Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-22)
Journal Track

Ethics and Governance of Artificial Intelligence: A Survey of Machine Learning Researchers (Extended Abstract) *

Baobao Zhang[1,2] , **Markus Anderljung**[2] , **Lauren Kahn**[3] , **Noemi Dreksler**[2] , **Michael C. Horowitz**[4] , **Allan Dafoe**[2,5]
[1]Syracuse University
[2]Centre for the Governance of AI
[3]Council on Foreign Relations
[4]University of Pennsylvania
[5]DeepMind
*baobaozhangresearch@gmail.com, markus.anderljung@governance.ai, lkahn@cfr.org, noemi.dreksler@gmail.com, horom@sas.upenn.edu, allandafoe@deepmind.com

Abstract

Machine learning (ML) and artificial intelligence (AI) researchers play an important role in the ethics and governance of AI, including through their work, advocacy, and choice of employment. Nevertheless, this influential group's attitudes are not well understood, undermining our ability to discern consensuses or disagreements between AI/ML researchers. To examine these researchers' views, we conducted a survey of those who published in two top AI/ML conferences ($N = 524$). We compare these results with those from a 2016 survey of AI/ML researchers and a 2018 survey of the US public. We find that AI/ML researchers place high levels of trust in international organizations and scientific organizations to shape the development and use of AI in the public interest; moderate trust in most Western tech companies; and low trust in national militaries, Chinese tech companies, and Facebook. While the respondents were overwhelmingly opposed to AI/ML researchers working on lethal autonomous weapons, they are less opposed to researchers working on other military applications of AI, particularly logistics algorithms. A strong majority of respondents think that AI safety research should be prioritized more and a majority that ML institutions should conduct pre-publication review to assess potential harms. Being closer to the technology itself, AI/ML researchers are well placed to highlight new risks and develop technical solutions, so this novel data has broad relevance. The findings should help to improve how researchers, private sector executives, and policymakers think about regulations, governance frameworks, guiding principles, and national and international governance strategies for AI.

*This paper is an extended abstract of an article published in the *Journal of Artificial Intelligence Research*. This work has not been presented at a major AI conference with archival proceedings before.

1 Introduction

Tech companies and governments alike see the potential for artificial intelligence (AI) and have moved to develop machine learning (ML), particularly deep learning, applications across a variety of sectors — from healthcare to national security [Kanaan, 2020; Horowitz, 2018]. Civil society groups, governments, and academic researchers have expressed concerns about AI related to safety [Amodei *et al.*, 2016; Russell, 2019], discrimination and racial bias [Noble, 2018; Barocas *et al.*, 2019], and risks associated with uses of AI in a military and government context [Brundage *et al.*, 2018; Horowitz, 2019; Zwetsloot and Dafoe, 2019].

There is a small but growing literature that surveys the AI/ML community. Most existing surveys focus on eliciting researcher forecasts on AI progress, such as when specific milestones will be reached or when AI will surpass human performance at nearly all tasks [Sandberg and Bostrom, 2011; Grace *et al.*, 2018; Gruetzemacher *et al.*, 2020]. Others have focused on how computer scientists define AI [Krafft *et al.*, 2020] or the impact of AI on society [Anderson *et al.*, 2018]. AI/ML professionals have also been surveyed in regard to their views on working on military-related projects [Aiken *et al.*, 2020b] and their immigration pathways and intentions [Aiken *et al.*, 2020a; Zwetsloot *et al.*, 2021].

To better understand the attitudes of this critical community, which will impact future AI governance, we surveyed 524 technical experts' attitudes about the governance of AI. Key results from our survey include:

- Relative to the American public, AI/ML researchers place high levels of trust in international organizations (e.g., the UN, EU, etc.) to shape the development and use of AI in the public interest. While the American public rated the US military as one of the most trustworthy actors, AI/ML researchers place relatively low levels of trust in the militaries of countries where they do research.
- The majority of AI/ML researchers (68%) indicate that AI safety, broadly defined, should be prioritized more than it is presently.

