The Ethics of OpenAI/ChatGPT

Jacques Rousseau, University of Cape Town, South Africa

The European Conference on Education 2023
Official Conference Proceedings

Abstract

The OpenAI Playground and ChatGPT use GPT-3.5 to produce text using an AI language model that is capable of routinely producing texts that would appear to have been written by humans at a level of sophistication that would meet typical benchmarks for competence in those fields. Policy responses at universities currently speak to the capacity these tools have at present. But AI models for text-generation will keep improving, resulting in an arms race that educators cannot win. A further concern for many educators is that students who have greater familiarity with computers and the Internet might better be able to exploit these tools in formulating “better” generative commands, which would in turn further exacerbate the "digital divide" between students with historical advantages compared to others. While some universities are responding by increasing the number of assignments written in class or oral examinations, these potential solutions cannot be implemented in large classes, such as those with an enrolment of 900+ students as are common in South African universities. The range and severity of the possible consequences of OpenAI (and related tools) for teaching, learning and research is significant enough to merit reflection and response at the highest levels of decision-making, and this paper will offer reflections on possible responses to this challenge.

Keywords: ChatGPT, Plagiarism, Digital Divide, Assessment, Research Integrity

Introduction

Concerns about artificial intelligence (AI) tools such as ChatGPT and their possible impact on education are of course simply new variants or evolutions of prior concerns, rather than new ones. In popular culture, these concerns have been depicted in cinema, perhaps most famously with the Terminator movies, while in the academic literature notable examples include the articulation of problems such as the “Paperclip maximiser”.

If you don’t know about the Paperclip maximizer, it is a thought experiment described by Swedish philosopher Nick Bostrom (2003). It highlights the possibility that an AI can be programmed to pursue goals that seem harmless – like making paperclips – and that it could take that job so seriously, and so literally, that it threatens human society or even human existence.

Given enough power, it might try to turn everything into paperclips, or into machines that can help it produce paperclips. Humans might decide to turn it off, so they are already a clear threat to the AI’s mission. There is of course iron in our blood, which could usefully be harvested. To cut a terrifying story short, such an AI would need to somehow be programmed to value human life too.

Bostrom highlights the possible threats unintended consequences carry, and therefore, the need to try to anticipate future problems, and to establish mitigation strategies in advance.

One variant or evolution of AI challenge introduced with Generative Artificial Intelligence tools like the OpenAI Playground, ChatGPT, Google’s Bard or Microsoft Bing is that they are freely accessible; have generalized application to activities that most of us could benefit from; and that they are getting better rapidly and frequently.

In summary, they are already capable of routinely producing outputs that would appear to have been created by humans, at a level of sophistication that would frequently evade detection, and that would also meet typical benchmarks for competence in those fields, even if not excellence.

1. Ethical Challenges Posed by AI in Education

Because generative AI models will keep improving, and because tools for detecting non-human-generated output are currently all imperfect, we will always be playing catch-up - this is an arms race that we cannot win.

So, the response to ethical challenges cannot be premised on control – they must be built from the ground up to focus on responsible, appropriate and productive use of these tools. This requires a long-term commitment and vision, as you would need to invest time in policy-shaping; and on education about long-term possible problems you want your staff or your students to be aware of, so that the threats can be mitigated by informed and ethically-aware stakeholders.

There is a danger in any response that suggests “we can think about this later”. An apparently useful tool, once embedded in organizations at scale and in a casual or unplanned manner will be very difficult to unwind, because hypothetical risks might seem trivial when compared to obvious benefits. Consider the case of the Chinese “social credit” system (Feng, 2022), where all of your activity is tracked, and then used to categorize you as meriting a loan at preferential or prejudicial interest rates – or even as eligible for a loan or not. While this provides enormous possibility for efficiency, it also threatens to diminish the moral value of treating people as sentient beings, rather than objects in a database.

However, what if these sorts of determinations can take place in the background, based on the data from all of our creative or work outputs; alongside responses from course evaluations; citation metrics or grant funding success; tracking "productivity" based on a metric defined by bureaucrats; which is all then assessed by a Paperclip maximiser?

