Revue de “Garantir une IA à notre service ? ou “ Dompter la Silicon Valley ”, de Gary Marcus, MIT Press

Philippe Dessus

Garantir une IA à notre service ? ou "Dompter la *Silicon Valley*"

Philippe Dessus

LaRAC, LIG, & Inspé, Univ. Grenoble Alpes, Grenoble, France

Philippe.Dessus@univ-grenoble-alpes.fr

Référence

Gary Marcus (2024). *Taming Silicon Valley: How we can ensure that AI works for us*. MIT Press, 240 p.

L'ouvrage dont il est question ici paraît à une date (mi-2024) tout à fait appropriée. Le battage autour de l'intelligence artificielle (IA) générative est un peu retombé, certaines de ses limites, depuis longtemps pointées dans des travaux spécialisés, sont maintenant diffusées dans le grand public, et le rôle de l'IA dans la désinformation sur les réseaux sociaux est bien documenté.

Cet assez court ouvrage, de lecture aisée pour des non-spécialistes du domaine, a pour but de faire le point sur les avancées de l'IA et de lancer des alertes sur ce qu'il faudrait faire pour que l'IA soit « *équitable, juste, et digne de confiance* » (p. 5). Son point de vue est à la première personne. Gary Marcus, professeur émérite de psychologie de l'université de New York, a pris une part non négligeable dans les débats sur l'IA fondée sur les réseaux de neurones artificiels. Il a été une des personnes qui ont montré (Marcus, 2018) les limites de l'approche informatique de l'apprentissage profond (*deep learning*).

L'ouvrage débute par une introduction qui brosse un état des lieux de l'impact social de l'IA, et où l'auteur reconnaît qu'à ce jour il est difficile de prédire si cet impact sera pour le meilleur ou pour le pire. Il énonce quatre problèmes majeurs qui seront considérés dans le reste de l'ouvrage : – l'IA générative est profondément biaisée ; – les entreprises développant des systèmes d'IA (générative ou autre) ne sont pas « responsables », c'est-à-dire qu'elles ne s'intéressent pas à leur impact social (*e.g.*, si ces systèmes peuvent servir à des buts criminels ou à propager des infox, ou *fake news*) ; – ces mêmes entreprises ont des intérêts à créer et maintenir une bulle financière euphorique à propos de l'IA générative, en ne mettant en avant que ses aspects positifs et en masquant ses problèmes, qui sont importants selon l'auteur ; – les États (à la notable exception de l'Europe) ne s'empressent pas à réguler cette industrie, lui laissant une grande marge de manœuvre, que ce soit pour lancer des systèmes à effets délétères

ou pour entraîner leurs systèmes avec des documents récupérés sans tenir compte des droits d'auteur.

L'ouvrage comprend ensuite trois grandes parties. La première est un point de vue sur les avancées actuelles de l'IA générative et ses principaux écueils. La deuxième détaille, à un niveau socio-économico-politique, le pouvoir de lobbying des principales entreprises d'IA, tant auprès du public que du gouvernement étatsunien. La troisième propose quelques solutions pratiques aux problèmes évoqués ci-avant. Un épilogue en forme d'appel à l'action clôt l'ouvrage.

L'argument principal de cette première partie est de montrer que, si certaines performances des systèmes d'IA générative sont dignes d'intérêt, on les a lancés prématurément. Le premier chapitre décrit justement quelques applications intéressantes de l'IA générative : générer un poème, une illustration, ou le premier jet d'un programme informatique. Mais ce n'est pas sans problèmes, que Marcus commence à lister. Personne, pas même ses concepteur·es, ne comprend comment l'IA générative fonctionne. Et ces systèmes sont entraînés *via* le recours intensif à des travailleur·es sous-payé·es (*e.g.*, Casilli, 2021). Le deuxième chapitre détaille d'un peu plus près les erreurs que ces systèmes génèrent, improprement nommées « hallucinations » (*e.g.*, des biographies de personnes comportant des éléments totalement inexacts) puisqu'une hallucination est un état altéré de conscience, ce qu'aucun système ne peut simuler. Comme ces erreurs sont inhérentes à la manière dont fonctionne l'apprentissage profond (*deep learning*) et les grands modèles de langage (*large language models*), Marcus estime qu'elles ne pourront jamais être évitées (voir aussi Marcus, 2018). Il indique aussi que les garde-fous mis en place, par exemple pour éviter que le système indique la procédure pour construire une bombe, sont aisément contournables (voir p. 44). Le troisième chapitre liste brièvement, mais de manière convaincante, les douze principaux risques que l'IA générative fait peser sur ses utilisateur·es et la société en les rendant plus faciles et moins détectables. Ces risques peuvent être regroupés en thèmes : désinformation et manipulation, diffamation et productions de fausses vidéos et enregistrements audio, erreurs et discriminations, non-respect de la vie privée et du droit d'auteur, problèmes de sécurité et enfin, coûts environnementaux, dus à la dépense énergétique et en eau de ces systèmes.