- Researchers reveal nuanced views about the appropriate sharing of research. While most researchers believe that "researchers should be encouraged to share" all aspects of research, there is considerable variation among the aspects of research that they feel "must be shared every time": 84% think that high-level description of the methods must be shared every time while only 22% think that of the trained model. Further, a majority of AI/ML researchers (59%) support "pre-publication review" for "work that has some chance of adverse impact."
- The respondents are wary of AI/ML researchers working on certain military applications of AI. Respondents are the most opposed to other researchers working on lethal autonomous weapons (58% strongly oppose) but far fewer are opposed to others working on logistics algorithms (6% strongly oppose) for the military. 31% of researchers indicate that they would resign or threaten to resign from their jobs, and 25% indicate that they would speak out publicly to the media or online, if their organization decided to work on lethal autonomous weapons.

2 Method

To study attitudes about trust and governance in AI, we contacted 3,030 AI/ML researchers between September 16 and October 13, 2019, of whom 524 researchers responded (17%). The researchers were selected based on having papers accepted at two top AI research conferences, following the sampling frame of Grace *et al.* 2018. One group of respondents had papers accepted to the 2018 NeurIPS conference and the other to the 2018 ICML conference. Another group had papers accepted at NeurIPS and ICML in 2015 and participated in a 2016 researcher survey on AI [Grace *et al.*, 2018]. We chose the sample to match that of Grace *et al.* (2018), which chose ICML and NeurIPS as they were the two largest, widely cited, and general conferences [Zhang *et al.*, 2021b]. The complete methodology, results, figures, and tables of the survey can be found in Zhang *et al.* (2021a).

3 Results

Figure 1 shows the mean importance of AI governance challenges, along with the corresponding 95% confidence interval for both AI/ML researchers and the general public [Zhang and Dafoe, 2020]. For the AI/ML researcher group, almost all issues were rated as having a mean importance of around 2.5, between "somewhat important" and "very important," with the top five issues including preventing criminal justice bias, ensuring autonomous vehicles are safe, preventing critical AI system failure, protecting data privacy, and preventing mass surveillance. Hiring bias and technological unemployment are rated slightly (about 0.3 points) lower than other issues. The one outlier is "Reducing risks from US-China competition over AI," rated significantly below the other challenges at 1.8 (just below "somewhat important"); this result may be an artifact of our question phrasing, in that AI/ML researchers may believe that risks from US-China competition are real, but not one that is helped by "tech companies and governments" trying to "carefully manage" them.

