Managing the Risks of AI Research

Six Recommendations for Responsible Publication

May 6, 2021

PARTNERSHIP ON AI

Table Of Contents

Executive Summary

Once a niche research interest, artificial intelligence (AI) has quickly become a pervasive aspect of society with increasing influence over our lives. In turn, open questions about this technology have, in recent years, transformed into urgent ethical considerations. This white paper by the Partnership on AI (PAI) offers recommendations addressing one such question: **Given AI's potential for misuse, how can AI research be disseminated responsibly?**

Many research communities, such as biosecurity and cybersecurity, routinely work with information that could be used to cause harm, either maliciously or accidentally. These fields have thus established their own norms and procedures for publishing high-risk research. Thanks to breakthrough advances, AI technology has progressed rapidly in the past decade, giving the AI community less time to develop similar practices.

Recent pilots, such as OpenAI's "staged release" of GPT-2 and the "broader impact statement" requirement at the 2020 NeurIPS conference, demonstrate a growing interest in responsible AI publication norms. Effectively anticipating and mitigating the potential negative impacts of AI research, however, will require a community-wide effort. As a first step towards developing responsible publication practices, **this white paper provides recommendations for three key groups in the AI research ecosystem:**

- **Individual researchers,** who should disclose and report additional information in their papers and normalize discussion about the downstream consequences of research.
- **Research leadership,** which should review potential downstream consequences earlier in the research pipeline and commend researchers who identify negative downstream consequences.
- **Conferences and journals,** which should expand peer review criteria to include engagement with potential downstream consequences and establish separate review processes to evaluate papers based on risk and downstream consequences.

Additionally, this white paper includes an appendix which seeks to disambiguate a variety of terms related to responsible research which are often conflated: "research integrity," "research ethics," "research culture," "downstream consequences," and "broader impacts."

This document represents an artifact that can be used as a basis for further discussion, and we seek feedback on it to inform future iterations of the recommendations it contains. Our aim is to help build our capacity as a field to anticipate downstream consequences and mitigate potential risks.

Introduction

Context

In the past two years, PAI has hosted a variety of multistakeholder convenings exploring how advances in AI can be disseminated in a responsible manner that is mindful of downstream consequences. These took place against a backdrop of pilots on responsible dissemination within the AI community such as OpenAI's staged release approach for GPT-2,[1] the 2020 NeurIPS "broader impact statement" requirement,[2] and the 2021 NeurIPS Paper Checklist.[3]

Through these convenings, we quickly discovered that there were a wide variety of views within the AI community on this issue. Some researchers saw the consideration of downstream consequences as outside their remit and believed their focus should be on scientific progress alone. Others expressed a sense of personal or professional responsibility about the impacts of their work, but felt ill-equipped or confused when it came to what, if anything, to do about it.[4] And while many more in the community agreed on the importance of ensuring beneficial outcomes from AI, they often strongly disagreed about which strategies would be most effective, the appropriate balance between different values such as open scientific inquiry and mitigating possible harm,[5] and how to weigh uncertain positive and negative consequences against each other.

Despite the lack of community consensus on the responsible dissemination of AI research, the pace of this research imparts an urgency which requires action under uncertainty. As more fully explained in Appendix I, we have come to the conclusion that there is, in fact, an important role for the AI research community to play in mitigating negative impacts. As members of society and domain experts, researchers cannot completely dissociate the potential harmful consequences of the tools they enable from their moral and professional obligations.

This white paper synthesizes key themes that emerged during our multistakeholder convenings and research,[6] providing specific recommendations for fostering a responsible research culture that includes the consideration of downstream consequences. Our aim is to help build our capacity as a field to anticipate downstream consequences and mitigate potential risks. This white paper represents an artifact that can be used as a basis for further discussion, and we seek feedback on it to inform future iterations of the recommendations it contains.

1 *Better language models and their implications.* (2019, February 14). OpenAI. https://openai.com/blog/better-language-models/

2 *NeurIPS 2020 FAQ. (n.d.). Retrieved April 8, 2021*, from https://nips.cc/Conferences/2020/PaperInformation/NeurIPS-FAQ

3 Conference, N. I. P. S. (2021, March 26). *Introducing the NeurIPS 2021 Paper Checklist.* Medium. https://neuripsconf.medium.com/introducing-the-neurips-2021-paper-checklist-3220d6df500b

4 Abuhamad, G., & Rheault, C. (2020). Like a researcher stating broader impact for the very first time. *ArXiv:2011.13032 [Cs].* http://arxiv.org/abs/2011.13032

5 Whittlestone, J., & Ovadya, A. (2020). The tension between openness and prudence in AI research. ArXiv:1910.01170 [Cs]. http://arxiv.org/abs/1910.01170

6 Recordings and papers from our 2020 NeurIPS workshop on Navigating the Broader Impacts of AI Research can be found here. In addition to small group consultations, we hosted events at a variety of other venues, including a workshop with the Montreal AI Ethics Institute, and conducted research into other high-stakes fields, including biosecurity and cybersecurity.

Scope of this document

This white paper focuses specifically on the challenge of anticipating the downstream consequences of AI research[7] and mitigating potential negative impacts such as accidents, unintended consequences, inappropriate applications, malicious use, or other systemic harms.[8] We have not primarily focused on other aspects of responsible research such as research integrity and research ethics, though there are certainly areas where these overlap. In conversations on responsible research practices, we've found these terms can often get conflated, so we have provided more detail to help disambiguate them in Appendix II.

We have primarily focused on interventions at the point of publication. There are other important points in the research pipeline (such as conception, funding, and productization),[9] but these are largely beyond the scope of this document (though we do touch on the benefits of anticipating downstream consequences in advance of publication). Since AI research is disseminated in a variety of ways (such as through peer-reviewed conferences and journals, preprint archives, blog posts, and open-source code), we use the term "publication" broadly, inclusive of non-traditional publication methods.

Through our work, it has become clear that effectively anticipating and mitigating downstream consequences of AI research requires community-wide effort; it cannot be the responsibility of any one group alone. The AI research ecosystem includes both industry and academia, and comprises researchers, engineers, reviewers, conferences, journals, grantmakers, team leads, product managers, administrators, communicators, institutional leaders, data scientists, social scientists, policymakers, and others. Addressing all of these different roles is beyond the scope of this report, which focuses on three key groups: individual researchers, research leadership, and conferences and journals.[10] In future iterations, we hope to expand our consideration to include other groups.

The recommendations in this white paper are presented as a first step towards responsible research practices that better equip the field to navigate the downstream consequences of AI. In the section on "Outstanding challenges and further work," we present some thoughts on additional activities that might be required.

7 By AI research, we mean fundamental advances in AI technology, investigations seeking to gain deeper analytical insight into these advances, or significant advances in how AI technologies are applied, regardless of whether this is undertaken in an academic or commercial setting. Defining AI is a rabbit hole in itself, but we broadly take it to mean the use of computational processes to synthesize the appearance of intelligence. Currently, Machine Learning (ML) is the leading approach to this, and examples include identifying and deciphering the objects in images ("computer vision (CV)"), interpreting text in various ways ("natural language processing (NLP)"), controlling robots or game agents via strong feedback loops ("reinforcement learning"), or synthesizing convincing text and images.

8 Some examples of how AI research might result in negative impacts include: AI tools for the identification of humans that are deployed in surveillance applications, bias in datasets and algorithms that are amplified when misapplied in a different context, privacy-sensitive tools developed with certain safety promises that are later disregarded due to mission creep. For real-world examples, see the PAI AI Incidents Database. Much has been written about the unintended consequences of AI and the risks of prematurely applying it in certain domains; see the citations on the Navigating the Broader Impacts of AI research workshop webpage. For a fuller discussion of malicious uses, see: Brundage, et al. (2018). The Malicious Use of Artificial Intelligence. https://maliciousaireport.com/

9 See: Decision points in AI governance. (2020). UC Berkeley Center for Long-Term Cybersecurity. https://cltc.berkeley.edu/2020/05/05/decision-points-in-ai-governance/ for further reading on intervention points for AI research

10 We chose to initially focus on these groups because they are the closest to the research and peer review process, and are well-placed to effect cultural and policy changes. They are also well-represented in the PAI community, and have been most actively involved in our convenings and consultations on this topic.

Recommendations

Recommendations for individual researchers

Researcher expertise is essential to anticipating and mitigating the potential risks of AI technologies. While institutional-level actors can support an ecosystem that more broadly incentivizes responsible practices, many steps of the research process exist beyond the scope of their policies. Furthermore, the precise application of these policies depends on the fidelity of the information provided by researchers.

