Advancing ethics review practices in AI research

Madhulika Srikumar, Rebecca Finlay, Grace Abuhamad, Carolyn Ashurst, Rosie Campbell, Emily Campbell-Ratcliffe, Hudson Hongo, Sarah R. Jordan, Joseph Lindley, Aviv Ovadya & Joelle Pineau

The implementation of ethics review processes is an important frst step for anticipating and mitigating the potential harms of AI research. Its long-term success, however, requires a coordinated community efort, to support experimentation with diferent ethics review processes, to study their efect, and to provide opportunities for diverse voices from the community to share insights and foster norms.

As artificial intelligence (AI) and machine learning (ML) technologies continue to advance, awareness of the potential negative consequences on society of AI or ML research has grown. Anticipating and mitigating these consequences can only be accomplished with the help of the leading experts on this work: researchers themselves. Several leading AI and ML organizations, conferences and journals have therefore started to implement governance mechanisms that require researchers to directly confront risks related to their work that can range from malicious use to unintended harms. Some have initiated new ethics review processes, integrated within peer review, which primarily facilitate a reflection on the potential risks and effects on society after the research is conducted. This is distinct from other responsibilities that researchers undertake earlier in the research process, such as the protection of the welfare of human participants, which are governed by bodies such as institutional review boards (IRBs). Although these initiatives are commendable, they have yet to be widely adopted. They are being pursued largely without the benefit of community alignment. As researchers and practitioners from academia, industry and non-profit organizations in the field of AI and its governance, we believe that community coordination is needed to ensure that critical reflection is meaningfully integrated within AI research to mitigate its harmful downstream consequences. The pace of AI and ML research and its growing potential for misuse necessitates that this coordination happen today. Writing in Nature Machine Intelligence, Prunkl et al. [1] argue that the AI research community needs to encourage public deliberation on the merits and future of impact statements and other self-governance mechanisms in conference submissions. We agree. Here, we build on this suggestion, and provide three recommendations to enable this effective community coordination, as more ethics review approaches begin to emerge across conferences and journals. We believe that a coordinated community effort will require: (1) more research on the effects of ethics review processes; (2) more experimentation with such processes themselves; and (3) the creation of venues in which diverse voices both within and beyond the AI or ML community can share insights and foster norms. Although many of the challenges we address have been previously highlighted [1–6] , this Comment takes a wider view, calling for collaboration between different conferences and journals by contextualizing this conversation against more recent studies [7–11] and developments.

Developments in AI research ethics

In the past, many applied scientific communities have contended with the potential harmful societal effects of their research. The infamous anthrax attacks in 2001, for example, catalysed the creation of the National Science Advisory Board for Biosecurity to prevent the

