Is openness in AI research always the answer?

*As research into AI has become more developed, so too has the understanding that AI research might be misused by bad actors. Discussing OpenAI's recent decision to withhold the source code for an algorithm designed to replicate handwriting, citing concerns for the public wellbeing, **Gabrielle Samuel** argues that blanket commitments to openness are insufficient to protect against the potential 'dual-use' of AI research and that AI researchers need to develop a shared ethical code of conduct for releasing their research findings.*

The explosion of Artificial intelligence ("AI") research, has triggered concerns and significant speculation about what an AI future might bring. Concerns which were further underscored by news of a massive £150 million investment in an Oxford University AI ethics centre. In a move to control and regulate the field, a plethora of AI ethics guidelines and recommendations have emerged. One of the key principles cited in these guidelines is the need for openness to ensure fairness, equity and transparency.

However, blanket principles such as these can also serve to hide complexities and moral tensions. One example of this is the widespread controversy which ensued from OpenAI's recent decision to not release the source code for their GPT-2 algorithm. This algorithm, they say, has been trained to predict the next word in a text, given all the previous words, and as such can generate realistic and coherent text in a variety of styles that looks like a human has written it. The non-profit consortium's decision was based on concerns about "malicious applications of the technology", by which they mean they had concerns that GPT-2 could be used to automate fake content online (i.e., generate fake news), impersonate people online, or create spam or propaganda online. Instead of releasing the source code, they released a much smaller AI model as "an experiment in responsible disclosure".

Academics and the private sector have already been in heated discussions about whether such algorithms should or should not be open-sourced, and this widely publicised decision has only amplified the debate and added further controversies to a field which is already struggling to find shared understandings of ethics practice.

Putting issues of whether OpenAI's algorithm works and/or whether Open AI are only withholding their data as a publicity stunt (a hype which is certainly not unfamiliar to the AI world), this incident causes us to re-visit the question: with demands for openness, are there any contexts within which AI research should be withheld from society, particularly if it is likely to lead to harm?

This question is not new, researchers and policymakers have been grappling with it for years in other fields, whether in relation to dual-use research in the biosecurity and bioterrorism arena (ie. research developed for social good but which can be directly misapplied to pose significant threat to society), or in relation to withholding health/medical research findings which may have potential detrimental impacts on patient behaviour and health outcomes.

While a range of commentators argue for more regulation and control over what gets published, others support the need for scientific freedom in the scientific endeavour. This has led to a stalemate in terms of how these issues can be addressed, and a number of unresolved questions. Are there any solutions?

As part of Wellcome Trust-funded research project, we interviewed a range of UK University AI researchers to explore the types of ethical decisions AI Academy researchers make during their research. Our research shows how our participants are ethically aware and concerned with the societal impact of their work. They also show that they are aware of the limitations of University ethics governance systems currently in place to monitor this. These systems, they say, are designed to respond to questions more focused on the ethics of data use in AI research, rather than on the AI tool itself, or its broader societal impacts.

This leaves researchers - and possibly reviewers and/or Editors of scientific journals - in unchartered territory when making decisions about publication in terms of the types of ethical issues they need to consider (does it need ethics approval? Should AI research be published if it could be considered dual use? What type of ethical issues should the authors have considered?). In fact, our previous research has shown how reviewers and Editors have inconsistent practices in terms of whether any ethical issues at all are considered during the publication of digital research. More generally, Ploug has also shown that publication ethics guidelines provide little instruction in the area of dual use research in medical research.

The European Commission's *Ethics guidelines for Trustworthy AI* has tried to deal with this. It asks AI researchers to consider: '*to what degree your system could be dual-use? If so, did you take suitable preventative measures against this case (including for instance not publishing the research or deploying the system)?'*

But, raising the question almost never provides a solution, without shared understandings of what to do if you answer "yes" to the question. When there is an absence of shared understandings, a culture of "personal ethics" can emerge in which researchers draw on their own ethical principles to determine how best to act.

If we really do want to think seriously about these questions, and about if or when it would be morally responsible to withhold AI research findings, we need to do more than pose questions to AI researchers. We need to think long and hard about the answers.

The challenge, as has previously been argued in relation to other dual use technologies, is to devise a legislative or regulatory system that balances security and safety without imposing unreasonable bureaucratic burdens. A self-regulatory code of conduct at the point of dissemination is one possibility, in which there are editorial processes to scrutinise manuscripts that might pose security threats and an independent authority to oversee this. Another suggestion could be only releasing the AI code at a specific time point, which would leave enough time for other researchers to create a counter-AI tool.

At the very least we should think about how we would determine who decides. In this case it was OpenAI, but should it always be the researcher, especially if they are making decisions based on a "personal ethics"? Also, how should we define which potential harms warrant not publishing research, and what does their magnitude and likelihood need to be?

These questions remind us that while AI may have become 'ethicised', underneath, there are some very real moral questions which need attending to.