Over the last few years, a number of AI researchers have faced significant criticism for overlooking or disregarding harmful impacts.[11] Many of these incidents have related to research that involves the identification, classification, or prediction of human characteristics or socio-economic behaviors, such as the prediction of sexual orientation[12] or "trustworthiness"[13] from facial images. Facial recognition research has also come under fire for enabling the surveillance of vulnerable populations, with many researchers expressing discomfort at these troubling uses of AI.[14] As AI continues to advance and be deployed in new domains, it will become increasingly necessary for researchers to reflect on the possible impacts of their work to avoid being blindsided by harmful consequences.

Researchers remain the leading experts on their own work. This makes them uniquely qualified to describe the relevance of their research and weigh in on downstream consequences – actions that will both contribute to responsible AI practices and protect their work from mischaracterization.[15]

1. Disclose and report additional information in your papers

A significant way individual researchers can contribute to a culture of responsible publication in AI is by explicitly reporting relevant additional details about their work.[16] Well-documented research not only helps with the anticipation of potential impacts, but also, once the impacts are apparent, allows for reflection on the choices that were made so that more informed decisions can be made in the future.

11 Hutson, M. (2021, February 15). *Who should stop unethical A.I.?* The New Yorker. https://www.newyorker.com/tech/annals-of-technology/who-should-stop-unethical-ai

12 Arcas, B. A. y, Todorov, A., & Mitchell, M. (2018, January 18). *Do algorithms reveal sexual orientation or just expose our stereotypes?* Medium. https://medium.com/@blaisea/do-algorithms-reveal-sexual-orientation-or-just-expose-our-stereotypes-d998fafdf477

13 Ongweso Jr, E. (2020, September 28). An ai paper published in a major journal dabbles in phrenology.Vice. https://www.vice.com/en/article/g5pawq/an-ai-paper-published-in-a-major-journal-dabbles-in-phrenology

14 Noorden, R. V. (2020). The ethical questions that haunt facial-recognition research. *Nature, 587*(7834),354–358. https://doi.org/10.1038/d41586-020-03187-3

15 For more on why researchers are well positioned to anticipate risks, see: Prunkl, C. E. A., Ashurst, C., Anderljung, M., Webb, H., Leike, J., & Dafoe, A. (2021). Institutionalizing ethics in AI through broader impact requirements. *Nature Machine Intelligence,* 3(2), 104–110. https://doi.org/10.1038/s42256-021-00298-y

16 This complements other ongoing efforts to improve transparency and documentation for Machine Learning models, such as PAI's About ML initiative.

For its 2021 conference, NeurIPS introduced a checklist "designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact,"[17] providing a template for how researchers can begin considering these issues. However, this kind of reporting is not yet a common requirement at conferences, and (as previously discussed) AI research sometimes bypasses the peer review process entirely. We therefore encourage authors to voluntarily include additional information in their papers even when it has not been explicitly requested, including when the research is being disseminated outside of the peer review process.

There are three key items of additional information we suggest researchers report: contribution and motivation, consideration of downstream consequences, and the amount of computational resources used. While the NeurIPS checklist covers many aspects of responsible research, we believe these items are particularly important to include for the purposes of anticipating and mitigating downstream consequences

Contribution and motivation

It is not always clear how much of a contribution a paper is claiming to make. Is it an incremental improvement on an existing benchmark? Or an entirely new technique? This is important because, in general, as discussed below, the greater the contribution of a research paper, the greater the responsibility to consider its potential impacts.

Additionally, AI researchers often report results on an innovative deep learning technique or improved performance without being clear about the motivation for the work or its intended applications. For research that is particularly sensitive (such as the kind of research mentioned previously that involves predicting human characteristics), it is important to weigh the potential benefits against the risks, yet this is made hard by a lack of clear reporting on motivation for the work or its intended application? This isn't to suggest that all research needs to explicitly have socially beneficial applications to be worthwhile, but for research that clearly has significant risks (e.g. video manipulation techniques that could be used to create non-consensual deepfakes), it's helpful to indicate what the possible legitimate applications are (e.g. face swapping for digital effects in movies), and where appropriate to signpost mitigations (e.g. algorithms that can detect manipulated images) to enable a more thorough risk analysis. This should be balanced by a discussion of possible negative applications as part of the consideration of downstream consequences, discussed below.

Clearly reporting the level of contribution and motivating applications provides valuable information for those evaluating the research with respect to its potential impact and developing mitigation strategies for potential risks.

17 *NeurIPS 2021 Paper Checklist Guidelines*. (n.d.). NeurIPS 2021. Retrieved April 8, 2021, from https://neurips.cc/Conferences/2021/PaperInformation/PaperChecklist

Consideration of downstream consequences

We suggest researchers adopt a habit of reflecting on the downstream consequences of their research in all of their papers, whether required by the publication venue or not.[18] We understand many researchers may not feel well-equipped to anticipate the downstream consequences of their work, especially when it comes to the second- or third-order effects. That's okay; the exercise is not supposed to comprehensively capture all the ways that research could be harmful. Instead, the goal is to encourage more thoughtful engagement with these issues, and to exercise the muscle of anticipating the impacts of AI research so that it can gradually grow stronger. (See Appendix III for resources for getting started.) In some circumstances, this exercise may even surface interesting and socially beneficial research directions.

It is likely that for research on similar techniques or in similar domains, a lot of the potential downstream consequences will be shared. For example, all researchers working on enhancing the accuracy of facial recognition systems will have in common the risk of those systems being deployed for questionable surveillance purposes, even if the techniques and approaches to the research are quite different. Focusing only on these kinds of shared, broad risks could result in discussions of downstream consequences being generic and superficial. For this reason, we recommend that researchers focus their reflection on the implications of the marginal capability increase enabled by their contribution, while citing other work that contains a thorough discussion of the downstream consequences of the underlying technology or application. We suggest a rule of thumb that the level of original reflection should be proportional to the contribution of the research. For incremental advances, a short statement citing work that discusses the consequences of similar research in more detail may be sufficient. For more significant advances, a more substantial discussion is warranted.

The amount of computational resources used

There are a number of reasons why it would be beneficial to have a norm of reporting the computation used in research projects. Firstly, from a research integrity perspective, it is an important part of the methodology and is relevant to those who might want to reproduce the research. Secondly, the environmental effects of training large ML models is a key component of a research project's broader impacts (both in terms of the computation used to conduct the research and the potential computation required if the research was applied in practice). Finally, and perhaps most relevant to the concept of downstream consequences, it gives an indication of the resources required for others to train models of similar capability, which is a useful factor to consider when modeling potential threats from malicious actors.

2. Normalize discussion about the downstream consequences of research

We've found that while some researchers are keen to engage in conversations about the potential downstream consequences of their research, it's currently not a standard part of the culture at most research institutions. Ideally these discussions should happen throughout the research process (not just in preparation for publication), especially for research that could be particularly high stakes.

18 The potential benefits of this kind of reflection are enumerated in Prunkl, C. E. A., Ashurst, C., Anderljung, M., Webb, H., Leike, J., & Dafoe, A. (2021). Institutionalizing ethics in AI through broader impact requirements. *Nature Machine Intelligence, 3*(2), 104–110. https://doi.org/10.1038/s42256-021-00298-y

Researchers themselves are in a position to help shape this culture. Here, we present some ideas for how to do this:[19]

- Host a seminar discussion on the potential downstream consequences of your work.
- Use the resources in Appendix III to engage in techniques such as design fiction and critical design to think creatively about possible consequences.
- Build relationships and collaborations with peers in other disciplines, who can bring a new perspective to the potential impacts of your research.
- Ask your peers and your advisor or manager about what they see as the potential downstream consequences or applications of your research.

A culture of openly discussing potential downstream consequences during the research process may open up interesting new research questions, identify opportunities to rework research methods and outputs to be less risky overall, and increase the chance that the research will have a positive impact on society.[20]

19 Some advice on making these conversations productive can be found here: Bruckman, A. (2020). "Have you thought about...": Talking about ethical implications of research. *Communications of the ACM, 63(9),* 38–40. https://cacm.acm.org/magazines/2020/9/246935-have-you-thought-about/fulltext

20 Some researchers have raised concerns that a more public discussion of risks could (by providing ideas and templates for malicious actors) make it easier for tech to be misused. We discuss this further in the section on "Risks and potential downstream consequences of these recommendations."

