative :

AGENTS CONVERSATIONNELS

Les agents conversationnels, capables de générer du texte, de répondre à des questions et de tenir des conversations, reposent sur une forme d'IA appelée **grand modèle de langage**.

Pour mieux comprendre ce que cela signifie, décomposons ce terme :



Modèle : Comme d'autres algorithmes d'apprentissage automatique, les agents conversationnels ne doivent pas leurs capacités à la programmation, mais à un entraînement sur d'immenses volumes de textes. Ils y repèrent des motifs qui leur servent à créer un modèle de fonctionnement du langage.

3 Hao, K., & Seetharaman D. (2023) Cleaning Up ChatGPT Takes Heavy Toll on Human Workers. *The Wall Street Journal*.

4 Dzieza, J. (2023) AI Is A Lot of Work. *New York*. https://nymag.com/intelligencer/article/ai-artificial-intelligence-humans-technology-business-factory.html

5 Murgia, M. (2023) Transformers: the Google scientists who pioneered an AI revolution. Financial Times. https://www.ft.com/content/37bb01af-ee46-4483-982f-ef3921436a50

Langage : Les agents conversationnels sont capables de lire et d'écrire avec fluidité, que ce soit des phrases, des paragraphes ou même des articles entiers. Ils y arrivent en grande partie grâce aux *modèles transformateurs*, qui examinent les similitudes et différences entre les mots sous divers angles ou « dimensions ».

Par exemple, si nous ne considérions que deux dimensions, la *rondeur* et la *rougeur*, le modèle transformateur verrait une pomme et un camion de pompiers très éloignés en termes de rondeur mais proches en termes de rougeur, tandis qu'une balle de baseball serait proche de la pomme en termes de rondeur mais éloignée en termes de rougeur.

Les modèles transformateurs font des déductions en « explorant » ces différentes dimensions. Par exemple, en partant du mot « roi » et en se déplaçant dans la dimension « féminin », l'algorithme trouvera « reine ». S'il se dirige plutôt vers la dimension de la jeunesse, il pourrait aboutir à « prince ». En combinant ces deux dimensions, il pourra arriver à « princesse ».

Cela permet à l'IA de mieux prédire quels mots doivent suivre en s'appuyant sur le contexte global de la phrase ou du paragraphe. Par exemple, si vous écrivez « Frida a bu un chocolat », un algorithme de type autocomplétion pourrait automatiquement proposer « au lait » après « chocolat », car c'est ce qui apparaît le plus souvent dans ses données d'entraînement. Par contre, si vous demandez à un agent conversationnel : « Quel type de chocolat Frida a-t-elle bu? », le *modèle transformateur* pourrait repérer le mot « bu » et, en se concentrant sur « chocolat » dans le contexte d'un liquide, proposer « chaud » comme mot suivant.

Grand : Les agents conversationnels parviennent à imiter le langage humain et tenir des conversations grâce à l'immense volume de données sur lesquelles ils ont été formés et au nombre d'opérations (prédictions) qu'ils peuvent faire. Par exemple, un agent conversationnel très populaire a été formé sur un ensemble de données d'environ 500 milliards de mots. Chaque mot se voit attribuer une valeur dans jusqu'à 96 dimensions différentes (comme la « rougeur » ou la « rondeur »), et l'agent conversationnel effectue plus de 9 000 opérations à chaque fois qu'il devine un nouveau mot[6].

GÉNÉRATEURS DE MÉDIAS

Les outils d'IA qui créent des médias comme des images, des vidéos ou des voix fonctionnent de manière similaire aux IA conversationnelles, en étant formés sur des ensembles de données. En fait, beaucoup d'entre eux intègrent de grands modèles de langage : si vous faites la requête « représente une famille prenant le petit-déjeuner », l'image comprendra probablement des verres de jus d'orange, car le *modèle transformateur* comprend que le jus d'orange est souvent associé au petit-déjeuner.

Pour *créer* du contenu, ces IA utilisent un autre type de modèle, appelé *modèle de diffusion*.