La deuxième partie de l'ouvrage s'intéresse aux buts économiques, rhétoriques et politiques des des GAFAM (*Google*, *Amazon*, *Facebook*, *Apple* et *Microsoft*), en montrant qu'ils sont devenus de plus en plus capitalistiques et monopolistiques au cours des années (ce que Zuboff, 2020 détaille plus précisément). Pour asseoir ces monopoles, il faut manipuler l'opinion par des annonces et des promesses les plus fracassantes possible, mais aussi minimiser les

inconvénients et risques des produits mis en avant. Même si elles ne sont pas suivies d'effets, ces annonces permettront à ces entreprises de maintenir un haut niveau boursier. Il ne faut bien sûr pas oublier le lobbying actif qu'elles mènent auprès des différents gouvernements (étatsunien et européen), en étant vent debout contre toute régulation de leurs produits. Un exemple cité (p. 103) concerne la France, où Cédric O défendait la régulation du secteur en tant que secrétaire d'état au numérique du gouvernement de Jean Castex, … pour défendre l'inverse quelques années après, cette fois en tant que lobbyiste et cofondateur d'une startup d'IA générative.

La troisième partie de l'ouvrage est une suite de douze courts chapitres proposant des solutions aux problèmes abordés précédemment. Nous ne les détaillerons pas faute de place, mais elles évoquent tour à tour la nécessité de préserver les droits des auteur·es des œuvres servant à l'entraînement des systèmes et les données personnelles de leurs utilisateur·es, d'aller vers plus de transparence, de responsabilité, vers une régulation indépendante, internationale et promouvant une IA raisonnée. Que les systèmes puissent fournir des outputs fiables est l'un des vœux les plus souvent exprimés par leurs utilisateur·es, et ils en sont loin, comme le montre l'un des derniers articles co-écrits par Marcus testant ces systèmes sur des tâches de compréhension (Dentella *et al*., 2024).

Le livre se clôt sur un épilogue plaidant pour des actions citoyennes pour amener les entreprises et les gouvernements à se positionner clairement sur les points évoqués dans l'ouvrage, et peut-être arriver un jour à « dompter la Silicon Valley »…

L'ouvrage de Marcus est remarquable car il résume clairement et en peu de pages les enjeux intellectuels, sociaux et politiques de l'IA, notamment générative. Il restera aux lecteurs et lectrices à réfléchir à ces enjeux dans un contexte éducatif, ce qui est rarement fait directement. En effet, le chapitre 11 sur la littératie de l'IA est plutôt rapide, et est évoqué en passant le risque que les humains « sortent de la boucle », si à la fois les élèves et les enseignant·es recourent à l'IA générative pour respectivement produire et évaluer les devoirs.

De mon point de vue, deux points en relation avec l'éducation auraient pu être approfondis dans l'ouvrage. Tout d'abord, le fait que « […] *la technologie de l'intelligence artificielle n'est pas magique* » (p. 23), contrairement à l'image que les entreprises qui les commercialisent veulent propager, et qu'il est nécessaire d'élaborer des moyens pédagogiques pour faire réfléchir les personnes utilisatrices. À ce sujet, le travail de Lupetti et Murray-Rust (2024) est exemplaire et prometteur, dans lequel ces chercheurs organisent des ateliers de design de produits utilisant l'IA où des étudiant·es passent successivement par une phase d'enchantement dans lequel le

produit est décrit avec des métaphores du champ sémantique de la magie, puis par une phase de désenchantement, où le fonctionnement du produit est détaillé et où l'on invite à la réflexion.

Le second point concerne la description d'applications éducatives et efficaces de l'IA générative. Même s'il est tôt pour en dresser une liste convaincante et scientifiquement prouvée, les travaux suivants sont également prometteurs : une application co-pilote aidant les tuteur·es humain·es à enseigner les mathématiques (Wang *et al*., 2024) ou un robot conversationnel pour débusquer des mythes conspirationnistes (Costello *et al*., 2024). Ce livre intéressera assurément toute personne (étudiant·e, enseignant·e, même non spécialiste de ces questions) voulant réfléchir aux enjeux éducatifs, sociaux et politiques de l'intelligence artificielle.

Références

Casilli, A. A. (2021). *En attendant les robots. Enquête sur le travail du clic*. Seuil.

Costello, T. H., Pennycook, G., & Rand, D. G. (2024). Durably reducing conspiracy beliefs through dialogues with AI. *ArXiv Preprint*. https://doi.org/10.31234/osf.io/xcwdn

Dentella, V., Gunther, F., Murphy, E., Marcus, G., & Leivada, E. (2024). Testing AI on language comprehension tasks reveals insensitivity to underlying meaning. *Scientific Reports*, *14*(1), 28083. https://doi.org/10.1038/s41598-024-79531-8

Lupetti, M. L., & Murray-Rust, D. (2024). (Un)making AI magic: A design taxonomy. *Proc. CHI Conf. on Human Factors in Computing Systems (CHI'24)*. http://dx.doi.org/10.1145/3613904.3641954

Marcus, G. (2018). Deep Learning: A critical appraisal. *ArXiv Preprint.* https://arxiv.org/abs/1801.00631

Wang, R. E., Ribeiro, A. T., Robinson, C. D., Loeb, S., & Demszky, D. (2024). Tutor CoPilot: A human-AI approach for scaling real-time expertise. *ArXiv Preprint*. https://arxiv.org/abs/2410.03017

Zuboff, S. (2020). *L'âge du capitalisme de surveillance* (B. Formentelli & A.-S. Homassel, Trad.). Zulma.