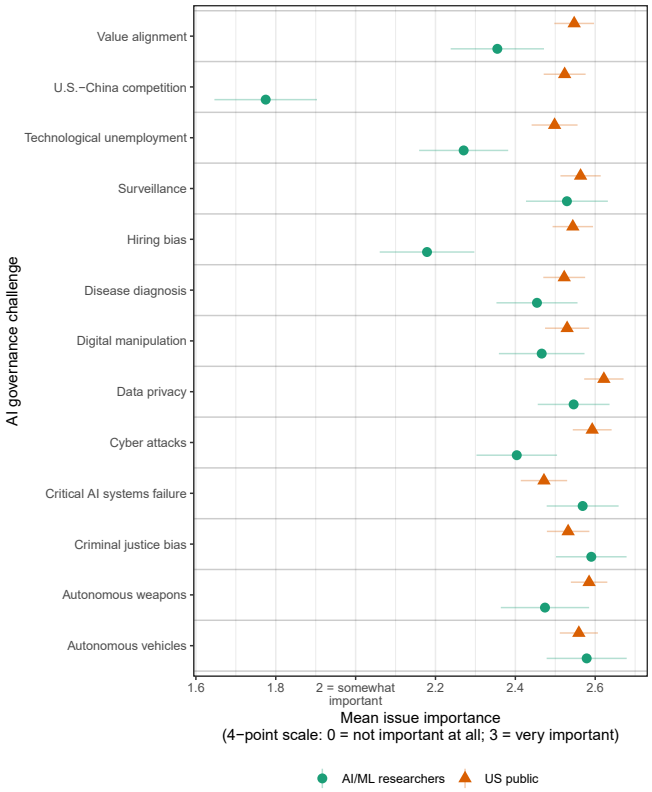

Figure 1: Perception of "how important it is for tech companies and governments to carefully manage" AI governance challenges. We compare AI/ML researchers' and the US public's responses. Each respondent was presented with five AI governance challenges randomly selected from a list of 13. Respondents were asked to evaluate the importance of each governance challenge using a four-point scale (the slider scale allows respondents to input values to the tenth decimal point): 0 = not important, 1 = not too important, 2 = somewhat important, 3 = very important. We present the mean responses for each governance challenge (by respondent type) along with the corresponding 95% confidence intervals.

There is considerable overlap between the assessment of AI governance challenges by AI/ML researchers and the US public. Both groups rate protecting data privacy, preventing mass surveillance, and ensuring that autonomous vehicles are safe among the five most important governance challenges. AI/ML researchers placed significantly less importance on value alignment[1], technological unemployment, and hiring bias, and slightly more importance on critical AI systems failure, criminal justice bias, and autonomous vehicles, than the public. The gap between AI/ML researchers and the US public is particularly large when it comes to preventing the risks from US-China competition in AI. In contrast to AI/ML researchers' relatively low mean rating of 1.77 out of 3, the US public gave US-China competition a mean rating of 2.52 out of 3.

Good governance benefits from understanding what institutions and organizations AI/ML researchers (and other

[1]Defined as "AI systems are safe, trustworthy, and aligned with human values."

stakeholders) trust. To test AI/ML researcher trust in different organizations and institutions, we ask: "Suppose the following organizations were in a position to strongly shape the development and use of advanced AI. How much trust do you have in each of these organizations to do so in the best interests of the public?" Respondents were shown five randomly selected actors. For each actor, they then assigned a number value on a four-point scale ranging from 0 "no trust at all" to 3 "a great deal of trust." For AI/ML researchers, the most trusted actors, with a mean score above 2, were non-governmental scientific associations and intergovernmental research organizations. The Partnership on AI, a consortium of tech companies, academics, and civil society groups, is also rated relatively highly (mean score of 1.89). It is noteworthy that these more neutral, scientific organizations received the highest trust ratings but currently play a relatively small role in AI development and management.

Relative to the American public, AI/ML researchers place high levels of trust in international organizations (e.g., the UN, EU, etc.) to shape the development and use of AI in the public interest. While the American public rated the US military as one of the most trustworthy actors, AI/ML researchers place relatively low levels of trust in the militaries of countries where they do research.

The safety of AI systems may be a critical factor in their development and adoption. We asked respondents about their familiarity with and prioritization of AI safety. We described AI safety in a broad way, as focused on "making AI systems more robust, more trustworthy, and better at behaving in accordance with the operator's intentions," and also provided examples. We first sought to understand how familiar researchers were with AI safety research. We asked them to make a self-assessment using a five-point scale, ranging from 0 "not familiar at all (first time hearing of the concept)" to 4 "very familiar (worked on the topic)." To evaluate views about the value of AI safety research, we asked respondents, "How much should AI safety be prioritized relative to today?" Respondents selected answers on a five-point Likert scale, ranging from -2 "much less" to 2 "much more" with 0 meaning "about the same."

The AI/ML researchers we surveyed report, on average, moderate familiarity with AI safety as a concept. The distribution follows an approximately normal distribution, although it is right-skewed. 3% of respondents say that they are "not familiar at all" with AI safety while 15% say they are "very familiar."