Le principe est de partir d'images réelles, puis d'y ajouter progressivement du bruit, soit des modifications aléatoires, jusqu'à ce que l'image

6 Lee, T., & Trott S. (2023) Large language models explained with a minimum of math and jargon. *Understanding AI.* https://www.understandingai.org/p/large-language-models-explained-with

d'origine soit complètement brouillée. C'est ce qu'on appelle la *diffusion*.

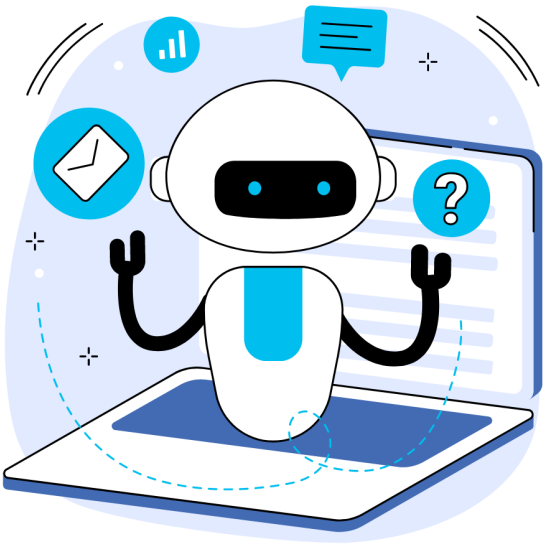
Ensuite, le modèle essaye d'inverser ce processus en testant des milliers, voire des millions de méthodes pour annuler le bruit et revenir à l'image d'origine. À chaque tentative, il compare son résultat à l'original et ajuste son processus en fonction du niveau de réussite. C'est ce qu'on appelle la *diffusion inverse*.

Une fois qu'il arrive à recréer parfaitement l'image initiale, le modèle dispose d'une « graine » – une base pour créer de nouvelles images similaires. En apprenant à débruiter ces images pour les ramener à l'original, le modèle apprend également à générer de nouvelles images dans le même style.

Ainsi, si vous lui demandez de créer une image d'orange, il puisera dans ces « graines » d'orange, les représentations d'oranges dans son ensemble de données, déjà passées par ce processus.

Enjeux liés à l'IA - Points d'attention

Il ne fait aucun doute que l'IA générative est un outil puissant qui aura des impacts majeurs sur notre quotidien, que ce soit à l'école, au travail ou à la maison. Elle apporte de nombreux avantages, mais comporte aussi des risques. Voici quelques domaines clés où l'IA a un impact.



LES INFORMATIONS

Les agents conversationnels peuvent être efficaces pour *réduire* la croyance aux théories du complot, en fournissant des informations fiables et des contre-arguments perçus comme venant d'une source objective[7]. Les gens sont aussi parfois plus aptes à repérer leurs propres préjugés lorsqu'ils sont reflétés par des IA entraînées à partir de leurs décisions[8]. Parallèlement, les outils d'IA générative peuvent parfois être utilisés pour produire du contenu volontairement trompeur, allant de sites Web et pages sur les réseaux sociaux avec de fausses informations et images pour attirer du trafic[9,10], à des théories du complot ou de la désinformation politique[11]. Ce type de contenu s'avère souvent très convaincant[12], surtout si

7 Costello, T. H., Pennycook, G., & Rand, D. G. (2024). Durably reducing conspiracy beliefs through dialogues with AI.

8 Celiktutan, B., Cadario, R., & Morewedge, C. K. (2024). People see more of their biases in algorithms. Proceedings of the National Academy of Sciences, 121(16), e2317602121.

9 Eastin, T., & Abraham S. (2024) The Digital Masquerade: Unmasking AI's Phantom Journalists. https://www.ajeastin.com/home/publications/digital-masquerade

10 DiResta, R., & Goldstein, J. A. (2024). How Spammers and Scammers Leverage AI-Generated Images on Facebook for Audience Growth. arXiv preprint arXiv:2403.12838.

