



© Liliya-AdobeStock

S. Societal Impact

L'éthique intellectuelle au temps de l'IA forte

ESCP Impact Paper No.2023-30-FR

Erwan LAMY
ESCP Business School

ESCP RESEARCH INSTITUTE OF MANAGEMENT (ERIM)



[L'éthique intellectuelle au temps de l'IA forte

Erwan Lamy*

ESCP Business School

Résumé

Avec les récents développements de l'intelligence artificielle de « Open AI », qui vient de rendre publique GPT-4, commence à poindre de manière plus crédible la perspective d'une IA forte, c'est-à-dire d'une IA dotée de capacités cognitives semblables à celles des humains (Bubeck et al. 2023). Si cela arrive, ce sera un second décentrement de l'humain, après la révolution copernicienne, et il faudra apprendre à vivre et à travailler avec d'autres intelligences que celles des humains. Cette perspective soulève évidemment des questionnements éthiques considérables, qui sont aujourd'hui abondamment discutés. Mais le volet épistémique de ces questionnements éthiques est souvent laissé dans l'ombre. De même que nous devons cultiver certaines vertus intellectuelles, comme l'ouverture d'esprit ou l'honnêteté intellectuelle, pour travailler dans un collectif organisationnel composé d'humains, il faudra également cultiver certaines vertus qui nous permettront de travailler avec les IA forte. Je propose de commencer à y réfléchir.

Mots clés (keywords): Intelligence artificielle, IA forte, GPT-4, éthique intellectuelle, vertu épistémique.

*Associate Professor, ESCP Business School

ESCP Impact Papers are in draft form. This paper is circulated for the purposes of comment and discussion only. Hence, it does not preclude simultaneous or subsequent publication elsewhere. ESCP Impact Papers are not refereed. The form and content of papers are the responsibility of individual authors. ESCP Business School does not bear any responsibility for views expressed in the articles. Copyright for the paper is held by the individual authors.

L'éthique intellectuelle au temps de l'IA forte

Une nouvelle révolution copernicienne?

*[The] new Philosophy calls all in doubt,
The Element of fire is quite put out;
The Sun is lost, and th'earth, and no man's wit
Can well direct him where to look for it.
And freely men confess that this world's spent,
When in the Planets, and the Firmament
They seek so many new; then see that this
Is crumbled out again to his Atomies.
Tis all in pieces, all coherence gone;
All just supply, and all Relation:
Prince, Subject, Father, Son, are things forgot,
For every man alone thinks he hath got
To be a Phoenix, and that then can be
None of that kind, of which he is, but he.*

Ces vers sombres, cités par Thomas Kuhn dans son étude de la révolution copernicienne (Kuhn 1957), sont de John Donne, un poète anglais de la fin du XV^{ème} siècle et du début du XVI^{ème}. Ils disent l'effet qu'a eu sur les esprits du temps l'idée que la Terre, et donc l'humain, n'était peut-être pas au centre du monde. Ce n'est pas seulement une position géographique qui fut déplacée, c'est tout un ordre du monde, incluant les hiérarchies mondaines ("Prince, Subject, Father, Son"), qui fut ébranlé.

Quatre siècles plus tard, nous commençons à prendre conscience que l'humain pourrait ne plus être au centre de la vie intellectuelle, et l'ébranlement pourrait être aussi considérable que celui provoqué par Copernic. Ce qui pourrait ainsi nous décentrer une seconde fois, ce sont les nouvelles générations d'intelligences artificielles (IA), révélées au monde il y a quelques mois à peine, et surtout leurs successeurs. Ce qui signale le trouble profond que provoquent déjà ces nouvelles machines, comme le trouble que provoquaient chez nos aïeux les nouvelles idées coperniciennes, ce sont les réactions qu'elles suscitent. On ne compte déjà plus les tribunes ou les discours voulant rappeler que les IA n'ont d'intelligence que le nom, et que l'esprit humain est incomparable. Jean-Gabriel Ganascia, président de l'Association française pour l'avancement des sciences et ancien pilote du comité d'éthique du CNRS, expliquait ainsi récemment que « ChatGPT est un perroquet qui répond de manière aléatoire » (2023). D'autres, dont je suis, rappellent que les IA ne savent pas distinguer le vrai du faux, et peuvent être comparées à des « générateurs de bullshit » (Lamy 2023).