When asked about prioritizing AI safety an overwhelming majority of our respondents (68%) say that the field should be prioritized more than at present. These results demonstrate significant growth in the reported prioritization of AI safety in the research community, though different definitions may have caused these differences. In a similar survey of AI/ML researchers conducted in 2016, 49% of respondents believed that AI safety should be prioritized more than it was at the time [Grace *et al.*, 2018].[2]

[2]We updated the definition of AI safety research from Grace *et al.* 2018 after consultation with AI/ML researchers working in AI safety research.

The AI/ML research community has recently seen innovation and subsequent controversy regarding publication norms, which also relate to questions of trust. Such norms concern when, how, and where research is published. OpenAI's release strategy for GPT-2, a large language model, is a prime example. Citing concerns the system could be used for "malicious purposes", they employed a staged release strategy; the initial paper was accompanied by a smaller version of GPT-2, the full model only being released eight months later [Solaiman *et al.*, 2019].

We asked questions to generate insights into AI/ML researchers' views on publication norms. First, we assessed how much they agree or disagree that "machine learning research institutions (including firms, governments, and universities) should practice pre-publication review," which involves "a strong norm or policy" to have discussions that are "informed, substantive, and serious" about "the ethical implications of publication". A majority of respondents agree (20% strongly agree; 39% somewhat agree) with the statement. Additionally, both familiarity with AI safety and prioritization of AI safety significantly predict support for pre-publication review. These results speak to interest amongst AI/ML researchers to address the risks of misuse of their work.

Next, we asked respondents about the importance of sharing various aspects of AI/ML research. Respondents were shown three aspects of research, randomly selected from a list of seven (e.g., high-level description of methods, code, and training data). For each aspect of research, respondents could select from six levels of sharing, ranging from "it doesn't matter" to "it must be shared every time."

The respondents think that high-level descriptions of the methods, the results, and a detailed description of the methods should almost always be openly shared. However, support declines for requiring the sharing of other information that would be essential for replication, such as the code, the training data, or the trained model. Researchers felt that sharing these aspects of research should be encouraged but not required. On the high end, 84% indicated that high-level description of methods must be shared every time; on the low end, only 22% indicated that the trained model must be shared every time. We do not find significant differences in responses between researchers who work in academia versus in industry.

We also investigated researchers' views toward military applications of AI. Working on military uses of AI requires a great deal of trust in how they might be used, given the central role that some think AI could play in the future of military power [Scharre, 2018]. We asked about three areas of military applications of AI that have received public scrutiny: lethal autonomous weapon systems, surveillance technologies for intelligence agencies, and military logistics. Respondents were asked to evaluate two randomly selected military applications out of the three. They were asked whether they would support or oppose researchers working on the application in the country where the respondent currently works or studies. Respondents selected answers on a Likert scale, ranging from -2 "strongly oppose" to 2 "strongly support." Those who answered that they "strongly oppose" or "some-

what oppose" researchers working on the applications were asked what types of collective actions (e.g., signing a petition or protesting) they would take if their organization decided to conduct such research.

Our results show researchers have substantial concerns regarding working on some military applications of AI. Nevertheless, there are nuances to their views. Researchers, on average, more than somewhat oppose work on lethal autonomous weapon systems (-1.3), very weakly oppose work on surveillance applications (-0.3), and very weakly support work on logistics applications (0.5). 58% strongly oppose other researchers working on lethal autonomous weapons, 20% strongly oppose others working on surveillance tools, but only 6% strongly oppose others working on military logistics. This is consistent with work by Aiken, Kagan, and Page (2020b), which focuses just on US-based AI professionals and finds that US-based AI professionals are more opposed to working on battlefield applications of AI than other applications.

A majority of researchers who said they opposed others working on each application said they would actively avoid working on the project, express their concern to a superior in their organization involved in the decision, or sign a petition against the decision. 75% of researchers who said they opposed others working on lethal autonomous weapons said they would avoid working on lethal autonomous weapons themselves, and 42% of those respondents said they would resign or threaten to resign from their jobs. In absolute terms, 31% of researchers indicated that they would resign or threaten to resign from their jobs, and 25% indicated that they would speak out publicly to the media or online if their organization decided to work on lethal autonomous weapons. Of those who say they oppose other researchers working on lethal autonomous weapons, less than 1% said they would do nothing. The percentages for surveillance and logistical software are 3.5% and 7.5%, respectively.

