



© Liliya-AdobeStock

S. Societal Impact

Intellectual ethics in an era of strong AI

ESCP Impact Paper No.2023-20-EN

Erwan LAMY
ESCP Business School

[Intellectual ethics in an era of strong AI

Erwan Lamy*

ESCP Business School

Abstract

With the recent developments in artificial intelligence made by Open AI, which has just made GPT-4 public, the prospect of strong AI, meaning AI with cognitive abilities similar to those of humans, is beginning to appear more credible (Bubeck et al. 2023). If this happens, it would be the second time, after the Copernican revolution, that humans found themselves decentred, and we would need to learn to live and work with non-human intelligences. This perspective obviously raises significant ethical questions that are being discussed at length, but the epistemic aspect of these ethical questions is often overlooked. Just as we need to nurture certain intellectual virtues, such as open-mindedness and intellectual honesty, to be able to work in groups of humans, we must also cultivate certain virtues that will allow us to work with strong AIs. I suggest we begin to reflect on this.

Keywords: Artificial intelligence, strong AI, GPT-4, intellectual ethics, epistemic virtue.

*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.

Intellectual ethics in an era of strong AI

A new Copernican revolution?

*[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.*

These bleak verses, quoted by Thomas Kuhn in his study of the Copernican revolution (Kuhn 1957), were written by John Donne, an English poet from the late 15th and early 16th centuries. They speak of the effect on contemporary minds of the idea that the Earth, and therefore human beings, might not be at the centre of the universe. It was not simply a geographical position that was displaced, but an entire world order, including worldly hierarchies ("Prince, Subject, Father, Son"), that was undermined.

Four centuries later, we are beginning to become aware that humans may no longer be the focus of intellectual life, and the disruption could be just as significant as the one caused by Copernicus. What could decentre us for a second time are the new generations of artificial intelligence (AI), those that were revealed to the world just months ago and, more importantly, their successors. What demonstrates the profound unease arising from these new machines, in the same way that our ancestors were uneasy about the new Copernican ideas, are the reactions they are causing. We have lost count of the number of editorials and speeches trying to remind us that AIs are intelligent in name only, and that the human mind is incomparable. Jean-Gabriel Ganascia, President of the Association Française pour l'Avancement des Sciences and former head of the CNRS Ethics Committee, recently explained that "ChatGPT is a parrot that responds randomly" (2023). Others, including myself, point out that AIs cannot distinguish truth from falsehood and can be likened to "bullshit generators" (Lamy 2023).

All these comments are well-founded, but they should not detract from these technologies' potentially disruptive nature. The first point to bear in mind is that this is only the beginning, and the beginning is already spectacular. The second point is that, just a few months ago, few people imagined that machines could produce cogent, structured speech and that we could converse with them as we do with other humans. Already this is fabulous, and we would be unwise to underestimate it. Finally, several indicators suggest that these machines, with all their limitations, are credibly paving the way for strong AI (Bubeck et al. 2023), which seemed to be mere science fiction not so long ago.

Strong AI is characterised by its ability to reason, learn and adapt on its own, exceeding the limitations of weak AI, which can only carry out specific tasks. It is not clear whether this emerging strong AI is truly "intelligent" in the same way as human intelligence. But this philosophical debate is undoubtedly of little practical interest, no more than asking ourselves if an engine performs the same sort of "work" as a human. What is important is that it might be capable of performing the same intellectual tasks as a human in a way that is perfectly indistinguishable.

We need to be prepared, and one of the most forgotten aspects of this preparation concerns intellectual ethics.

A new intellectual ethics

For several years now, an abundance of literature has been written on the ethical issues surrounding artificial intelligence. A number of major themes regularly feature in these debates:

- The biases and stereotypes that AI absorbs as it learns can perpetuate and amplify existing discrimination and systems of power.
- AI's lack of transparency and explicability makes it difficult to justify any decisions that might be taken with its help and complicates the issue of accountability.
- The widespread use of personal data in AI raises serious problems about privacy and the risk of generalised surveillance.
- AIs can be used for manipulation or to obtain consent in a bad-faith way.
- AI can help to reinforce the already significant monopolistic power of a few large companies.
- The social benefits of a technology whose environmental effects could be significant also raises ethical questions.

We could also add concerns that are more directly related to the latest generation of AI, especially ChatGPT and GPT-4:

- The ability to use AI to produce fake news or images could considerably exacerbate already worrying epistemic disturbances.
- The ease of using AI to produce content could encourage current or future knowledge workers, such as students, researchers and journalists, to cheat.
- These new AIs appear to threaten a wide spectrum of even the most highly qualified jobs and raise concerns for these AIs' unprecedented economic consequences.

Finally, and in a much more speculative way, there is discussion of the concerns fuelled by film and literature about the potential for conflict between humans and AIs.

