The tension between openness and prudence in responsible AI research

Jess Whittlestone[1] Aviv Ovadya[2]

[1]Leverhulme Centre for the Future of Intelligence, University of Cambridge
[2]The Thoughtful Technology Project

Introduction

The AI research community has historically had strong norms around openness. This is clear in, for example, community backlash against the closed access Nature Machine Intelligence journal (Dietterich 2018), and the fact that many top conferences are moving towards open review and pushing for code and dataset releases. Openness in research is valuable for many reasons, including ensuring that the benefits of research are distributed widely, and enabling scientific progress by helping researchers build on one another's work. However, concerns about potential harms arising from the misuse of AI research are growing (Brundage 2018; Ovadya and Whittlestone 2019), prompting some to consider whether the field should reconsider norms around the publication and dissemination of research.

Core disagreements

AI researchers hold varying opinions on how we should balance the tension between openness and prudence in AI research. In order to have a more productive conversation about this disagreement, we highlight three types of belief which seem particularly central to assessing the relative importance of openness vs. prudence.

Beliefs about risks/harms

- How significant are the potential harms of research being misused?
- How significant are the potential harms of decreased openness in research, such as reduced inclusivity?
- How can we balance these harms against each other?

Beliefs about efficacy

- How effective is reducing research release likely to be at mitigating malicious use or other harms in practice?

Beliefs about future needs

- Are more advanced systems capable of greater harm likely to be developed soon?
- Should/can we start exploring options now to prepare for potential future misuse?

Questions for publication norms

(1) **What different options** are there for how research is released?
(2) **Under what circumstances** should different types of release be used?
(3) **What processes** should govern how these decisions about release type are made?
(4) **Who should be involved** in making these decisions?
(5) **Who or what should manage** (and fund) all of the above?

Learning from other fields

Both biology and computer security, which have some precedent for restricting release of outputs of potentially harmful research, have established procedures and institutions underpinning these decisions. Biosafety practices include processes for classifying the risk level of different microorganisms, determined by specialist organisations (Atlas and Dando, 2006). Similarly, computer security has processes for responsibly disclosing critical information with potential for misuse, which are managed by entities such as Information Security and Analysis Centres (European Union Agency for Network and Information Security, 2018).

Recommendations

We suggest that future work should:

- **Aim to better understand risks of misuse** across different areas of AI research
- **More thoroughly investigate potential harms of reduced openness** in AI research, including by (a) more substantively engaging with different communities to understand concerns; and (b) better understanding how harms have arisen and been dealt with in other fields
- **Identify areas where a deeper tension between openness and prudence exists**, and so identify specific questions in need of deeper ethical and philosophical analysis
- **Explore different options for publication norms and processes** and their real-world impacts in much more detail
- **Build a community** around exploring and acting on these issues, with established venues for discussions, such as specific workshops on responsible AI research and publication