misuse of biomedical research. Virology, in particular, has had long-running debates about the responsibility of individual researchers conducting gain-of-function research. Today, the field of AI research finds itself at a similar juncture [12]. Algorithmic systems are now being deployed for high-stakes applications such as law enforcement and automated decision-making, in which the tools have the potential to increase bias, injustice, misuse and other harms at scale. The recent adoption of ethics and impact statements and checklists at some AI conferences and journals signals a much-needed willingness to deal with these issues. However, these ethics review practices are still evolving and are experimental in nature. The developments acknowledge gaps in existing, well-established governance mechanisms, such as IRBs, which focus on risks to human participants rather than risks to society as a whole. This limited focus leaves ethical issues such as the welfare of data workers and non-participants, and the implications of data generated by or about people outside of their scope [6] . We acknowledge that such ethical reflection, beyond IRB mechanisms, may also be relevant to other academic disciplines, particularly those for whom large datasets created by or about people are increasingly common, but such a discussion is beyond the scope of this piece. The need to reflect on ethical concerns seems particularly pertinent within AI, because of its relative infancy as a field, the rapid development of its capabilities and outputs, and its increasing effects on society. In 2020, the NeurIPS ML conference required all papers to carry a 'broader impact' statement examining the ethical and societal effects of the research. The conference updated its approach in 2021, asking authors to complete a checklist and to document potential downstream consequences of their work. In the same year, the Partnership on AI released a white paper calling for the field to expand peer review criteria to consider the potential effects of AI research on society, including accidents, unintended consequences, inappropriate applications and malicious uses [3]. In an editorial citing the white paper, Nature Machine Intelligence announced that it would ask submissions to carry an ethical statement when the research involves the identification of individuals and related sensitive data [13], recognizing that mitigating downstream consequences of AI research cannot be completely disentangled from how the research itself is conducted. In another recent development, Stanford University's Ethics and Society Review (ESR) requires AI researchers who apply for funding to identify if their research poses any risks to society and also explain how those risks will be mitigated through research design14. Other developments include the rising popularity of interdisciplinary conferences examining the effects of AI, such as the ACM Conference on Fairness, Accountability, and Transparency (FAccT), and the emergence of ethical codes of conduct for professional associations in computer science, such as the Association for Computing Machinery (ACM). Other actors have focused on upstream initiatives such as the integration of ethics reflection into all levels of the computer science curriculum. Reactions from the AI research community to the introduction of ethics review practices include fears that these processes could restrict open scientific inquiry [3]. Scholars also note the inherent difficulty of anticipating the consequences of research [1] , with some AI researchers expressing concern that they do not have the expertise to perform such evaluations [7]. Other challenges include concerns about the lack of transparency in review practices at corporate research labs (which increasingly contribute to the most highly cited papers at premier AI conferences such as NeurIPS and ICML [9]) as well as academic research culture and incentives supporting the 'publish or perish' mentality that may not allow time for ethical reflection. With the emergence of these new attempts to acknowledge and articulate unique ethical considerations in AI research and the resulting concerns from some

researchers, the need for the AI research community to come together to experiment, share knowledge and establish shared best practices is all the more urgent. We recommend the following three steps.

Study community behaviour and share learnings

So far, there are limited studies that have explored the responses of ML researchers to the launch of experimental ethics review practices. To understand how behaviour is changing and how to align practice with intended effect, we need to study what is happening and share learnings iteratively to advance innovation. For example, in response to the NeurIPS 2020 requirement for broader impact statements, a paper found that most researchers surveyed spent fewer than two hours working on this process [7], perhaps retroactively towards the end of their research, making it difficult to know whether this reflection influenced or shifted research directions or not. Surveyed researchers also expressed scepticism about the mandated reflection on societal impacts [7]. An analysis of preprints found that researchers assessed impact through the narrow lens of technical contributions (that is, describing their work in the context of how it contributes to the research space and not how it may affect society), thereby overlooking potential effects on vulnerable stakeholders [8]. A qualitative analysis of a larger sample [10] and a quantitative analysis of all submitted papers [11] found that engagement was highly variable, and that researchers tended to favour the discussion of positive effects over negative effects. We need to understand what works. These findings, all drawn from studies examining the implementation of ethics review at NeurIPS 2020, point to a pressing need to review actual versus intended community behaviour more thoroughly and consistently to evaluate the effectiveness of ethics review practices. We recognize that other fields have considered ethics in research in different ways. To get started, we propose the following approach, building on and expanding the analysis of Prunkl et al. [1]. First, clear articulation of the purposes behind impact statements and other ethics review requirements is needed to evaluate efficacy and motivate future iterations by the community. Publication venues that organize ethics review must communicate expectations of this process comprehensively both at the level of individual contribution and for the community at large. At the individual level, goals could include encouraging researchers to reflect on the anticipated effects on society. At the community level, goals could include creating a culture of shared responsibility among researchers and (in the longer run) identifying and mitigating harms. Second, because the exercise of anticipating downstream effects can be abstract and risks being reduced to a box-ticking endeavour, we need more data to ascertain whether they effectively promote reflection. Similar to the studies above, conference organizers and journal editors must monitor community behaviour through surveys with researchers and reviewers, partner with information scientists to analyse the responses [15], and share their findings with the larger community. Reviewing community attitudes more systematically can provide data both on the process and effect of reflecting on harms for individual researchers, the quality of exploration encountered by reviewers, and uncover systemic challenges to practicing thoughtful ethical reflection. Work to better understand how AI researchers view their responsibility about the effects of their work in light of changing social contexts is also crucial. Evaluating whether AI or ML researchers are more explicit about the downsides of their research in their papers is a preliminary metric for measuring change in community behaviour at large [2]. An analysis of the potential negative