Recommendations for research leadership

Leaders across industry and academia (including principal investigators, managers, decision-makers, and others in senior positions) have a key part to play in fostering a responsible research culture and helping to promote beneficial societal outcomes. Actors in these roles can provide the institutional support that encourages the anticipation and mitigation of negative downstream consequences. Accordingly, leaders can also discourage such behavior, bearing a greater responsibility when unidentified risks reflect poorly on their organizations and teams.

Companies and institutions have sparked a number of controversies by releasing AI research that had harmful outputs. In 2016, Microsoft suspended its Tay Twitter chatbot, which quickly picked up offensive and abusive language from other users, less than 24 hours after it launched.[21] Since then, Natural Language Processing models have been criticized for producing toxic outputs[22] and potentially automating the spread of misinformation.[23] In 2020, one medical chatbot reportedly encouraged a simulated patient to take their own life.[24] Such incidents aren't just a problem for corporate institutions: MIT recently removed an image-labeling dataset they built that contained derogatory slurs, issuing an apology.[25] It's possible that some of these incidents could be avoided or mitigated by thinking more thoroughly in advance about possible risks.

While we explore several interventions at the point of publication in this report, anticipating potential downstream consequences earlier in the research process makes it possible to amend plans so that they can be pursued in a more socially responsible way, and avoid difficulties later when it comes to publication. If research teams can mitigate potential risks in advance, conferences and journals should rarely need to withhold the publication of papers based on concerns about downstream consequences.

3. Review potential downstream consequences earlier in the research pipeline

To support responsible AI and avert future contention, we recommend research leaders build in opportunities for their teams to consider downstream consequences and harm mitigation strategies at various points in the research process, including when initially formulating research ideas. This can take different levels of formality, from developing a standardized review and approval process, to simply initiating regular, intentional discussions with researchers about the potential downstream consequences of their work. In general, for research that could have high-stakes applications, or is likely to significantly advance the current state of the art, a more thorough review of the potential consequences is warranted.

21 Schwartz, O. (2019, November 25). In 2016, *Microsoft's racist chatbot revealed the dangers of online conversation.* IEEE Spectrum: Technology, Engineering, and Science News. https://spectrum.ieee.org/tech-talk/artificial-intelligence/machine-learning/in-2016-microsofts-racist-chat bot-revealed-the-dangers-of-online-conversation

22 Strickland, E. (2021, February 1). *OpenAI's GPT-3 Speaks! (Kindly Disregard Toxic Language).* IEEE Spectrum: Technology, Engineering, and Science News. https://spectrum.ieee.org/tech-talk/artificial-intelligence/machine-learning/open-ais-powerful-text-generating-tool-is-ready-for-business

23 Hsu, J. (2019, November 4). *Microsoft's AI research draws controversy over possible disinformation use.* IEEE Spectrum: Technology, Engineering, and Science News. https://spectrum.ieee.org/tech-talk/artificial-intelligence/machine-learning/microsofts-ai-research-drawscontroversy-over-possible-disinformation-use

24 Quach, K. (2020, October 29). *Researchers made an OpenAI GPT-3 medical chatbot as an experiment. It told a mock patient to kill themselves.* https://www.theregister.com/2020/10/28/gpt3_medical_chatbot_experiment/

25 Quach, K. (2020, July 1). *MIT apologizes, permanently pulls offline huge dataset that taught AI systems to use racist, misogynistic slurs.* https://www.theregister.com/2020/07/01/mit_dataset_removed/

There are examples of analogous review processes in the scientific research community at large. "Institutional Review Boards" (IRBs) are a regulatory mechanism in the US to ensure research funded by the federal government is conducted ethically and the welfare of participants is protected. Other countries have similar institutions, though they may go by different names, such as "ethics review boards." There are, however, some limitations with IRBs as a mechanism for mitigating downstream consequences of AI research, as described in Appendix III. On the other side of the Atlantic, the EU's Horizon 2020 "Responsible Research and Innovation" process[26] encourages reflection on societal impacts and downstream consequences, along with other practices more focused on research ethics and integrity.

Some companies doing AI research have already implemented their own internal review processes, though since the details are rarely made public[27] it is often not clear the extent to which downstream societal consequences are explicitly considered, and at what point in the research process the reviews occur. Outside of IRBs, in academia there generally are not processes in place for reviewing and mitigating research risks. We recommend that AI research leaders (in both industry and academia) start establishing a process that considers downstream consequences and encourages discussion between researchers. We provide the following suggestions as a starting point:

- Start asking the researchers you oversee about the potential downstream consequences of their work, especially when early-stage research ideas are being explored.[28]
- Hold seminar discussions where researchers are asked to present their research and discuss potential downstream consequences, receiving feedback from peers.[29]
- Consider using techniques that encourage creative thinking and help identify externalities that might otherwise get overlooked. See Appendix III for resources to help with this.
- Build cross-disciplinary relationships with experts in other fields who can provide an alternative perspective on the potential impacts of AI research. If possible, consider bringing a social scientist onto your team, as recommended by Hecht, 2020.[30]
- Try to get input from underrepresented groups, especially vulnerable or impacted communities. This could involve utilizing initiatives such as Diverse Voices[31] and other focus group methods, as well as ensuring your research team is itself diverse.
- If any concerning risks are surfaced, work with the researchers to see if the research direction can be amended to avoid or mitigate them. For particularly challenging cases, it may also be worth contacting or employing people with expertise in ethics for advice. In the section on "Outstanding challenges and further work," we propose a mechanism to make this process easier.
- Ideally, publicly discuss the steps and techniques involved in your review process, share any lessons learned, and learn from others doing the same.

26 *Responsible research and innovation in practice*. (n.d.). Retrieved April 8, 2021, from https://www.rri-practice.eu/

27 Though it's possible to find some examples that are laid out at a high level, such as: *Review process*. (n.d.). Google AI. Retrieved April 8, 2021, from https://ai.google/responsibilities/review-process/

28 Some advice on making these conversations productive can be found here: Bruckman, A. (2020). "Have you thought about...": Talking about ethical implications of research. *Communications of the ACM, 63*(9), 38–40. https://cacm.acm.org/magazines/2020/9/246935-have-you-thought-about/fulltext

29 To make this more engaging, organizations could use a "red-team" dynamic where some participants are tasked with thinking creatively about all the possible ways the research could impact the world and others are tasked with coming up with creative mitigation strategies for any harms.

30 Hecht, B. (2020, December 13). *Suggestions for writing NeurIPS 2020 broader impacts statements*. Medium. https://brenthecht.medium.com/suggestions-for-writing-neurips-2020-broader-impacts-statements-121da1b765bf

31 Diverse Voices | Tech Policy Lab. (n.d.). Retrieved April 8, 2021, from https://techpolicylab.uw.edu/project/diverse-voices/

While such a process lacks the rigor of more formal mechanisms like IRBs, it is also less burdensome and more flexible. It might not catch every risk at first, but it provides an opportunity to practice the skills needed to anticipate and mitigate negative downstream consequences, while also surfacing implementation challenges that can inform the design of more formal accountability mechanisms.

4. Commend researchers who identify negative downstream consequences

Researchers and institutions may sometimes feel wary of drawing attention to the potential risks or negative impacts of their work or that of their colleagues. They may be worried about losing funding or revenue (if grantmakers or clients are put off), missing out on publications (if their paper is rejected on ethical grounds or if their paper is held up in a review process and they are scooped by another researcher), or the reputational risks of a public backlash.

However, the AI field will only build its capacity to anticipate and mitigate risks if researchers can talk openly about the downsides of AI research without fear of being penalized. Research leaders have an opportunity to set a positive example by supporting and commending individuals who call attention to negative downstream consequences, recognizing this as a valuable role.

As well as potentially opening up interesting new research opportunities, identifying risks early can open the door to harm mitigation steps and provide an opportunity to address risks before they become major issues. This is better not only for society, but also for individual researchers and institutions. Failing to identify risks early on might force these groups to withdraw or course correct their research at a later stage, which is likely to be more difficult, or potentially face public backlash. In addition, we suspect increased transparency and openness on the downsides of AI research will be better for the AI ecosystem as a whole and could improve trust in organizations that make good-faith efforts to engage with these issues.

Recommendations for conferences and journals

As previously discussed, AI research is disseminated and deployed in a variety of venues outside of the traditional publication process. Nevertheless, conferences and journals remain the primary venues for the formal presentation of findings. Researchers are still highly incentivized to pursue formal publication; it leads to jobs, tenure, grants, funding, citations, and prestige. As such, this makes the academic publication process a key intervention point and it makes conferences and journals especially capable of influencing responsible research norms through their policies. Thoughtful and consistent policies will not only contribute to building a research culture that includes the consideration of downstream consequences, but will also minimize the reputational risks incurred by publishing research widely regarded as problematic.