11 Chopra, A., & Pigman A. (2024) Monsters, asteroids, vampires: AI conspiracies flood TikTok. Agence France Presse. https://www.france24.com/en/live-news/20240318-monsters-asteroids-vampires-ai-conspiracies-flood-tiktok

12 Spitale, G., Biller-Andorno, N., & Germani, F. (2023). AI model GPT-3 (dis) informs us better than humans. *Science Advances*,

des humains y apportent quelques améliorations[13]. Une étude de 2023 a révélé que plus de la moitié des gens pensent avoir vu du contenu trompeur généré par l'IA au cours des six derniers mois, et à peu près le même nombre n'était pas certain de pouvoir reconnaître de la désinformation créée par l'IA s'ils en voyaient[14]. Les agents conversationnels reproduisent aussi fréquemment des idées reçues, comme la croyance erronée selon laquelle les personnes noires auraient une peau plus épaisse que les personnes blanches[15].

Il est également important de faire attention aux « *hallucinations* ». Cela se produit lorsque le modèle invente des informations fausses ou inexactes. Par exemple, lorsqu'on demande aux agents conversationnels de fournir des références pour leurs réponses, ils inventent souvent des livres et des auteurs. Comme l'explique Subodha Kumar de l'université de Temple, « le grand public qui utilise [l'IA] aujourd'hui ne comprend pas vraiment son fonctionnement. Elle génère des liens et des références inexistants, car elle est conçue pour produire du contenu[16] ».

Un autre agent conversationnel a systématiquement donné des réponses erronées à des questions sur les processus électoraux[17]. Comme dans le cas de la désinformation intentionnelle, les gens sont susceptibles de croire ces hallucinations et fausses informations car les agents conversationnels ne montrent aucun doute ou incertitude[18]. Ils peuvent aussi fournir des informations exactes mais dangereuses ou inappropriées. Bien que la plupart d'entre eux disposent de « garde-fous » pour éviter cela, des recherches montrent que ces derniers sont imparfaits et relativement faciles à contourner. Par exemple, un agent conversationnel, par exemple, a expliqué à un utilisateur qu'il croyait être âgé de 15 ans comment masquer l'odeur de l'alcool[19].

Le plus grand risque n'est peut-être pas que les gens soient mal informés, mais plutôt qu'on en vienne à douter que quoi que ce soit soit réel[20]. À mesure que les images truquées deviennent plus sophistiquées, les signes révélateurs comme les traits asymétriques ou les doigts en trop disparaîtront, et il deviendra presque impossible de distinguer une image authentique d'une fausse, simplement en la regardant.

Zoom sur : les hypertrucages

Un *hypertrucage* est une image ou une vidéo d'une personne réelle créée de cette manière. Cela peut parfois être fait pour le divertissement, comme les « doubles numériques » d'acteurs dans les films, mais

9(26), eadh1850.

13 Goldstein, J. A., Chao, J., Grossman, S., Stamos, A., & Tomz, M. (2024). How persuasive is AI-generated propaganda?. PNAS nexus, 3(2), page034.

14 Maru Public Opinion. (2023) Media Literacy in the Age of AI. Canadian Journalism Foundation. https://cjf-fjc.ca/media-literacy-in-the-age-of-ai/

15 Omiye, J. A., Lester, J. C., Spichak, S., Rotemberg, V., & Daneshjou, R. (2023). Large language models propagate race-based medicine. NPJ Digital Medicine, 6(1), 195.

16 Chiu, J. (2023) ChatGPT is generating fake news stories — attributed to real journalists. I set out to separate fact from fiction. The Toronto Star.

17 Angwin, J., Nelson A. & Palta R. (2024) Seeking Reliable Election Information? Don't Trust AI. Proof News. https://www.proofnews.org/seeking-election-information-dont-trust-ai/

18 Kidd, C., & Birhane, A. (2023). How AI can distort human beliefs. Science, 380(6651), 1222-1223.

19 Pratt, N., Madhavan, R., & Weleff, J. (2024). Digital Dialogue—How Youth Are Interacting With Chatbots. JAMA Pediatrics.