In all these debates, however, there is one topic that is never discussed: how intellectual ethics will be transformed by contact with AI, especially strong AI.

Intellectual work does not just consist in applying certain methods; it is guided by moral rules that must be incorporated. Not telling the truth or refusing to respond to criticism are not just misconduct with potentially damaging consequences, they are culpable faults. Intellectual ethics consists in cultivating certain virtues and combatting certain vices in a way that contributes to a healthy intellectual life. Some examples of these virtues can be intellectual honesty, rigour, open-mindedness and epistemic responsibility (Lamy 2022).

These virtues are crucial to smooth-running organisations. They help to ensure that the information that circulates within them can be trusted to guide enlightened decision-making. They support efficient forms of collaboration and communication within the organisation by encouraging everyone to share reliable, sincere information. Finally, they support organisations' reputation and credibility. Conversely, epistemic vices can have catastrophic consequences. Boudewijn de Bruin (2015) has shown that financial institutions' epistemically perverse behaviour may have led to the 2008 financial crisis.

This issue of epistemic vices and virtues has begun to attract attention in recent years, starting with de Bruin's seminal work (2013, 2015). With the arrival of strong AIs, these questions need to be approached in a new light: what will become of these virtues when strong AIs take the stage?

What I am proposing here is an exercise in foresight. The point is not to think about the epistemic virtues that need to be cultivated today in order to work with ChatGPT or GPT-4. These are already beginning to emerge. The advent of these new AIs involves fostering a form of "digital sensitivity", which would consist of "paying extra attention when reading an online text...without falling into a form of generalised scepticism" (Lamy 2023). These AIs still have many flaws and can produce a number of false information that can appear to be true. The situation could be very different with strong AIs. We need to imagine that they no longer have these flaws and that their intellectual output will be at least as valuable as that of humans.

With such machines, intellectual ethics would no longer demand a certain mistrust. It would, however, demand that we should be able to overcome this moment of scepticism that will accompany this second Copernican revolution that is our intellectual decentring to accept the arrival of non-human intelligence.

This would already entail an acute form of open-mindedness. Open-mindedness is the intellectual attitude that leads us to consider what may seem strange as worthy of interest. Open-mindedness means accepting having one's intellectual habits disrupted in order to broaden one's epistemic horizons. This virtue is already important when working as a team. With strong AI, it will no doubt be crucial. These non-human entities will undoubtedly follow non-human lines of reasoning, and the results of this reasoning may appear quite unusual. We need to be ready to listen to them. This could cause us to radically question our most deeply rooted certainties.

This would also imply accepting this second decentring, a new form of intellectual humility. This virtue, which of course would only be valid for as long as it takes us to get used to this decentring, would involve recognising our limits as a species (and no longer just as individuals) and accepting the loss of our monopoly on intelligence. In the face of a strong AI, this humility would allow us to accept that these entities may possess knowledge or intellectual skills equal to or superior to our own, that they may be capable of a genuine intellectual life as important and worthy as ours, albeit different. Since it seems difficult to have a fruitful exchange of ideas and expertise with beings that are considered inferior, or to whom we even refuse to acknowledge any form of intellectual life, this virtue would be the prerequisite for any true collaboration with strong AIs. This intellectual humility would be a kind of lucid letting go. Lucid, because we will also need to understand that their intellectual abilities come with the ability to be wrong. They are no longer computers who can only make the mistakes of their programmers or creators.

Finally, we can imagine a third intellectual virtue for working with strong AI: the ability to summarise and the capability of epistemic leadership. We will need to be able to direct AIs

who are smarter than we are, like a business manager directs the combined collective efforts of their employees, who may also be more intelligent than they are. Managers are not experts, but they need to be able to direct experts. Today, we are imagining that the emergence of a superior intelligence would necessarily create conflict. This turns the intellectual hierarchy into something that is necessarily conflictual. This is not true. This hierarchy may simply entail a form of division of labour that cannot be reduced to a relationship of domination. But the virtue of epistemic leadership is not limited to understanding this reality. It consists of an ability to stimulate intelligences, directing them without trampling them, leading them while respecting their autonomy. It also entails the ability to harvest the fruits of these intelligences by summarising their work. Working with strong AIs does not require understanding their reasoning in detail, but being able to summarise.

Conclusion

Clearly, the new intellectual ethics that may accompany the rise of strong AIs is not limited to these three virtues. The aim of this short text is obviously not to offer an exhaustive list of epistemic virtues that need to be cultivated in order to work with future strong AIs. It exists to demonstrate the type of thinking that will soon need to take place within each organisation. Of course, this has not happened yet. Strong AIs are not here yet. But their arrival is becoming more credible each day, and we need to be prepared. This preparation must not be limited to the usual technical discussions or ethical reflections. It must involve a whole epistemological aspect with the problem of intellectual ethics.

References

- 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 January 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 January 2023.