In June 2020, a large collective of AI researchers and experts wrote an open letter to Springer Publishing to contest a paper which claimed to predict criminality based on facial photos.[32] (The previous month, a university press release said the paper would soon be published by Springer—the publisher later stated that the paper was merely under consideration for a conference and associated proceedings series and ultimately rejected.)[33] The open letter called for the alleged offer for publication to be rescinded, all publishers to refrain from publishing similar studies, and Springer to "issue a statement condemning the use of criminal justice statistics to predict criminality, and acknowledging their role in incentivizing such harmful scholarship in the past." This response demonstrates how a lack of consideration for downstream consequences can incur reputational risks. Further examples of how the lack of clearly defined ethical review protocols can lead to problems for publication venues can be found in Prunkl et al., 2021.[34]

By requiring paper submissions to engage thoughtfully with downstream consequences, conferences and journals can contribute significantly to responsible AI research. At the same time, this must be implemented in a manner that ultimately promotes open and honest discussion.

5 . Expand peer review criteria to include engagement with potential downstream consequences

The peer review process, which leverages the unique expertise of specific research communities, is an invaluable tool for mitigating potential negative downstream consequences. Designed to assess whether research is suitable for publication, the peer review process is ultimately a form of self-regulation. It depends on independent experts with relevant competencies evaluating submitted works in accordance with publishers' guidelines. Common criteria provided by publishers include validity, novelty, and significance.

We suggest that publishers additionally integrate engagement with downstream consequences into their existing peer review processes. Depending on the publication venue, this could mean requiring the inclusion of a "broader impact statement," requesting this information be included as supplementary material rather than in the paper itself, or simply expecting authors to embed these considerations somewhere within or throughout their paper.[35] Just as the NeurIPS 2020 Broader Impact Statement did not count towards the submission's page limit, publishers may need to consider whether it is appropriate to be more flexible about the length of papers to accommodate this additional information.

32 Coalition for critical technology. (2020, June 29). Abolish the *#techtoprisonpipeline.* Medium. https://medium.com/@CoalitionForCriticalTechnology/abolish-the-techtoprisonpipeline-9b5b14366b16

33 https://twitter.com/SpringerNature/status/1275818113410510849

34 Prunkl, C. E. A., Ashurst, C., Anderljung, M., Webb, H., Leike, J., & Dafoe, A. (2021). Institutionalizing ethics in AI through broader impact requirements. *Nature Machine Intelligence, 3*(2), 104–110. https://doi.org/10.1038/s42256-021-00298-y

35 This builds on recommendations proposed by: It's time to do something: Mitigating the negative impacts of computing through a change to the peer review process. (2018, March 29). *ACM FCA*. https://acm-fca.org/2018/03/29/negativeimpacts/

Reviewers should aim to assess whether the author has sufficiently engaged with the potential downstream consequences of their research (which may include criteria such as breadth, depth, quality, relevance, and appropriate citations). This can be thought of as similar to assessing whether an author has sufficiently engaged with prior literature. In cases where the discussion is not sufficient, this should inform the reviewer's assessment of the work and be provided as feedback to the authors. Note that this process does not require reviewers to make ethical judgement calls about whether the potential risks of the research warrant not publishing it: this should be a separate process, as described in the next recommendation.

Not all research requires the same level of engagement with downstream consequences. There are a number of factors that may influence the expected engagement, including the proximity of the research to applications, and whether those applications are in high-stakes domains.[36] We believe that one of the most important factors, however, is the level of contribution of the paper. For incremental advances, a lightweight statement that cites other work discussing similar downstream consequences might be sufficient. For major new breakthroughs, or work that involves novel applications, more substantial discussion is warranted. In general, the expected level of engagement with downstream consequences should be proportional to the contribution a paper claims to make.[37]

Since it has only recently been asked for, evaluating whether an author has sufficiently engaged with downstream consequences may feel challenging for reviewers at first. Over time, however, we are optimistic that norms will emerge and reviewers will become more confident in their ability to make consistent judgements. We echo the calls by Prunkl et al., 2021[38] to provide guidance and example statements (appropriate to the publication venue) to help both researchers and reviewers in this task, and see the new NeurIPS Paper Checklist[39] as a positive step towards providing more structure for researchers and reviewers.

6. Establish a separate review process to evaluate papers based on risk and downstream consequences

Ideally, any major risks, downstream consequences, or ethical issues connected to AI research should be addressed before a paper is submitted for publication. However, this may not always be the case — especially since the AI field is in the early stages of developing responsible research practices. Journals and conferences may find themselves in the position of needing to evaluate whether a paper carries significant potential harms that warrant withholding publication.

36 For example, a new gradient descent technique is so removed from applications that the possible consequences are combinatorially many, whereas the prediction of socio-economic behaviors is much more targeted in its potential applications, many of which could be high-stakes.

37 As discussed in the Recommendations for Individual researchers, we emphasize this factor in particular because focusing on the marginal capability increase enabled by their contribution reduces the need for researchers to simply restate generic risks that apply to all research on similar techniques or in similar domains, which will likely have been covered elsewhere (though they should cite this prior work). This is not to suggest that a small advance can't lead to sizable effects: an incremental improvement in a technology that is well-known to have harmful applications could push it past a tipping point that leads to its widespread application. However, we believe this is better addressed via a separate review process (as discussed in the next section) that can deliberate on mitigation strategies (such as delayed or withholding publication), since if the harms are already well-known there is not likely much to be gained from simply asking the authors to re-enumerate them. However, if the contribution of the research involves making this technology vastly more scalable than was previously possible, it is very useful to have the researchers reflect on the implications of that aspect in particular.

38 Prunkl, C. E. A., Ashurst, C., Anderljung, M., Webb, H., Leike, J., & Dafoe, A. (2021). Institutionalizing ethics in AI through broader impact requirements. *Nature Machine Intelligence, 3*(2), 104–110. https://doi.org/10.1038/s42256-021-00298-y

39 Conference, N. I. P. S. (2021, March 26). *Introducing the NeurIPS 2021 Paper Checklist.* Medium. https://neuripsconf.medium.com/introducing-the-neurips-2021-paper-checklist-3220d6df500b

The peer review process is designed to assess merit and rigor, which may include the rigor of engagement with downstream consequences. Evaluating the downstream consequences themselves and strategizing about risk mitigation is a different task that requires a different set of expertise. We therefore recommend establishing a separate review process for this that includes a diverse, multidisciplinary set of experts. For papers that are flagged as posing a significant risk but otherwise meet the bar for acceptance, mitigation strategies should be considered to see if there are ways to publish the work in some form.[40] While this cannot always be guaranteed, risks could be mitigated by, for example, redacting parts of the research to prevent malicious use, using a staged release approach, or restricting access to code to vetted individuals.[41]

We also strongly caution against basing the evaluation solely on the severity of risks identified by the author in their discussion of downstream consequences. This would be counterproductive to the goals of encouraging open and frank discussions about the broader impacts of AI research, as authors may be tempted to minimize and conceal significant risks for fear of rejection or other sanctions. The goal is to reward authors that do a good job of comprehensively identifying potential risks, not penalize them. In addition, reviewers should not assume that the author's discussion of downstream consequences contains all the information necessary to make an informed judgement call about the risk level of the paper; there will likely be other factors that influence this decision.

The variety of publishing venues in AI (which include conferences, workshops, journals, and more) makes it difficult to provide recommendations for universal editorial policies on this issue. We encourage publishing venues to engage with the wider AI community on these matters, and to consider how various roles (such as editors, program chairs, referees, and ethics experts) might best contribute to or oversee this review process while being mindful of the considerations discussed above. An additional challenge is the lack of consensus on what constitutes risky research, which we discuss further in the section on "Categorizing AI research by risk level."

40 This could be particularly important for papers that demonstrate AI methods for attacks or malicious use. Researchers need to be able to warn the community of vulnerabilities and dangers without inadvertently providing the means for malicious users to easily execute attacks.

41 Ovadya, A., & Whittlestone, J. (2019). Reducing malicious use of synthetic media research: Considerations and potential release practices for machine learning. *ArXiv:1907.11274 [Cs].* http://arxiv.org/abs/1907.11274

Outstanding challenges and further work

As an initial step towards a responsible research culture that considers downstream consequences, this white paper has recommended contributions three groups in the AI research community can make. However, there is more work to be done. Here, we highlight some of the remaining challenges and suggest next steps. Some of these may constitute further work by PAI, though we additionally welcome efforts to address them from the community at large.