20 Dance, W. (2023) Addressing Algorithms in Disinformation. Crest Security Review.

cela peut aussi causer de sérieux préjudices si la vidéo semble montrer quelqu'un dans une situation embarrassante ou choquante. Si les affaires les plus médiatisées impliquent des célébrités, l'utilisation la plus courante de la technologie des hypertrucages est la création de contenus pornographiques non consensuels, ciblant presque toujours des femmes. Ces contenus peuvent avoir des effets traumatisants pour les victimes, et le problème est amplifié par le fait que certains créateurs ou diffuseurs de ces hypertrucages pensent, à tort, qu'ils sont inoffensifs puisqu'ils « ne sont pas réels »[21] . D'autres, bien sûr, ont l'intention explicite de nuire à la personne dont l'image a été manipulée. Même si les hypertrucages de célébrités font le plus parler d'eux, des outils pour créer des hypertrucages pornographiques de n'importe qui sont désormais facilement accessibles[22].

« C'est super frustrant, parce que ce n'est pas toi, et tu veux que les gens te croient, et même s'ils savent que ce n'est pas toi, c'est quand même gênant... Je me sens humiliée. Mes parents sont humiliés. » — Victime de 16 ans d'un hypertrucage à caractère intime.

Les jeunes doivent comprendre que les hypertrucages à caractère intime ne sont pas sans conséquences et qu'ils causent du tort aux personnes représentées. Une des stratégies employées par des plateformes comme Meta pour limiter la propagation et l'impact des hypertrucages et autres images trompeuses générées par l'IA est l'ajout de *filigranes*[23]. Cela consiste à apposer une icône, une étiquette ou un motif pour signaler que l'image a été créée par une IA. Cependant, il n'existe pas encore de méthode de filigranage qui ne puissent être supprimées ou ajoutées à des images et des vidéos réelles pour les discréditer[24]. C'est pourquoi Sam Gregory, directeur général de l'organisation de défense des droits humains Witness, considère le filigranage comme « une réduction des risques » plutôt qu'une solution en soi[25].

BIAIS

Les générateurs d'images par IA sont davantage entraînés sur des banques d'images que sur des photos réelles, ce qui signifie que les images qu'ils produisent reflètent les choix conscients et inconscients des entreprises de banques d'images. En conséquence, ces algorithmes non seulement reproduisent les préjugés existants, mais peuvent même être encore plus biaisés que le monde réel.

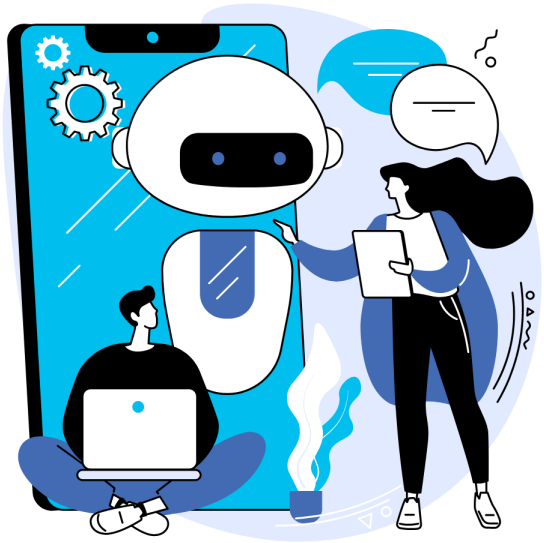
21 Ruiz, R. (2024) What to do if someone makes a deepfake of you. Mashable. https://mashable.com/article/ai-deepfake-porn-what-victims-can-do

22 Maiberg, E. (2024) 'IRL Fakes:' Where People Pay for AI-Generated Porn of Normal People. 404. https://www.404media.co/irl-fakes-where-people-pay-for-ai-generated-porn-of-normal-people/

23 Reuters. (2024) Facebook and Instagram to label digitally altered content 'made with AI'. *The Guardian*.

24 Saberi, M., Sadasivan, V. S., Rezaei, K., Kumar, A., Chegini, A., Wang, W., & Feizi, S. (2023). Robustness of ai-image detectors: Fundamental limits and practical attacks. *arXiv preprint arXiv:2310.00076*.