This outcome would likely be attractive to the corporate part of a university, but staff would quite quickly find themselves being assessed in completely opaque ways. What that would do to our incentive structures and our relationships with each other is largely unpredictable, except for the prediction that the consequences are not likely to be good. What follows articulates a selection of the concerns we should be attentive to.

1.1 Inhumane Treatment and Dependence

We do not have to wait for the pervasive presence of advanced AI to see the some consequences of passing responsibility away from humans. Consider existing automated (or at least unstaffed) ticketing stations or helplines. Ordinarily they work well, perhaps providing services in a wider range of languages more quickly than an human agent could. But when something goes wrong, there might be no one to help make alternative arrangements, or even to apologize.

Scale that automation up to a point where a significant proportion of a strategic and other decisions in a university are either themselves generated by AI or made on the basis of reports generated by AI, but where we simultaneously have a shortage of people who are able to strategize for – or even explain – complex issues, because their ability to do so has become atrophied through disuse.

While it used to be the case that AI was expensive, genuine human interaction and deliberation may well become the more scarce and expensive good. We have serious challenges in how we think about responsibility and accountability for decisions made by AI systems, not least in terms of its impact on trust and respect amongst humans.

That impact includes possibilities ranging from the devaluation of a range of tasks such as enrollment or curriculum planning, to full disrespect for those that perform those tasks – which we can imagine expressed in dehumanizing phrases like "an AI could do your job". This effect could be seen at all levels of authority, in that if stakeholders know or suspect that you make decisions based on inputs *provided to you* by people who have generated them via AI, and where you perhaps did the same, it becomes increasingly difficult to trust that you are doing your job – or perhaps more importantly, whether they need you at all.

Social and character skills will start to matter more. When jobs are perishable, and technologies come and go while people's working lives become ever-longer, social skills are a foundation that can give humans a comparative advantage, as they could help them do work that calls for empathy and human interaction—traits that are (at least currently) beyond machines.

1.2 Bias in Data and Algorithms

This technology is only as unbiased as the data it is trained on. If the data used to train the AI system are biased, the system will produce biased results. Organizations must ensure that they are using diverse and representative data sets to train their AI systems, and remain alert to the possibility of bias in their data as well as the decisions that flow from it.

In education, subtle differences in competencies based on culture, language and worldview require sensitive and experienced educators to make a range of choices that are often invisible to those making them, because they might have been making them for decades – can an algorithm do as well at this task? Excellent students might be undetectable in a dataset, yet clearly apparent to you when personally engaging with their commitment to learning.

1.3 Privacy, Security and Data Protection

The technology relies on large amounts of data to function effectively, so organizations must ensure that they are collecting and using these data in an ethical and responsible manner. They must also consider the potential risks of data breaches and take steps to protect sensitive information.

1.4 Accountability and Responsibility

Finally, AI raises important questions about accountability and responsibility. As organizations adopt this technology, they must ensure that they are taking responsibility for its actions and their consequences. They must be transparent about how the technology is being used and be prepared to take responsibility for any negative consequences that may arise.

The tendency to anthropomorphizing AI must be resisted – legal and moral responsibility for AI outputs should be no different to the responsibility we assign to other software tools, and ultimate responsibility must lie in the decisions and authority that created the environment they operate in, and those people who chose to deploy the AI tools in question.

2. Responses to Potential Ethical Concerns

So what do we do? Best practices for ethical implementation of OpenAI in organizations would seem to include, at a minimum, transparency and explicability of AI systems; the involvement of diverse stakeholders in the development and deployment of AI systems; regular monitoring and evaluation of AI systems for potential biases and ethical concerns; and the development of clear guidelines and policies for the use of AI systems.

But any response premised on a binary choice of rejecting or using AI should adapt to recognize that it’s likely that – at least for the moment – it’s not AI or humans that will be most efficient, but rather a combination of the two. Therefore, our policy, HR, PR, and other responses to the emergence of these new challenges need to incentivize humans to use the tools more effectively, rather than be made to feel worthless or replaceable by those tools. This is not only because we still need careful and creative judgment in order to implement the output of those tools judiciously, but also because control that is ceded is much more difficult to regain.

2.1 Complicating our Responses: The State of Humans

In “The Enigma of Reason” (2017), cognitive scientists Hugo Mercier and Dan Sperber argue that reason is an evolved trait. Their argument in summary is that the primary advantage humans have over other species is our ability to cooperate, and that the tools used by humans in reasoning and argumentation were not developed to solve logical problems so much as to resolve problems that arise when living in collaborative (and competitive) environments.

