Exploring the impact of broader impact requirements for AI governance

March 29 2021, by Ingrid Fadelli

Transparency	▪ Reasons and expectations ▪ Procedural transparency ▪ Accountability mechanisms
Guidance	▪ Guidance for researchers ▪ Guidance for reviewers ▪ Communication channels
Incentives	▪ Peer-review ▪ Outside expert involvement ▪ Encouragement to cite ▪ Prizes
Deliberation	▪ Forums for deliberation ▪ Data-driven deliberation ▪ Minimisation of reputational or legal costs

Credit: Prunkl et al.

As machine learning algorithms and other artificial intelligence (AI) tools become increasingly widespread, some governments and institutions have started introducing regulations aimed at ensuring that they are ethically designed and implemented. Last year, for instance, the Neural Information Processing Systems (NeurIPS) conference introduced a new ethics-related requirement for all authors submitting AI-related research.

Researchers at University of Oxford's Institute for Ethics in AI, the department of Computer Science and the Future of Humanity Institute have recently published a perspective paper that discusses the possible impact and implications of requirements such as the one introduced by the NeurIPS conference. This paper, published in *Nature Machine Intelligence*, also recommends a series of measures that may maximize these requirements' chance of success.

"Last year, NeurIPS introduced a requirement that submitting authors include a broader impact statement in their papers," Carina E. Prunkl, one of the researchers who carried out the study, told TechXplore. "A lot of people—including us—were taken by surprise. In response, we decided to write two pieces on the topic: a guide for researchers on how to start thinking about the broader impacts of their research and write a broader impact statement, as well as this perspective article, which really is about drawing out some of the potential impacts of such broader impact requirements."

Predicting and summarizing the possible impacts of a given research study is a highly complex and challenging task. It can be even more challenging in cases where a given technological tool or technique could have a variety of applications across a wide range of settings.

In their paper, Prunkl and her colleagues build on findings of studies that examined different governance mechanisms to delineate the possible

benefits, risks and challenges of the requirement introduced by NeurIPS. In addition, they propose a series of strategies that could mitigate potential challenges, dividing them into four key categories: transparency, guidance, incentives and deliberation.

"Our overall objective was to contribute to the ongoing discussion on community-led governance mechanisms by raising awareness of some of the potential pitfalls, and to provide constructive suggestions to improve the process," Prunkl said. "We begin the discussion by looking at the effects of other governance initiatives, such as institutional review boards, that are similar in nature and also involve researchers writing statements on the impacts of their research."

Prunkl and her colleagues considered previous AI governance initiatives that asked researchers to prepare statements about the impact of their work and highlighted some of the lessons learnt about such statements. They then discussed the potential benefits and risks of NeurIPS' broader impact statement requirement. Finally, they prepared a list of suggestions for conference organizers and the ML community at large, which could help them to improve the likelihood that such statements will have positive effects on the development of AI.

"Some of the benefits we list are improved anticipation and mitigation of potential harmful impacts from AI, as well as improved communication between research communities and policy makers," Prunkl said. "If not implemented carefully, there is a risk that statements will be of low quality, that ethics is regarded as a box-ticking-exercise or even that ethics is being trivialized, by suggesting that it is in fact possible to fully anticipate impacts in this way."

To assess and predict the broader impact of a given technology, researchers should ideally have a background in disciplines such as ethics or sociology and a robust knowledge of theoretical frameworks

and previous empirical results. In their paper, Prunkl and her colleagues outline a series of possible root causes for the failure or negative effects of past governance initiatives. These causes include the inherent difficulties encountered when trying to identify the broader impacts of a given study or technological tool, as well as institutional or social pressures and a lack of general guidelines to assist researchers in writing their statements.

"Our main suggestions focus on four key themes: first, improving transparency and setting expectations, which includes communication of the purpose, motivation, and expectation as well as procedural transparency in how these statements are being evaluated," Prunkl said. "Second, providing guidance, which includes both guidance on how to write these statements, as well as guidance for referees on how to evaluate them."

In their paper, Prunkl and her colleagues also highlight the importance of setting incentives. Preparing high-quality statements can be expensive and time-consuming, thus they feel that institutions should introduce incentives that encourage more researchers to invest significant time and effort on reflecting about the impact of their work.

"One solution would be to integrate the evaluation of statements into the peer-review process," Prunkl explained. "Other options include creating designated prizes and to encourage authors to cite other impact statements."

The fourth theme emphasized by Prunkl and her colleagues relates to public and community deliberation. This final point reaches beyond the context of broader impact statements and the researchers feel that it should be at the basis of any intervention aimed at governing AI. They specifically highlight the need for more forums that allow the ML community to deliberate on potential measures aimed at addressing the

risks of AI.

"Finding governance solutions that effectively ensure the safe and responsible development of AI is one of the most pressing challenges these days," Prunkl said. "Our article highlights the need to think critically about such governance mechanisms and reflect carefully on the risks and challenges that might arise and that could undermine the anticipated benefits. Finally, our article emphasizes the need for community deliberation on such governance mechanisms."

Prunkl and her colleagues hope that the list of suggestions they prepared will help conference organizers who are planning to introduce broader impact requirements to navigate possible challenges associated with AI development. The researchers are currently planning to intensify their work with ML researchers, in order to further assist them with preparing research impact statements. For instance, they plan to co-design sessions with researchers where they will collaboratively create resources that could help teams to prepare these statements and identify the broader impacts of their work.

"The debate around impact statements has really highlighted the lack of consensus about which governance mechanisms should be adopted, and how they should be implemented," Prunkl said. "In our paper, we highlight the need for continued, constructive deliberation around such mechanisms. In response to this need, one of the authors, Carolyn Ashurst, (along with Solon Barocas, Rosie Campbell, Deborah Raji and Stuart Russell) organized a NeurIPS workshop on the topic of 'Navigating the Broader Impacts of AI Research.'"

During the workshop organized by Ashurst and her colleagues, participants discussed NeurIPS impact statements and ethical reviews, as well as broader questions around the idea of responsible research and development. Moreover, the organizers explored the roles that different

parties within the ML research ecosystem can play in navigating the preparation of broader impact statements.

In the future, Prunkl and her colleagues plan to create more opportunities for constructive deliberation and discussion related to AI governance. Their hope is that the ML community and other parties involved in AI use will continue working together to establish norms and mechanisms aimed at effectively addressing issues that can arise from ML research. In addition, the researchers will conduct further studies aimed at analyzing impact statements and general attitudes towards these statements.

"Work to analyze the impact statements from conference preprints has already surfaced both encouraging and concerning trends," Prunkl said. "Now that the final versions of conference papers are publicly available, we/GovAI/our research group have started to analyze these statements, to understand how researchers responded to the requirement in practice. Alongside this, more work is needed to understand the current attitudes of ML researchers towards this requirement. Work by researchers at ElementAI found a mixed response from NeurIPS authors; while some found the process valuable, others alluded to many of the challenges highlighted in our paper, for example describing the requirement as 'one more burden that falls on the shoulders of already overworked researchers.'"

More information: Institutionalizing ethics in AI through broader impact requirements. *Nature Machine Intelligence*(2021). DOI: 10.1038/s42256-021-00298-y.

Like a researcher stating broader impact for the very first time. arXiv:2011.13032 [cs.CY]. arxiv.org/abs/2011.13032