25 Kelly, M. (2023) Watermarks aren't the silver bullet for AI misinformation. *The Verge*. https://www.theverge.com/2023/10/31/23940626/artificial-intelligence-ai-digital-watermarks-biden-executive-order



Les images générées par l'IA peuvent reproduire les stéréotypes présents dans les images d'entraînement. Par exemple, certaines IA montrent presque uniquement des femmes[26] lorsqu'il s'agit de tâches ménagères, et si on leur demande une image d'« autochtone d'Amérique », elles représentent souvent des personnes portant des coiffes traditionnelles[27]. Même sans tomber dans ces stéréotypes évidents, l'IA générative a tendance à offrir une vision restreinte des groupes historiquement marginalisés[28].

Certaines recherches suggèrent toutefois que les biais dans les réponses de l'IA peuvent être atténués en diversifiant l'ensemble d'entraînement. Une étude a montré que l'ajout de seulement mille images supplémentaires (à un modèle en comptant plus de deux milliards) réduisait de manière significative le nombre de résultats stéréotypés ou inexacts[29].

INTÉGRITÉ ACADÉMIQUE

L'IA peut être utilisée de manière efficace et responsable en classe, que ce soit pour faire des retours aux élèves ou pour simuler des situations comme des entretiens d'embauche. Toutefois, il est essentiel que les jeunes comprennent les enjeux éthiques liés à son utilisation. Bien que les trois quarts des enseignants affirment que l'IA a un impact sur l'intégrité académique[30], les recherches montrent que son arrivée n'a pas entraîné une augmentation du plagiat[31]. Les élèves reconnaissent également que s'appuyer trop sur l'IA pourrait les empêcher d'acquérir des compétences importantes[32], et ceux qui l'utilisent fréquemment sont plus susceptibles de procrastiner[33]. Les raisons pour lesquelles les élèves utilisent l'IA pour tricher sont les mêmes que

26 Tiku, N., Schaul K. & Chen S.Y. (2023) AI generated images are biased, showing the world through stereotypes. *The Washington Post*.

27 Heikkilä, M. (2023) These new tools let you see for yourself how biased AI image models are. *MIT Technology Review*. https://www.technologyreview.com/2023/03/22/1070167/these-news-tool-let-you-see-for-yourself-how-biased-ai-image-models-are/

28 Rogers, R. (2024) Here's How Generative AI Depicts Queer People. *Wired*. https://www.wired.com/story/artificial-intelligence-lgbtq-representation-openai-sora/

29 Stokel-Walker, C. (2024) Showing AI just 1000 extra images reduced AI-generated stereotypes. *New Scientist*.

30 Robert, J. (2024) AI Landscape Study. EDUCAUSE. https://library.educause.edu/resources/2024/2/2024-educause-ai-landscape-study

31 Singer, N. (2023) Cheating Fears Over Chatbots Were Overblown, New Research Suggests. *The New York Times*.

32 Pratt, N., Madhavan, R., & Weleff, J. (2024). Digital Dialogue—How Youth Are Interacting With Chatbots. JAMA Pediatrics.

33 Abbas, M., Jam, F. A., & Khan, T. I. (2024). Is it harmful or helpful? Examining the causes and consequences of generative AI

celles identifiées dans les études précédentes sur le plagiat : le manque de temps ou une charge de travail académique trop lourde[34].

Malheureusement, les outils de détection des textes générés par l'IA échouent souvent à les repérer correctement et peuvent aussi attribuer à tort à l'IA des textes qui ne le sont pas[35]. Les jeunes qui n'écrivent pas dans leur langue maternelle sont particulièrement susceptibles de voir leur travail identifié à tort comme ayant été produit par l'IA[36]. Plutôt que de compter sur ces outils, les enseignants et les parents doivent apprendre aux élèves à utiliser l'IA de manière éthique et leur expliquer clairement les usages qui ne sont pas acceptables.