Tous ces commentaires sont fondés, mais ne doivent pas faire oublier ce que ces technologies ont de potentiellement bouleversant. Il faut rappeler, premier point, qu'elles n'en sont qu'à leur début, et que ces débuts sont déjà spectaculaires. Second point, bien peu imaginaient il y a quelques mois que des machines pourraient produire des discours intelligibles, structurés, et que nous pourrions échanger avec elles comme nous échangeons avec un humain. C'est déjà fabuleux, et il serait imprudent de sous-estimer cette nouveauté. Enfin, plusieurs indices suggèrent que ces machines, avec toutes leurs limites, ouvrent de manière crédible la voie à une IA forte (Bubeck et al. 2023), ce qui était encore de l'ordre de la science-fiction il y a peu.

L'IA forte se caractérise par sa capacité à raisonner, apprendre et s'adapter de manière autonome, dépassant les limites de l'IA faible, qui ne peut accomplir que des tâches spécifiques. Il n'est pas clair que cette IA forte que l'on voit poindre soit réellement "intelligente", au même sens que l'intelligence humaine. Mais ce débat philosophique n'a sans doute pas beaucoup d'intérêt pratique, pas plus qu'il n'y en aurait à se demander si un moteur produit le même genre de "travail" qu'un humain. Ce qui importe, c'est qu'elle pourrait être capable de réaliser les mêmes tâches intellectuelles qu'un humain, et de manière parfaitement indistinguable.

Il faut s'y préparer. Une des dimensions trop souvent oubliée de cette préparation concerne l'éthique intellectuelle.

Une nouvelle éthique intellectuelle.

Voilà déjà plusieurs années qu'une très abondante littérature s'est développée autour des enjeux éthiques des intelligences artificielles. Plusieurs grands thèmes reviennent régulièrement dans les débats :

- Les biais et stéréotypes assimilés par l'IA au cours de son apprentissage peuvent perpétuer et amplifier des discriminations et des systèmes de domination existants.
- Le manque de transparence et d'explicabilité des IA rendent difficiles à justifier les décisions qui pourraient être prises avec leur aide, et complique la question de la responsabilité.
- L'utilisation généralisée des données personnelles dans l'IA soulève de graves problèmes concernant la vie privée et le risque d'une surveillance généralisée.
- Les IA peuvent être employées à des fins de manipulation, ou pour obtenir des consentements de manière déloyale.
- Les IA pourraient contribuer au renforcement du pouvoir monopolistique déjà considérable de quelques très grandes entreprises.
- L'utilité sociale d'une technologie dont les effets environnementaux risquent d'être importants soulève également des questionnements éthiques.

À ces sujets de préoccupation se sont ajoutés des questions plus directement liées aux toutes dernières générations d'IA, notamment chatGPT et GPT-4 :

- La possibilité d'employer les IA pour produire de fausses informations (ou de fausses photographies) pourrait considérablement aggraver des dérèglements épistémiques déjà inquiétants.
- La facilité d'emploi des IA pour produire des textes pourrait inciter les travailleurs ou futurs travailleurs du savoir (étudiants, chercheurs, journalistes, ...) à tricher.
- Ces nouvelles IA semblent menacer un spectre très large d'emplois, y compris parmi les plus qualifiés, et inquiètent donc pour leurs conséquences économiques inédites.

Enfin, et de manière beaucoup plus spéculative, sont discutées les inquiétudes, entretenues par le cinéma et la littérature, concernant de possibles rapports conflictuels entre les humains et les IA.