Providing resources to help researchers anticipate the potential downstream consequences of their work

This white paper includes recommendations that involve AI researchers attempting to anticipate the downstream consequences of their research. One of the challenges with this is that AI researchers often don't feel equipped to do this well.

We welcome efforts to provide guidance and make this process easier. As mentioned earlier, we endorse the recommendation from Prunkl et al., 2021[42] for publishers to provide examples of broader impact statements that researchers and reviewers can use as benchmarks. Other possible resources include comprehensive taxonomies of AI risks and concrete strategies for anticipating impacts, such as design-thinking exercises tailored to AI. See Appendix III for some existing guides and resources that may help.

42 Prunkl, C. E. A., Ashurst, C., Anderljung, M., Webb, H., Leike, J., & Dafoe, A. (2021). Institutionalizing ethics in AI through broader impact requirements. *Nature Machine Intelligence, 3*(2), 104–110. https://doi.org/10.1038/s42256-021-00298-y

Improving access to experts and reviewers who are able to assess the potential downstream consequences of AI research

The AI peer review system is already stretched thin and some of the recommendations in this white paper involve placing additional asks on reviewers or finding reviewers with particular kinds of expertise to help assess risks. But at the moment, researchers are generally incentivized to prioritize writing papers over reviewing them. Until a larger proportion of the community feels able to dedicate more time to reviewing, and more confident in their ability to consider potential downstream consequences, it will likely be difficult to find enough suitable reviewers. Additionally, it may be necessary to consult experts from other fields to get an accurate picture of downstream consequences.

A database of experts who can act as reviewers or provide input on potential downstream consequences would make this easier. This could include AI researchers, ethicists, and experts from other relevant fields. This may go some way to alleviating the challenge of sourcing qualified reviewers and also act as a resource for researchers or institutions which would value additional perspectives on downstream consequences earlier in the research process. One challenge to consider with this idea is that, in general, reviewers are not directly compensated for their time, though as researchers it is often considered part of their job. If the pool of experts is to include a diverse variety of contributors, it's possible that they won't all have the institutional resources or funding to take on this additional work.

Creating advisory bodies for AI research

The National Science Advisory Board for Biosecurity[43] (NSABB) in the US is tasked with providing "advice, guidance, and recommendations regarding biosecurity oversight of dual use research, defined as biological research with legitimate scientific purpose that may be misused to pose a biologic threat to public health and/or national security."

As described earlier in the report, AI can also be considered "dual-use research," since it has both socially beneficial and harmful applications. The NSABB has advised on publication decisions for life sciences research and also produces guidance on topics such as "Enhancing Personnel Reliability and Strengthening the Culture of Responsibility".[44]

Given the parallels with AI, we encourage the AI community to explore whether a similar body might be appropriate and how one tailored to the field of AI might need to differ. The NSABB was established due to national security implications of 9/11 and the 2001 Anthrax attacks; it would be preferable not to wait for an equivalent trigger event for AI.

There are important limitations to the NSABB model. It is a US body, so to handle the global nature of AI there would need to be equivalent institutions in other countries. Additionally, the ultimate efficacy of such an advisory body is unclear: In one case, the NSABB recommended redacting parts of some research but later reversed this decision after a revised manuscript was submitted,[45] leaving it uncertain whether the initial recommendation would have been adhered to.

These limitations demonstrate why further discussion is necessary to determine if and how advisory bodies might play a role in mitigating negative downstream consequences of AI research.

43 National Science Advisory Board for Biosecurity (NSABB). (n.d.). *Office of Science Policy.* Retrieved April 8, 2021, from https://osp.od.nih.gov/biotechnology/national-science-advisory-board-for-biosecurity-nsabb/

44 Guidance for enhancing personnel reliability and strengthening the culture of responsibility. (2011). National Science Advisory Board for Biosecurity. https://osp.od.nih.gov/wp-content/uploads/2013/06/CRWG_Report_final.pdf

45 Partnership on AI. (2020, December 9). *What the AI community can learn from sneezing ferrets and a mutant virus debate.* Medium. https://medium.com/partnership-on-ai/lessons-for-the-ai-community-from-the-h5n1-controversy-32432438a82e

Categorizing AI research by risk level

One of the challenges with the recommendations in this white paper is the subjectivity involved in making judgement calls about the possible risks posed by a piece of research. A more robust approach might involve categorizing AI research into risk levels[46] similar to the "biosafety levels" used in life sciences research.

Factors that might place research at a higher risk level include:

- techniques that make certain activities vastly more scalable;
- research that implicates vulnerable communities or involves protected characteristics;
- research that represents a significant advance in a contentious domain;
- research that radically departs from common ML architectures or represents a paradigm shift;
- research that enables the impersonation of humans or human capabilities; and
- research with applications that could infringe on human rights, civil liberties, or otherwise degrade our institutions.

This incomplete list is not to say that such research is inherently problematic, just that it should be subject to additional scrutiny to mitigate risks.

Clarity around what kind of research is considered high-or low-risk would be valuable for a number of reasons. It would help funders make more conscious decisions about the types of research they want to fund. It would help publishers determine which papers might warrant additional review or redactions.[47] And it would help researchers better understand their specific obligations with respect to downstream consequences.

46 Initial approaches to this include:
- Jordan, S. (2019). *Designing an Artificial Intelligence research review committee* (pp. 16–17). Future of Privacy Forum. https://fpf.org/wp-content/uploads/2019/10/DesigningAIResearchReviewCommittee.pdf
- Perset, K., Murdick, D., Clark, J., & Grobelnik, M. (2020, November). A first look at the OECD's Framework for the Classification of AI Systems, designed to give policymakers clarity. *OECD AI Policy Observatory.* https://www.oecd.ai/wonk/a-first-look-at-the-oecds-framework-for-the-classification-of-ai-systems-for-policymakers
- *White Paper on Artificial Intelligence: A European approach to excellence and trust.* (2020). European Commission. https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf

47 When journals developed ethical guidelines for stem cell research, they followed insights from an international stem cell society. The lack of consensus in the AI community on risks makes it difficult for publishing venues to develop consistent editorial policies.

Reconciling academic and industrial research norms

In AI, it's not uncommon for AI academics to hold dual positions, working for both universities and tech companies. Additionally, many paper authors and peer reviewers work in corporate AI labs.[48] Because of this, we have generally not made distinctions between academic and industry AI research in this report.

However, these circumstances can cause problems when the norms and expectations in one setting conflict with those of the other. For example, tech companies have been criticized[49] for not upholding traditional academic standards of openness in their publications, since they do not always release their data or code. This is further complicated by the fact that there may be legitimate concerns about downstream consequences that warrant withholding aspects of the research, but it is difficult to distinguish this from efforts to simply protect proprietary tech or valuable IP.

There is also concern about the influence of "big tech" money on academic research priorities, and possible conflicts of interest. Recent events in the community[50] have surfaced concerns about corporate influence on papers by industry researchers published in academic venues, and the extent to which industry researchers can expect to have academic freedom. An additional disclosure norm which requires authors to specify funding sources more explicitly and also report any corporate edits that the paper underwent may be useful here, though how the latter would be implemented and enforced requires further consideration.

These issues have implications not just for downstream consequences but also research integrity and ethics. We encourage the AI community to reflect further on the issue of reconciling academic and industrial research norms.

48 Hagendorff, T., & Meding, K. (2020). Ethical considerations and statistical analysis of industry involvement in machine learning research. *ArXiv:2006.04541 [Cs].* http://arxiv.org/abs/2006.04541

49 AI is wrestling with a replication crisis. (2020, November 12). MIT Technology Review. https://www.technologyreview.com/2020/11/12/1011944/artificial-intelligence-replication-crisis-sciencebig-tech-google-deepmind-facebook-openai/

50 Hutson, M. (2021, February 15). *Who should stop unethical A.I.?* The New Yorker. https://www.newyorker.com/tech/annals-of-technology/who-should-stop-unethical-ai

Developing common protocols for responsible product deployment

This white paper has primarily focused on research publication, but in AI (in part due to the role of industry research and the pace of development) there is often no clear line between "research dissemination" and "product deployment."

There is a lot more work to be done in addressing the product end of this spectrum. In particular, what safety protocols, processes, standards, and regulation are required for the responsible deployment of AI systems? Many traditional high-stakes engineering fields (such as the aviation industry)[51] have clearly established safety standards, risk analysis methodology, and accountability mechanisms. The AI industry currently lacks similarly established guidelines.[52]

In our recommendation for research leaders on reviewing downstream consequences earlier in the research process, we mentioned the value of organizations openly sharing their review processes and lessons learned. Similarly, we encourage companies deploying AI products to collaborate with the wider community to help pilot and iterate on responsible deployment protocols with the aim of creating more effective and standardized processes.