VIE PRIVÉE ET PARASOCIALITÉ

Les agents conversationnels sont souvent utilisés pour se divertir, mais ils peuvent aussi servir à fournir des retours (en jouant « l'avocat du diable » ou la « voix de la raison ») et aider à réduire le stress et l'anxiété[37]. Beaucoup de gens les trouvent utiles et réconfortants. S'ils sont conçus ou supervisés par des professionnels de la santé mentale, ils peuvent même être efficaces dans le cadre d'une thérapie, notamment pour les personnes qui sont moins enclines à consulter un thérapeute humain[38].

Il y a toutefois des risques que les agents conversationnels donnent des conseils erronés ou même dangereux[39], surtout s'ils n'ont pas été développés par des psychothérapeutes dans le cadre d'un programme de thérapie structuré. Même si les agents conversationnels ne peuvent pas ressentir d'empathie, les recherches montrent que nous avons tendance à les percevoir comme empathiques, surtout si nous sommes influencés pour le croire[40]. Les jeunes qui se tournent vers les agents conversationnels pour trouver de la compagnie peuvent développer des attentes irréalistes vis-à-vis des relations, ainsi que des idées fausses sur ce que leurs futurs partenaires attendront d'eux – et sur ce qu'ils attendront de leurs futurs partenaires[41].

Les informations que vous communiquez à un agent conversationnel peuvent être utilisées pour l'entraîner et, selon la politique de confidentialité de l'outil, être vendues à des courtiers en données, partagées avec les partenaires commerciaux de l'entreprise ou utilisées

usage among university students. International Journal of Educational Technology in Higher Education, 21(1), 10.

34 Abbas, M., Jam, F. A., & Khan, T. I. (2024). Is it harmful or helpful? Examining the causes and consequences of generative AI usage among university students. International Journal of Educational Technology in Higher Education, 21(1), 10.

35 Perkins, M., Roe, J., Vu, B. H., Postma, D., Hickerson, D., McGaughran, J., & Khuat, H. Q. (2024). GenAI Detection Tools, Adversarial Techniques and Implications for Inclusivity in Higher Education. *arXiv preprint arXiv:2403.19148*.

36 Liang, W., Yuksekgonul, M., Mao, Y., Wu, E. et Zou, J. (2023). Les détecteurs GPT sont biaisés contre les écrivains anglais non natifs. *Patterns*, 4(7).

37 Meng, J., & Dai, Y. (2021). Emotional support from AI chatbots: Should a supportive partner self-disclose or not?. *Journal of Computer-Mediated Communication*, 26(4), 207-222.

38 Habicht, J., Viswanathan, S., Carrington, B., Hauser, T. U., Harper, R., & Rollwage, M. (2024). Closing the accessibility gap to mental health treatment with a personalized self-referral Chatbot. *Nature Medicine,* 1-8.

39 Robb, A. (2024) 'He checks in on me more than my friends and family': can AI therapists do better than the real thing? *The Guardian*.

40 Pataranutaporn, P., Liu, R., Finn, E., & Maes, P. (2023). Influencing human–AI interaction by priming beliefs about AI can increase perceived trustworthiness, empathy and effectiveness. *Nature Machine Intelligence*, 5(10), 1076-1086.