While their analysis is compelling, it needs to be understood in light of various confounding factors, ranging from some we’ve been aware of for decades, such as confirmation bias (the tendency people have to embrace information that supports their beliefs, and to reject information that contradicts those beliefs) (Shermer, 2002, p.145), to more recent concerns such as misinformation and disinformation (including “fake news”, to use a term that gained traction in the time of U.S. President Donald Trump).

What these confounders illustrate is that if reasoning were intended to generate sound judgments, rather than to serve as a mechanism for social collaboration and improving one’s perceived standing in society, it would be difficult to imagine more serious impediments to achieving rational outcomes than confirmation bias and the prevalence of unreliable source information (via mis/disinformation).

The asymmetry described above reflects the task that reason evolved to perform, which is to optimize our existence within the context of existing in a group, whether that be a local community or an international community of scholars. For our purposes in this paper, though, it highlights something else, which is that analysis of evidence and the development of arguments drawn from AI-generated sources *removes* us from the collaborative sort of meaning-making described here, and thus could be said to be contrary to the purpose our reasoning tools and strategies evolved to serve.

One way to look at socially-engaged reasoning is thus as a system that partly corrects for our natural inclinations to stubbornly hold on to untested claims, or more generously, to be reluctant to see options besides the ones we are already familiar or comfortable with. The clearest example of this is in scientific disciplines – in an environment where empirical data are respected above all else, such as a laboratory, there’s very little room for confirmation bias or other mistakes attributable to subjectivity or misinformation.

And, while it would perhaps be counterproductive to social engagement, and also antithetical to the many fields of education that are not strictly empirical, this does perhaps point to a general lesson that our best reasoning – or at least the best outcomes of our reasoning – are the product of people engaging with each other in debate and deliberation, all committed to reaching the most justified conclusion that they can, under their particular circumstances.

3. The World, and Education, in an AI Future

What do you teach university students in a post-Google world? Within minutes, anyone with an internet connection can acquire basic knowledge about any topic, while simultaneously being connected to a community that will now reinforce any given belief as well as the value (sometimes even *virtue*) of holding that belief. Elsewhere, I have described this as *contextual rationality* (Rousseau, 2021), in that because one’s context might involve pre-filtering of evidence, and prior selection of which conclusions are desirable, we might believe ourselves to be thinking in fully rational ways, even as we are woefully uninformed, confused, or both.

This is a clear threat to subject specialists, and to universities – especially in relation to technical qualifications. But it also highlights a difference between those technical qualifications and the humanities and social science disciplines, which deal with big, abstract ideas and not just facts. The humanities are perhaps more relevant than ever, as we engage with the "fourth industrial revolution" and uncertainties regarding the role AI will play.

Specialist degrees – such as those offering technical training in subjects like accounting – run a particular risk. Students commit 3, 4 or 5 years to study, and then emerge into a job market that's quite different from what they expected. The job they were trained for might no longer exist; the skills required may have changed; or more likely, it would now be a job that is performed more quickly and competently by an AI.

An employer who is committed to the long-term sustainability of their enterprise should equally be aware of the importance of hiring people who can solve problems, rather than simply those who have technical skills, because so many technical skills will soon be better outsourced to AI.

So how does all this help prepare students for the fourth industrial revolution? First, in offering the reminder that specialists are often replaceable, or will soon be. Answers to difficult problems frequently emerge out of collaboration and debate amongst people who – while they might have a specialization – are also conversant with multiple and subtle skills related to their understanding of the World and the problems they have been brought in to discuss and hopefully solve.

In cases where some defined technical skill is required, these can always be bought in, or trained – our most valuable inputs from humans, rather than AI, will however come from those who can see the World, and think about what they see, in a way that is only *informed*, rather than *bound*, by a discipline-specific mindset. A university education that equips graduates for this reality is crucial.

Consider the example of autonomous vehicles, and the ethical quandaries they spark discussion of. For example, consider an autonomous vehicle that is in a situation where it must make one of two choices: swerving to avoid a collision, but doing so with a high probability of going off a bridge and killing its passengers; or continuing along its current path, which would involve a high probability of killing some number of pedestrians.