Because the field lacks common norms for responsible publication and deployment, there are cases where the actions of one organization appear to be in conflict with another's.[53] There is clearly a need for more community coordination on these issues.

51 Hunt, W. (2020). *The Flight to Safety-Critical AI: Lessons in AI safety from the aviation industry*. UC Berkeley Center for Long-Term Cybersecurity. https://cltc.berkeley.edu/wp-content/uploads/2020/08/Flight-to-Safety-Critical-AI.pdf

52 Though we welcome emerging efforts to address this such as:
- Cihon, P. (2019). *Standards for AI governance: International standards to enable global coordination in AI research & development*. Center for the Governance of AI, Future of Humanity Institute, University of Oxford. http://www.fhi.ox.ac.uk/wp-content/uploads/Standards_-FHI-Technical-Report.pdf
- Artificial intelligence standards by ISO/IEC JTC 1/SC 42. (n.d.). ISO. Retrieved April 8, 2021, from https://www.iso.org/committee/6794475/x/catalogue/,
- *AI standards*. (2019, March 14). National Institute of Standards and Technology (NIST). https://www.nist.gov/artificial-intelligence/ai-standards
- *Ethically aligned design*. (2016). The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v1.pdf
- *About ML*. (n.d.). The Partnership on AI. Retrieved April 8, 2021, from https://www.partnershiponai.org/about-ml/
- *P2840–Standard for responsible AI licensing*. (2019, May 9). IEEE Standards Association. https://standards.ieee.org/project/2840.html

53 https://twitter.com/jackclarkSF/status/1373774648874135554

Risks and potential downstream consequences of these recommendations

Any change in norms carries risks. In this section, we discuss some of the main risks and potential downstream consequences of the recommendations contained in this report.

Overall, these limitations highlight the need for these recommendations to be understood as a first step – and far from the last – towards developing responsible publication practices. Ultimately, they should be adjusted and updated as we learn more about how effective they are, what costs they impose, and how their limitations can be mitigated. By trialling and iterating on these recommendations, we expect to learn a lot about best practices for implementation, and how they can be tailored to better suit specific circumstances.

Opportunity cost

A significant risk of these recommendations is whether the time, effort, and resources they require outweighs the benefits.

Researchers are already under pressure to produce research and write it up in time for conference deadlines. Any time that they spend preparing the additional information recommended in this white paper is time they could otherwise have spent on other aspects of their research. These recommendations also place an additional burden on (already over-stretched) peer reviewers, which should also be taken into account. In the section on "Outstanding challenges and further work," we suggest that a database of reviewers might help mitigate this issue.

Resentment at poorly implemented interventions

Many researchers have had frustrating experiences with overly bureaucratic review boards. Some of the recommendations in this report, if poorly implemented, could feel like box-ticking exercises rather than things that meaningfully contribute to anticipating and mitigating negative downstream consequences. There is a risk that naive implementations of such policies could contribute to feelings of resentment in the AI community and potentially widen rifts between researchers with differing beliefs about ethical responsibility. This could not only thwart progress towards responsible AI butalso increase skepticism of effective mechanisms.

Examples of pitfalls that could provoke resentment include:

- Policies that impose disproportionate costs on researchers compared to the likely societal benefit (see the previous section on "Opportunity cost of additional overhead").
- Review processes that move too slowly compared to the pace of AI research.
- A lack of guidance (and examples) enabling researchers to effectively fulfill the requirements being asked of them.
- A lack of transparency and clarity around how decisions are being made, resulting in a perception of unfairness, arbitrariness, or ideological influence.
- Superficial policies that can easily be "gamed" to meet the "letter of the law" without engaging with the spirit.
- A lack of buy-in from senior leaders which undermines the credibility of the policies.

False sense of security

Another risk of the recommendations in this white paper is that they could contribute to a false sense of security. Anticipating downstream consequences is very difficult and even the most well-intentioned researchers or processes are unlikely to identify all of them. If researchers and reviewers believe that they have comprehensively addressed all possible risks, but have actually overlooked something significant, this could potentially be even more dangerous (especially if there are incentives to mask or reframe certain kinds of risks, as discussed in the next section). This is a particular concern if the process is seen as a "one-shot" effort at evaluation, rather than a continuous reflection that evolves both with the research itself and with our capacity as a community to more accurately anticipate impacts.

Perverse incentives and unintended consequences

In our recommendations, we have cautioned against using the author's enumeration of the risks of their research as the sole input when deciding whether a paper is too harmful to publish. This is because it could create perverse incentives for researchers to try to mask the more serious risks that could cause their paper to get rejected. It's also possible that companies might try to distract from risks that relate to their core business practices by focusing on less problematic issues. Both of these situations could act against the intention of the recommendations, which is to encourage a more frank and open discussion about downstream consequences.

It's also possible that emphasizing the rule of thumb "the greater the contribution of a research paper, the greater the responsibility to consider its impacts" could incentivize researchers to downplay the level of their contribution, to relieve themselves of the obligation to engage with the risks.

There is also the possibility that asking authors to reflect on the potential misuses of their research could result in the unintended consequence of providing a "roadmap" for malicious actors who want to use AI to cause harm. In this case, publishing a discussion of downstream consequences could constitute an "infohazard",[54] in that sharing that information is, itself, a risky act. This could potentially be mitigated by redacting parts of the discussion before publication (though still requiring it as part of the review process).

Additionally, there may be other ways these recommendations create perverse incentives or other unintended consequences that we haven't anticipated.

54 "Infohazard," or "informational hazard," is a term for "risks that arise from the dissemination or the potential dissemination of true information that may cause harm or enable some agent to cause harm." (Bostrom, N. (2011). Information hazards: A typology of potential harms from knowledge. *Review of Contemporary Philosophy,* 10, 44–79.)

Curtailing scientific inquiry and openness

The emphasis on considering downstream consequences may result in researchers feeling pressure to only pursue research with clear societal benefits. However, many of the advantages society enjoys today came out of fundamental research with no obvious beneficial applications. Review processes focused on downstream consequences (either within research institutions, funders, or publishers) could be seen by researchers as limiting their free inquiry. There's also a risk that harmless (and potentially beneficial) research could get mischaracterized, and end up being unfairly rejected or discarded.

In this report, we have primarily focused on increasing openness through greater disclosure. For most research, the "responsible" approach is to publish it as widely as possible, so that society can enjoy the beneficial applications and any potential harms can be uncovered, studied, and mitigated. But there are some interventions (such as redacting sensitive parts of research, or adopting a staged release approach) that arguably go against scientific norms of openness. This is a trade-off that will need to be continually assessed; such interventions should be used in rare cases where the risks outweigh the benefits of open publication.

Subjectivity and cultural differences

Ultimately, many of this report's recommendations rely on subjective judgements about potential impacts and risks. That's not to say that the impacts themselves are subjective, but that the valence people place on them, and how they weigh them against other factors, may vary. Individuals differ in preferences, values, ideology, and risk tolerance, and this may lead to significant differences in opinion about what papers are too risky to publish in full.[55]

Additionally, different cultures may have different ethical frameworks and value-systems. When asked to consider downstream consequences, the things that seem salient to different people could vary greatly. Furthermore, some nuanced ethical concepts don't translate easily into other languages. Since most AI research papers are written in English, this could make it hard for non-native English speakers to clearly express their perspective.

One way to mitigate these challenges is to support more forums during conferences and workshops for exploring and building understanding about these differences and disagreements. Where appropriate, for venues such as workshops, another approach could be to encourage a variety of formats to represent engagement with downstream consequences, such as vignettes, stories, parables, even references to literary or artistic works.[56]

That said, significant strides have been made under international law to determine what's ethically and legally unacceptable – determinations that bear relevance to any assessment of downstream consequences of AI research. We can look to peremptory norms such as the International Bill of Human Rights and the prohibition of "crimes against humanity"[57] while also taking note of ongoing rule-making, such as the necessity of businesses to undergo human rights due diligence in scientific research.[58] For instance, assessments could include reviewing downstream impacts against the rights set out in the Universal Declaration of Human Rights[59] and identifying the vulnerable populations most likely to be impacted.

55 In any discussion of "responsible" or "ethical" AI, questions like "Who decides what's responsible?" and "Ethical by whose values?" inevitably arise. Moral disagreements in philosophy have not been solved after millenia, yet humanity has found ways to build shared systems, institutions, and agreements despite our different ethical intuitions. There have already been some explorations of this in the context of AI. In the same way that the peer review process functions even though individual reviewers don't always agree on which papers meet the bar for acceptance, we are optimistic that we can make progress on responsible AI practices without needing to define in advance exactly what research should and shouldn't be considered ethical.