Dans tous ces débats, il est un sujet qui n'est jamais abordé : celui de la transformation de l'éthique intellectuelle au contact des IA, et notamment des IA fortes.

Le travail intellectuel ne consiste pas seulement en l'application de certaines méthodes, il est guidé par des règles morales qui doivent être incorporées. Ne pas dire la vérité, ou se

refuser à répondre à une critique, ce ne sont pas seulement des méconduites aux conséquences potentiellement dommageables, ce sont des fautes blâmables. L'éthique intellectuelle consiste à cultiver certaines vertus, et à combattre certains vices, qui contribuent à la poursuite d'une saine vie intellectuelle. Ces vertus peuvent être, par exemple, l'honnêteté intellectuelle, la rigueur, l'ouverture d'esprit, la responsabilité épistémique (Lamy 2022), etc.

Ces vertus sont essentielles au bon fonctionnement des organisations. Elles contribuent à assurer la fiabilité des informations qui y circulent, et donc facilitent la prise de décision éclairée. Elles soutiennent des formes de collaboration et de communication efficaces au sein de l'organisation, en y encourageant chacun à partager des informations fiables et sincères. Enfin, elles contribuent au renforcement de la réputation et de la crédibilité des organisations. À l'inverse, les vices épistémiques peuvent avoir des conséquences catastrophiques. Boudewijn de Bruin (2015) a ainsi montré que les comportements épistémiquement vicieux des institutions financières ont pu mener à la crise financière de 2008.

Cette question des vertus et des vices épistémiques commence à retenir l'attention depuis quelques années, à partir notamment des travaux séminaux de de Bruin (2013, 2015). Avec la venue des IA fortes, elle doit être abordée sous un nouveau jour : que deviendront ces vertus lorsque les IA fortes entreranno en scène ?

C'est un exercice de prospective que je propose ici. Il ne s'agit pas de réfléchir aux vertus épistémiques qu'il faut cultiver aujourd'hui pour travailler avec chatGPT ou GPT-4. Elles commencent à être déjà connues. La venue de ces nouvelles IA implique notamment de cultiver une forme de « sensibilité numérique » qui consisterait à « redoubler d'attention à la lecture d'un texte en ligne [...] sans pour autant tomber dans une forme de scepticisme généralisé » (Lamy 2023). Ces IA sont encore très défectueuses, et produisent beaucoup de faussetés avec toute l'apparence de la vérité. La situation serait très différente avec des IA fortes. Il faut envisager qu'elles n'aient plus ses faiblesses, et que leurs productions intellectuelles soient d'une valeur au moins égale à celles des humains.

Avec de telles machines, l'éthique intellectuelle ne commanderait plus une certaine méfiance. Elle commanderait au contraire d'être capable de passer ce moment d'incrédulité qui va accompagner l'avènement de notre décentrement intellectuel, cette seconde révolution copernicienne, pour accepter la venue d'intelligences non humaines.

Cela impliquerait déjà une forme aiguë d'ouverture d'esprit. L'ouverture d'esprit est cette disposition intellectuelle qui nous conduit à tenir pour digne d'intérêt ce qui peut avoir les apparences de l'étrangeté. Être ouvert d'esprit, c'est accepter d'être dérangé dans ses habitudes intellectuelles pour s'ouvrir de nouveaux horizons épistémiques. C'est une vertu déjà importante pour travailler en équipe. Avec les IA fortes, elle sera sans doute cruciale. Ces entités non humaines suivront sans doute des raisonnements non humains, et les résultats de ces raisonnements pourront paraître tout à fait inhabituels. Il faudra être prêt à les entendre. Cela pourrait passer par des remises en cause radicales de nos certitudes les plus ancrées.