Both options would no doubt be tragic, even as the calculus of *how many people and who they are* is omitted. Those details are omitted precisely to make the point that the *technological* decisions regarding risk-aversion programmed into the autonomous vehicle rest upon a myriad of assumptions about the relative value of life and the degree to which risk-aversion should trump efficiency *on top of* the engineering decisions that a philosopher such as myself could have no legitimate input into.

If it is only technologists who program machines that make decisions with serious implications, or only philosophers that do so, the outcomes are not likely to be favorable – we need both of those inputs (and more) in the room when these decisions are made. The solution is not simply AI, because if these choices are left in the hands of machine intelligence, we should still be concerned with who the people are who program the decisions, and what the justification or reasoning is behind the frameworks they use to allow the AI to make them.

In spite of these concerns, an increased interest in technology and computer-science related careers has correlated with a precipitous drop in the proportion of humanities majors at colleges in the USA and elsewhere (Heller, 2023). This should be of concern to all educators, in that we are in a time of epistemic crisis, with people retreating to more polarized and hardened views, and where collaboration is under strain, which is precisely when the collaborative and socially-motivated reasoning practices described here can be most valuable.

We of course need to educate people so they are productive and employable, but we also need to be educating people so that they're capable of helping to create a society that is livable and social, and where the value of human interaction is recognized for more than its sentimental value, but more because we know how important it is to reaching conclusions that accommodate our respective skills, and that respect the unique value that humans in collaborative engagement can add.

4. Conclusions

The low-hanging fruit, in terms of a list of obvious steps to take in response to AI in education, would include at least the following:

- Create a policy: create a policy on how ChatGPT should be used. This should cover areas such as data security, privacy, and responsible use of the technology.
- Educate users: we should educate users on the responsible use of ChatGPT and ensure that they understand the potential implications of using the technology.
- Monitor and review usage: we would ideally monitor how ChatGPT is being used, so as to take action, where possible, to encourage and empower employees to exercise their agency and creativity in doing so, rather than ceding their authority to those tools.

Steps such as those listed above do not, however, speak to the most significant challenge, which is that we need find ways to encourage and reward human creativity and collaboration. This is because humans deliberating together – and yes, making mistakes – is emblematic of the intellectual journey that teaching and learning offers.

Societies, and the humans that they comprise, are capable of feats of imagination and ingenuity that can result in unexpected insights in political theory or revolutionary scientific findings, many of which have arisen out of years of research, investigation, argument, and frustration of the sort that we might stop engaging in at all, once we become largely dependent on algorithmic outcomes.

Furthermore, societies in which AI resources are less available will likely be the same ones who are currently under-resourced in comparison to the Global North, in terms of their economies and educational systems.

The argument could therefore be made that there is a moral obligation to resist AI, for the sake of equality. This would not only be a tenuous argument – in that it's also possible that AI will lead to *increased* socio-economic equality – but is also not the primary argument made here, which is that the noise, the fuzziness, the mistakes in our communal deliberations add a value that cannot currently be served by AI, and that this is a value not be forsaken, even as we exploit the many opportunities that AI offers for improving our lives.

References

Bostrom, Nick (2003). "Ethical Issues in Advanced Artificial Intelligence". *Nickbostrom.com*. https://nickbostrom.com/ethics/ai

Feng, J. (2022, December 12). “How China's Social Credit System Works”. *Newsweek.* https://www.newsweek.com/china-social-credit-system-works-explained-1768726

Heller, N. (2023, February 27). “The End of the English Major”. *New Yorker.* https://www.newyorker.com/magazine/2023/03/06/the-end-of-the-english-major

Mercier, H., & Sperber, D. (2017). *The enigma of reason*. Harvard University Press. https://doi.org/10.4159/9780674977860

Rousseau, J. (2021). *Challenges to Science Communication in a Post-Truth World*. Communicatio, 47:2, 122-140, DOI: 10.1080/02500167.2021.1959363

Shermer, M. (2011). *The believing brain: From ghosts to gods to politics and conspiracies—How we construct beliefs and reinforce them as truths*. New York, NY: St. Martin’s Griffin.

Contact email: Jacques.Rousseau@uct.ac.za