56 An example of an event that accepted submissions in a variety of formats is the NeurIPS 2020 Resistance AI Workshop.

57 *United nations office on genocide prevention and the responsibility to protect.* (n.d.). Retrieved April 8, 2021, from https://www.un.org/en/genocideprevention/crimes-against-humanity.shtml

58 *General comment No. 25 on science and economic, social and cultural rights.* (2020). United Nations Economic and Social Council. https://undocs.org/E/C.12/GC/25

59 Nations, U. (n.d.). *Universal declaration of human rights.* United Nations. Retrieved April 8, 2021, from https://www.un.org/en/about-us/universal-declaration-of-human-rights

Acknowledgments

We'd like to thank the following people for their significant contributions to the production of this white paper:

Grace Abuhamad (ServiceNow), Emily Jarratt (Centre for Data Ethics and Innovation), Bran Knowles (Data Science Institute, Lancaster University), Joseph Lindley (Design Research Works and Imagination, Lancaster University), Aviv Ovadya (Thoughtful Technology Project), Hannah Quay-de la Vallee (Center for Democracy and Technology), Sara R. Jordan (Future of Privacy Forum), Liesbeth Venema (Nature Machine Intelligence, Springer Nature), Ben Zevenbergen (Google)

And the following people for providing feedback and insights that helped shape the ideas presented:

Dunstan Allison-Hope (Business for Social Responsibility), Markus Anderljung (Centre for the Governance of AI, Future of Humanity Institute), Carolyn Ashurst (University of Oxford), Solon Barocas (Microsoft Research and Cornell University), Haydn Belfield (Centre for the Study of Existential Risk, University of Cambridge), Y-Lan Boureau (Facebook AI Research), Miles Brundage (OpenAI), Will Carter (Google), Heather Douglas (Department of Philosophy, Michigan State University), Iason Gabriel (DeepMind), Brian Patrick Green (Markkula Center for Applied Ethics, Santa Clara University), Brent Hecht (Northwestern University and Microsoft), Reena Jana (Google), Jessica Newman (UC Berkeley AI Security Initiative), Carina Prunkl (Institute for Ethics in AI, University of Oxford), Inioluwa Deborah Raji (Mozilla Foundation), Stuart Russell (Center for Human-Compatible AI, UC Berkeley), Toby Shevlane (University of Oxford), Hanna Wallach (Microsoft Research), Jasmine Wang (Independent)

While individuals representing many of PAI's Partner organizations contributed text or suggestions to this document, it should not be read as representing the views of any specific member of the Partnership. Additionally, contributions from individuals do not necessarily reflect the views of their employers.

Many PAI staff contributed directly and indirectly to this work. In particular, Rosie Campbell led the project and writing, Madhulika Srikumar provided feedback, and Hudson Hongo was the editor.

Appendices

Appendix I: The role of the research community in navigating downstream consequences

This white paper rests on the underlying assumption that there is a role (and an accompanying responsibility) for the AI research community to play in considering the downstream consequences of its outputs.

It can be tempting for AI researchers to believe their sole duty to be expanding the capabilities of technology, with the societal impacts of their work lying outside their remit. As members of society, however, we cannot completely dissociate the impacts of our work from our moral obligations to that society, especially in realms where we have unique domain expertise.[60] Furthermore, the (public and private) resources devoted to AI research are premised on its possible (scientific and commercial) impacts, making the study and consequences of AI inherently entwined.

In an ideal world, advancing knowledge and the capabilities of technology might well only have positive effects. But in one of both accidents and malicious actors – where tools made possible by research can be misused – we must consider the potential harmful consequences as well.

In other high-stakes, dual-use fields such as biosecurity,[61] the idea that the research community bears some responsibility towards impacts is widely accepted, and there are norms, processes, and institutions aimed at upholding responsible research practices.[62] Additionally, it is not unusual for some occupations to have codes of professional ethics. In fact, scientists are already bound by certain obligations relating to research integrity and ethics (see Appendix II). It is therefore reasonable and necessary, in our view, to foster a culture where AI researchers understand how possible downstream consequences can play out and feel some stewardship over them.

As discussed in the introduction of this white paper, it is important to note that this is a shared burden that requires coordination across the variety of actors participating in the AI ecosystem on policies, processes, and incentive structures. Each part of the community has different mechanisms at their disposal to effect change and different domain expertise that can contribute to a responsible research culture.[63]

60 Douglas, H. E. (2003). The moral responsibilities of scientists (Tensions between autonomy and responsibility). A*merican Philosophical Quarterly*, 40(1), 59–68.

61 Partnership on AI. (2020, December 9). *What the AI community can learn from sneezing ferrets and a mutant virus debate*. Medium. https://medium.com/partnership-on-ai/lessons-for-the-ai-community-from-the-h5n1-controversy-324324 38a82e

62 Our thinking on the issues discussed in this white paper has been heavily influenced by how other high-stakes fields approach responsible research and deployment. Likewise, it's likely that the ideas discussed here may have relevance beyond the field of AI, though it is beyond the scope of this document to consider the nuances of how they might transfer.

63 An important factor to consider for effecting change in the AI community is regional influence. Since many researchers and conferences are based in North America and Europe, there is a particular opportunity and obligation for these regions to lead by example in developing responsible research and publication practices.

For instance, no one knows more about AI research than AI researchers themselves: their expertise, knowledge of limitations and potential applications, and ability to intervene early in the process are vital for helping to anticipate and mitigate negative consequences. At the same time, predicting the second- or third-order effects of research requires multidisciplinary expertise spanning many fields, and effective mitigation requires input and action from regulators, tech companies, publication venues, democratic deliberation, and other stakeholders. We therefore advocate for both "grassroots" norms-changes at the individual level and "top down" policy-based approaches at the institutional and leadership levels. If every actor in the research chain plays their part, the community will be more robust overall.

Appendix II: Disambiguating terms related to responsible research

When discussing responsible research practices, it is common for a number of different concepts to become conflated. Here, we try to disambiguate some of the different components, with the following caveats:

- The terms and definitions presented here are not yet commonly established, making it likely that others may use them slightly differently or make slightly different distinctions between them.
- This is by no means an exhaustive list of all the components involved in responsible research: a more thorough and nuanced taxonomy is welcomed.
- It is not always possible to entirely disentangle these different components. For example, one of the issues mentioned in the "Research culture" section is the increasing pace at which AI researchers are expected to publish. It's very possible that this puts pressure on researchers to publish before they've had a chance to thoroughly think through the downstream consequences of their work. Similarly, AI is commonly criticized for its lack of diversity, which could cause certain kinds of downstream consequences to be overlooked, particularly ones that impact underrepresented or vulnerable groups. This suggests that to successfully navigate downstream consequences in a way that maximizes benefits and minimizes harms, it may be necessary to address issues related to other aspects of responsible research practices.

Research integrity:

Research integrity is concerned with intellectual honesty and transparent reporting in the pursuit of truth, ultimately ensuring that findings are robust and replicable. This can include:

- not fabricating, falsifying, or misrepresenting data;
- providing the information necessary to reproduce the research;[64]
- using appropriate methodologies and analysis techniques (pre-registration, avoiding p-hacking, etc.);
- stating the assumptions under which the research holds, and describing limitations;
- contributing to open science practices such as open-sourcing code, providing open-access pre-prints, and using open-review platforms;
- following expectations and best practices when peer reviewing; and
- disclosing conflicts of interests and funding sources.

Research ethics:

Research ethics usually refers to principles surrounding how research is conducted. This may include the protection of the welfare of human participants, the responsible handling of data, and other considerations generally covered by IRBs. In the context of AI, relevant aspects of research ethics might include:

- ensuring training data is sourced responsibly and consensually;
- securely handling personal data so that privacy is protected;
- considering the welfare of laborers engaged in labeling data; and
- informing users of systems that are being used in studies.[65]

64 Gibney, E. (2019). This AI researcher is trying to ward off a reproducibility crisis. *Nature, 577*(7788), 14–14. https://doi.org/10.1038/d41586-019-03895-5

65 Meyer, R. (2014, June 28). Everything we know about facebook's secret mood manipulation experiment. The Atlantic. https://www.theatlantic.com/technology/archive/2014/06/everything-we-know-about-facebooks-secretmood-manipulation-experiment/373648/

Research culture:

Research culture, in this context, pertains to the issue of fostering a healthy and sustainable working environment that allows researchers and their ideas to flourish. This can include:

- producing and disseminating educational material to help "onboard" junior scientists;
- attracting a more diverse set of researchers and fostering an inclusive environment;
- improving incentive structures (e.g., reforming the "publish or perish" dynamics that many researchers experience);
- promoting productive interactions and eliminating toxic behaviors between researchers (especially when there is a seniority or power differential); and
- supporting intellectual exploration and productive disagreement.