consequences of AI research can consider the types of application the research can make possible, the potential uses of those applications, and the societal effects they can cause [4]. Building on the efforts at NeurIPS [16] and NAACL[17], we can openly share our learnings as conference organizers and ethics committee members to gain a better understanding of what does and does not work. Community behaviour in response to ethics review at the publication stage must also be studied to evaluate how structural and cultural forces throughout the research process can be reshaped towards more responsible research. The inclusion of diverse researchers and ethics reviewers, as well as people who face existing and potential harm, is a prerequisite to conduct research responsibly and improve our ability to anticipate harms.

Expand experimentation of ethical review

The low uptake of ethics review practices, and the lack of experimentation with such processes, limits our ability to evaluate the effectiveness of different approaches. Experimentation cannot be limited to a few conferences that focus on some subdomains of ML and computing research — especially for subdomains that envision real-world applications such as in employment, policing and healthcare settings. For instance, NeurIPS, which is largely considered a methods and theoretical conference, began an ethics review process in 2020, whereas conferences closer to applications, such as top-tier conferences in computer vision, have yet to implement such practices. Sustained experimentation across subfields of AI can help us to study actual community behaviour, including differences in researcher attitudes and the unique opportunities and challenges that come with each domain. In the absence of accepted best practices, implementing ethics review processes will require conference organizers and journal editors to act under uncertainty. For that reason, we recognize that it may be easier for publication venues to begin their ethics review process by making it voluntary for authors. This can provide researchers and reviewers with the opportunity to become familiar with ethical and societal reflection, remove incentives for researchers to 'game' the process, and help the organizers and wider community to get closer to identifying how they can best facilitate the reflection process.

Create venues for debate, alignment and collective action

This work requires considerable cultural and institutional change that goes beyond the submission of ethical statements or checklists at conferences. Ethical codes in scientific research have proven to be insufficient in the absence of community-wide norms and discussion [1]. Venues for open exchange can provide opportunities for researchers to share their experiences and challenges with ethical reflection. Such venues can be conducive to reflect on values as they evolve in AI or ML research, such as topics chosen for research, how research is conducted, and what values best reflect societal needs. The establishment of venues for dialogue where conference organizers and journal editors can regularly share experiences, monitor trends in attitudes, and exchange insights on actual community behaviour across domains, while considering the evolving research landscape and range of opinions, is crucial. These venues would bring together an international group of actors involved throughout the research process, from funders, research leaders, and publishers to interdisciplinary experts adopting a critical lens on AI impact, including social scientists, legal scholars, public interest advocates, and policymakers. In addition, reflection and dialogue can have a powerful role in influencing the future trajectory of a technology. Historically, gatherings convened by scientists have had far-reaching effects — setting the norms that