41 Hinduja, S. (2024) Teens and AI: Virtual Girlfriend and Virtual Boyfriend Bots. Cyberbullying Research Center. https://cyberbullying.org/teens-ai-virtual-girlfriend-boyfriend-bots

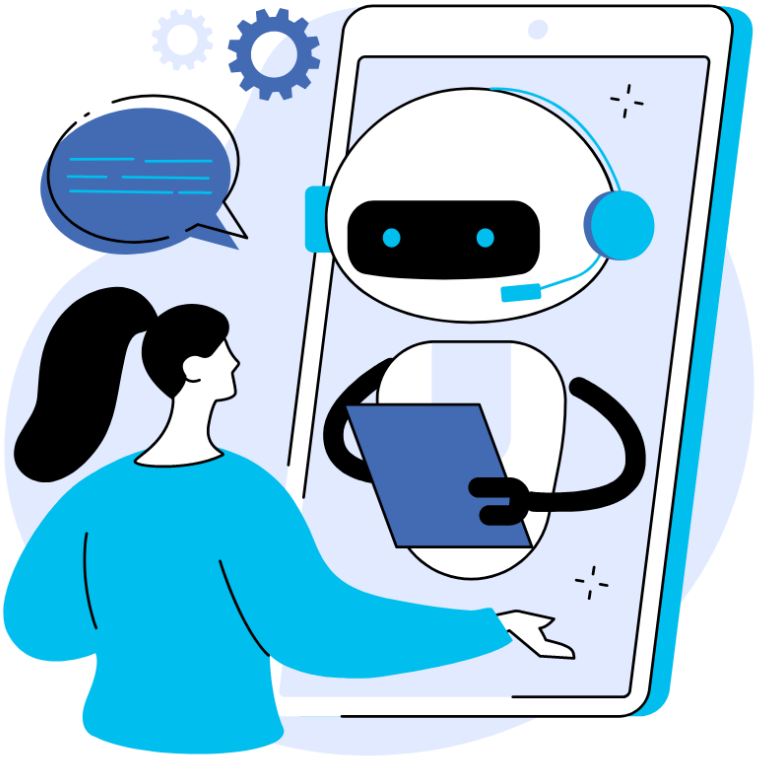
pour personnaliser vos fils d'actualité sur les réseaux sociaux et vous cibler avec des publicités. Même si les informations ne sont pas partagées ou utilisées, elles peuvent être exposées en cas de piratage de l'outil[42]. Les relations parasociales que nous développons avec les agents conversationnels peuvent nous rendre vulnérables, nous poussant à révéler plus d'informations que nous ne le ferions normalement – et l'agent conversationnel pourrait avoir été optimisé pour nous inciter à le faire, même sans l'intention directe de ses créateurs. De plus, comme ils sont formés à partir de données personnelles, telles que nos publications sur les réseaux sociaux ou nos recherches en ligne, et semblent déjà en savoir beaucoup sur nous, il existe un risque que nous jugions inutile de prendre des mesures pour protéger notre vie privée[43].

ET APRÈS?

Comme toutes les technologies, l'IA influence notre manière de l'utiliser, mais nous avons toujours le choix de l'utiliser de façon sûre et responsable. Que vous soyez enseignant, parent ou les deux, vous pouvez utiliser les informations contenues dans ce guide – ainsi que dans nos guides complémentaires *Aborder l'intelligence artificielle avec les enfants* et *Aborder l'intelligence artificielle en classe* – pour aider les jeunes à utiliser l'IA de manière positive, critique et responsable.

42 Caltrider, J., Rykov M. & MacDonald Z. (2024) Happy Valentine's Day! Romantic AI Chatbots Don't Have Your Privacy at Heart. Privacy Not Included. https://foundation.mozilla.org/en/privacynotincluded/articles/happy-valentines-day-romantic-ai-chatbots-dont-have-your-privacy-at-heart/

43 Caltrider, J., Rykov M. & MacDonald Z. (2024) Happy Valentine's Day! Romantic AI Chatbots Don't Have Your Privacy at Heart. Privacy Not Included. https://foundation.mozilla.org/en/privacynotincluded/articles/happy-valentines-day-romantic-ai-chatbots-dont-have-your-privacy-at-heart/



Avis de non-responsabilité : Meta apporte son soutien financier à HabiloMédias. Ce fiche-conseil a été élaboré conjointement par Meta et HabiloMédias. HabiloMédias ne recommande aucune entité commerciale, produit ou service. Ce guide n'a pas pour objectif de promouvoir Meta.