For AI in particular, there are some additional concerns related to the relentless pace of ML publications:[66] the exploding popularity of the field has made it incredibly competitive, and researchers are under increasing pressure to publish their ideas as fast as possible to beat others to it and rack up their publication count. Not only is this taking a toll on the wellbeing of researchers (a research culture issue), it also means papers are taking a quality hit (a research integrity issue), and gives researchers less time to properly consider (and mitigate) the downstream consequences of their work.

Downstream consequences:

Downstream consequences are the ultimate effects of the research on society, usually as a result of the research being applied. The term includes direct consequences of the research, as well as the possible second- and third-order effects. Some downstream consequences are immediately apparent, whereas others may only become noticeable further in the future. Downstream consequences might include impacts on the environment, individuals, different groups, and society at large.

It's important to note that downstream consequences might impact different people in different ways, and so any given downstream consequence may be perceived as positive by some and negative by others. However, for practical purposes, when we refer to negative downstream consequences, we generally mean things that would widely be viewed as accidents, unintended consequences, inappropriate applications, malicious use, or other systemic harms. We often refer to these as "risks" for the sake of brevity.

66 Time to rethink the publication process in machine learning. (2020, February 27). *Yoshua Bengio*. https://yoshuabengio.org/2020/02/26/time-to-rethink-the-publication-process-in-machine-learning/

Broader impacts:

The term "broader impacts" is used by the National Science Foundation[67] (NSF) as part of their funding criteria. In this context, it is generally used to mean the potential benefits of the research – funding applicants are encouraged to explain how their proposal contributes to the mission of the NSF. The term was also used by NeurIPS as part of the 2020 requirement for authors to include a "broader impacts statement" in their papers. Unfortunately, this appears to have resulted in some confusion.[68] While the instruction was to include societal impacts, both positive and negative, some people interpreted it to mean something similar to the NSF usage of the term, that is, explaining how the research could benefit society. Others understood this to be a prompt to solely consider the potential risks and harms of the research, a concept closer to the idea of negative downstream consequences. And some people considered "broader impacts" to incorporate aspects of research integrity, research culture, and research ethics, as well as downstream consequences, which may be closest to what the organizers had in mind. The updated guidance for 2021 involves a Paper Checklist that spans across all of these areas, and uses the term "negative societal impacts" to refer to downstream consequences.

67 *Broader impacts review criterion.* (n.d.). US National Science Foundation. Retrieved April 8, 2021, from https://www.nsf.gov/pubs/2007/nsf07046/nsf07046.jsp

68 Prunkl, C. E. A., Ashurst, C., Anderljung, M., Webb, H., Leike, J., & Dafoe, A. (2021). Institutionalizing ethics in AI through broader impact requirements. Nature Machine Intelligence, 3(2), 104–110. https://doi.org/10.1038/s42256-021-00298-y

Appendix III: Resources for anticipating downstream consequences

A collection of resources that can serve as a starting point for review processes or researchers aiming to anticipate downstream consequences of AI research.[69]

- The project page for the PAI Publication Norms workstream contains some "Questions for thinking about Risk" designed to prompt creative thinking about downstream consequences. They can be found in the "Resources" section https://www.partnershiponai.org/publication-norms

- The Responsible Research and Innovation Toolkit contains a database of resources for responsible research, some of which may be helpful for anticipating downstream consequences https://rri-tools.eu/search-engine

- Researchers and reviewers may wish to refer to the Universal Declaration of Human Rights when anticipating downstream consequences, to check for potential infringements on human rights https://www.un.org/en/about-us/universal-declaration-of-human-rights

- The 2021 NeurIPS Paper Checklist includes prompts on a variety of aspects of responsible research, including negative societal impacts https://neurips.cc/Conferences/2021/PaperInformation/PaperChecklist

- Two guides for writing "broader impact statements":
  - https://medium.com/@GovAI/a-guide-to-writing-the-neurips-impact-statement-4293b723f832
  - https://brenthecht.medium.com/suggestions-for-writing-neurips-2020-broader-impacts-statements-121da1b765bf

- DotEveryone's Consequence Scanning Kit is an interactive approach to anticipating research impacts https://doteveryone.org.uk/project/consequence-scanning/

- The Markkula Center's "Ethics in Technology Practice" materials constitute a set of resources for thinking about the impact of technology and the ethics of that impact https://www.scu.edu/ethics-in-technology-practice/

- Microsoft's "Harms Modeling" approach provides a guide to mapping stakeholders and anticipating harms[70] https://docs.microsoft.com/en-us/azure/architecture/guide/responsible-innovation/harms-modeling/

- "Heuristic Preventative Design" as described in section 3.2.3 of Li et al., 2018[71] describes a formalized process for anticipating and mitigating predictable negative impacts, and a case study demonstrating how they can then be mitigated.

69 This list should not be considered exhaustive: it is merely a starting point, and we welcome suggestions for additional resources. Inclusion in the list should not necessarily be considered endorsement; these are resources that members of our Partner community have found helpful, but we encourage readers to use their judgement.

70 In particular, see the taxonomy of types of harms.

71 Li, H., Alarcon, B., Espinosa, S. M., B, & Hecht. (2018). Out of site: Empowering a new approach to online boycotts. *Proceedings of the ACM on Human-Computer Interaction, 2*(CSCW), 1–28. https://doi.org/10.1145/3274375

- Techniques such as design fiction, critical design, and speculative design may help researchers think more creatively about the potential impacts of their work. Resources include:
    - The Little Book of Design Fiction covers how to do design fiction in 3 ways: addressing problems (e.g. bias), products (e.g. AI assistants), and technologies (e.g. NLP) https://www.research.lancs.ac.uk/portal/files/259562831/Little_Book_of_Design_Fiction.pdf
    - Lachlan Urquart's legal IT cards provide prompts for thinking about law, security, ethics, and privacy https://lachlansresearch.com/the-moral-it-legal-it-decks/
    - There are also various agencies and academic labs that specialize in this kind of work[72]

- Stakeholder mapping exercises may help reduce the risk of overlooking groups that might be affected by the impacts of the research

72 Agencies include: http://nearfuturelaboratory.com/, https://superflux.in/, https://changeist.com/, Academic labs include: https://designresearch.works/, http://imaginari.es/, http://eds.siat.sfu.ca/, https://www.designinformatics.org/

Appendix IV: AI research and Institutional Review Boards

"Institutional Review Boards" (IRBs) are a regulatory mechanism in the US to ensure research funded by the federal government is conducted ethically and the welfare of participants is protected. To address growing concerns about the ethics and impacts of AI research, there have been calls for IRBs for AI research.[73] There are, however, a number of challenges with this for the purpose of mitigating downstream consequences:

- A lot of AI research is privately funded, and so (at least in the US), would not be subject to IRB approval. In addition, there is uncertainty around the extent to which AI research would be covered under the current definitions, since there are exemptions for the use of second-hand data (such as that used to train ML models) and research that is for quality assurance purposes (which could apply to a lot of corporate AI research).

- Generally, IRBs focus only on risks to the participants of the research, rather than the potential downstream consequences for wider society.

- AI research is a global endeavor, and its impacts can reach far beyond the country in which it was conducted. In addition to the challenge of different countries having different conceptions of ethics, they also currently approach and implement IRBs (or their equivalent) very differently; there is no single global standard.

- Finally, IRBs are complex and resource-intensive, and it would not be trivial to put in place the regulation and resources needed to apply it to AI research.

  There are some efforts underway[74] to explore how independent processes and institutions might fulfill an IRB-like function for AI, though it is likely to be a while before these alternative processes are standardized and implemented.

While IRBs or a similar mechanism might in future might be more feasible, these limitations are why we have primarily focused on other types of interventions in this report.

73 Blackman, R. (2021, April 1). If your company uses AI, it needs an institutional review board. *Harvard Business Review*. https://hbr.org/2021/04/if-your-company-uses-ai-it-needs-an-institutional-review-board

74 FPF receives grant to design ethical review process for research access to corporate data. (2019, October 15). *Future of Privacy Forum*. https://fpf.org/blog/fpf-receives-grant-to-design-ethical-review-process-for-research-access-to-corporatedata/