Cela impliquerait également l'acceptation de ce second décentrement, c'est-à-dire une forme nouvelle d'humilité intellectuelle. Cette vertu, qui bien sûr ne pourrait valoir que le temps de nous habituer à ce décentrement, consisterait à reconnaître nos limites en tant qu'espèce (et non plus seulement en tant qu'individu) et à accepter la perte du monopole de l'intelligence. Face à une IA forte, cette humilité nous permettrait d'accepter que ces entités puissent détenir des connaissances ou des compétences intellectuelles égales ou

supérieures aux nôtres, qu'elles soient capables d'une véritable vie intellectuelle aussi importante et digne que la nôtre, quoi que différente. Parce qu'il semble difficile d'avoir un échange fertile d'idées et d'expertise avec des êtres que l'on tient pour inférieur, voire auxquels on dénie toute forme de vie intellectuelle, cette vertu serait la condition d'une véritable collaboration avec les IA fortes. Cette humilité intellectuelle serait une sorte de lâcher prise lucide. « Lucide, » car il faudra aussi comprendre que leurs facultés intellectuelles vont de pair avec la faculté de se tromper. Ce ne sont plus des ordinateurs, qui ne peuvent guère commettre que les erreurs de leurs programmeurs ou de leurs créateurs.

Enfin, on peut envisager une troisième vertu intellectuelle pour travailler avec les IA fortes : la capacité de synthèse et le leadership épistémique. Il faudra être capable de diriger des IA plus intelligentes que nous-mêmes, comme un chef d'entreprise dirige les efforts cognitifs combinés de collaborateurs et de collaboratrices qui peuvent également être plus intelligents qu'elle. Les dirigeants ne sont pas des experts, mais ils doivent être capables de diriger des experts. On imagine aujourd'hui que l'apparition d'une intelligence supérieure serait nécessairement source de conflit. Cela revient à faire de la hiérarchie intellectuelle quelque chose de nécessairement conflictuel. Ce n'est pas vrai. Cette hiérarchie peut simplement induire une forme de division du travail qui ne se réduit pas à une relation de domination. Mais la vertu du leadership épistémique ne se limite pas à la compréhension de cette réalité. Elle consiste en une capacité à stimuler les intelligences, à les diriger sans les écraser, à les conduire en respectant leur autonomie. C'est aussi la capacité à récupérer les fruits de ces intelligences en en faisant la synthèse. Travailler avec les IA fortes n'impliquera pas de comprendre le détail de leurs raisonnements, mais de savoir les synthétiser.

Conclusion

Il est clair que la nouvelle éthique intellectuelle qui pourrait venir avec l'avènement des IA fortes ne s'arrête pas à ces trois vertus. L'ambition de ce court texte n'était évidemment pas de livrer une liste exhaustive des vertus épistémiques qu'il faudrait cultiver pour travailler avec les IA fortes à venir. Elle était de montrer le genre de réflexion qu'il faudra bientôt mener dans chaque organisation. Bien sûr, ce n'est pas encore fait. Les IA fortes ne sont pas encore là. Mais leur venue se fait chaque jour plus crédible, et il faut s'y préparer. Cette préparation ne doit pas se limiter pas aux discussions techniques ou aux réflexions éthiques habituelles. Elle doit engager tout un volet épistémologique, avec le problème de l'éthique intellectuelle.

Références

Bubeck, S. et al. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. arXiv preprint arXiv:2303.12712.

De Bruin, B. (2013). Epistemic virtues in business. *Journal of business ethics*, 113(4), 583-595.

De Bruin, B. (2015). *Ethics and the global financial crisis*. Cambridge University Press.

Ganascia J.-C. (2023). ChatGPT est un perroquet qui répond de manière aléatoire. *Le Point*, 29 janvier 2023.

Kuhn, T. S. (1957). *The Copernican revolution: Planetary astronomy in the development of Western thought* (Vol. 16). Harvard University Press.

Lamy, E. (2022). Epistemic responsibility in business: An integrative framework for an epistemic ethics. *Journal of Business Ethics*, 183(1), 1-14.

Lamy, E. (2023). ChatGPT nous rendra-t-il moins crédules ? The Conversation, 26 janvier 2023.