guide research, and also creating practices and institutions to anticipate risks and inform downstream innovation. The Asilomar Conference on Recombinant DNA in 1975 and the Bermuda Meetings on genomic data sharing in the 1990s are instructive examples of scientists and funders, respectively, creating spaces for consensus-building [18,19]. Proposing a global forum for gene-editing, scholars Jasanoff and Hulburt argued that such a venue should promote reflection on "what questions should be asked, whose views must be heard, what imbalances of power should be made visible, and what diversity of views exist globally" [20]. A forum for global deliberation on ethical approaches to AI or ML research will also need to do this. By focusing on building the AI research field's capacity to measure behavioural change, exchange insights, and act together, we can amplify emerging ethical review and oversight efforts. Doing this will require coordination across the entire research community and, accordingly, will come with challenges that need to be considered by conference organizers and others in their funding strategies. That said, we believe that there are important incremental steps that can be taken today towards realizing this change. For example, hosting an annual workshop on ethics review at pre-eminent AI conferences, or holding public panels on this subject [21], hosting a workshop to review ethics statements [22], and bringing conference organizers together [23]. Recent initiatives undertaken by AI research teams at companies to implement ethics review processes [24], better understand societal impacts [25] and share learnings [26,27] also show how industry practitioners can have a positive effect. The AI community recognizes that more needs to be done to mitigate this technology's potential harms. Recent developments in ethics review in AI research demonstrate that we must take action together.

References

1. Prunkl, C. E. A. et al. Nat. Mach. Intell. 3, 104–110 (2021).
2. Hecht, B. et al. Preprint at https://doi.org/10.48550/arXiv.2112.09544 (2021).
3. Partnership on AI. https://go.nature.com/3UUX0p3 (2021).
4. Ashurst, C. et al. https://go.nature.com/3gsQfvp (2020).
5. Hecht, B. https://go.nature.com/3AASZhf (2020).
6. Ashurst, C., Barocas, S., Campbell, R., Raji, D. in FAccT '22: 2022 ACM Conf. on Fairness, Accountability, and Transparency 2057–2068 (2022).
7. Abuhamad, G. et al. Preprint at https://arxiv.org/abs/2011.13032 (2020).
8. Boyarskaya, M. et al. Preprint at https://arxiv.org/abs/2011.13416 (2020).
9. Birhane, A. et al. in FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency 173–184 (2022).
10. Nanayakkara, P. et al. in AIES '21: Proc. 2021 AAAI/ACM Conference on AI, Ethics, and Society 795–806 (2021).
11. Ashurst, C., Hine, E., Sedille, P. & Carlier, A. in FAccT '22: 2022 ACM Conf. on Fairness, Accountability, and Transparency 2047–2056 (2022).12. National Academies of Sciences, Engineering, and Medicine. https://go.nature.com/ 3UTKOEJ (date accessed 16 September 2022).
13. Nat. Mach. Intell. 3, 367 (2021).
14. Bernstein, M. S. et al. Proc. Natl Acad. Sci. USA 118, e2117261118 (2021).
15. Pineau, J. et al. J. Mach. Learn. Res. 22, 7459–7478 (2021).
16. Benjio, S. et al. Neural Information Processing Systems. https://go.nature.com/3tQxGEO (2021).
17. Bender, E. M. & Fort, K. https://go.nature.com/3TWnbua (2021).

18. Gregorowius, D., Biller-Andorno, N. & Deplazes-Zemp, A. EMBO Rep. 18, 355–358 (2017).
19. Jones, K. M., Ankeny, R. A. & Cook-Deegan, R. J. Hist. Biol. 51, 693–805 (2018).
20. Jasanof, S. & Hurlbut, J. B. Nature 555, 435–437 (2018).
21. Partnership on AI. https://go.nature.com/3EpQwY4 (2021).
22. Sturdee, M. et al. in CHI Conf. Human Factors in Computing Systems Extended Abstracts (CHI ’21 Extended Abstracts); https://doi.org/10.1145/3411763.3441330 (2021).
23. Partnership on AI. https://go.nature.com/3AzdNFW (2022).
24. DeepMind. https://go.nature.com/3EQyUWT (2022).
25. Meta AI. https://go.nature.com/3i3PBVX (2022).
26. Munoz Ferrandis, C. OpenRAIL; https://huggingface.co/blog/open_rail (2022). 27. OpenAI. https://go.nature.com/3GyZPYk (2022).