4 Limitations

It is important to recognize some of the limits to our findings. While our sample is more extensive than previous research, it has clear limits. Our sample strategy focused on those who publish in the top two AI/ML conferences; it thus may underweight the perspective of those subgroups of the AI/ML community who are less likely to publish there, such as product-focused industry researchers. Second, this survey captures the views of the researchers at a particular point in time, while the norms around AI research and publishing continue to evolve, and significant shifts in the psychological, political, and socioeconomic landscape continue to occur.

Another limitation might include demographic biases or response bias. The gender bias in our sample (91% of our respondents were men) needs to be noted, which mirrors the low gender diversity in the field itself. Further, a multiple regression that examined the association between demographic characteristics and response found that respondents have lower h-indexes (a measure of productivity and citation impact of researchers) and are more likely to work in academia compared with non-respondents. Future surveys could address these issues and also expand the sample frame to include related researchers, such as AI ethics experts and social scientists who study the societal impact of AI. This might also help address the gender bias in our sample. Nevertheless, the unique scope of the sample gives us the ability to speak to AI/ML research attitudes about AI governance in unique ways compared to previous literature.

5 Conclusion

As institutions, regulations, and norms of AI governance are forming, this survey of AI/ML researchers provides insight into how this emerging epistemic community views the ethical and governance issues related to the technology. The respondents place relatively high levels of trust in international organizations to manage the development and use of AI in the public interest. Researchers' trust some tech companies substantially more than others to develop and use AI in the public interest, a fact of potentially great relevance given the epistemic authority of AI researchers and the competition for AI talent. Compared with the US public who place high levels of trust in the US military, AI/ML researchers are relatively distrustful of the military. Furthermore, the AI/ML researchers we surveyed are opposed to working on lethal autonomous weapon systems in particular. Given their responses about publication norms, the respondents are also aware of the potential adverse impacts of their research. Finally, a majority of respondents think that AI safety research should be prioritized more and researchers should conduct pre-publication reviews to assess the potential harms their research could cause. This line of research could help guide policymakers, tech companies, civil society, and the AI/ML community in building and deploying safe and ethical AI systems.

Acknowledgements

We want to thank Charlie Giattino, Emmie Hine, Tegan McCaslin, Kwan Yee Ng, and Catherine Peng for their research assistance. For helpful feedback and input, we want to thank: Catherine Aiken, Carolyn Ashurst, Miles Brundage, Rosie Campbell, Alexis Carlier, Jeff Ding, Owain Evans, Ben Garfinkel, Katja Grace, Ross Gruetzemacher, Jade Leung, Alex Lintz, Max Negele, Toby Shevlane, Brian Tse, Eva Vivalt, Waqar Zaidi, Remco Zwetsloot, our colleagues at our respective institutions, and our anonymous reviewers. We are also grateful for research support from the Center for Security and Emerging Technology at Georgetown University and the Berkeley Existential Risk Initiative. This research was supported by: the Ethics and Governance of AI Fund, the Open Philanthropy Project grant for "Oxford University – Research on the Global Politics of AI," the Minerva Research Initiative under Grant #FA9550-18-1-0194, and the CIFAR Azrieli Global Scholars Program. The research reported here should solely be attributed to the authors; all errors are the responsibilities of the authors.

References

[Aiken *et al.*, 2020a] Catherina Aiken, James Dunham, and Remco Zwetsloot. Immigration pathways and plans of AI

talent. Technical report, Center for Security and Emerging Technology, 2020.

[Aiken *et al.*, 2020b] Catherina Aiken, Rebecca Kagan, and Michael Page. "Cool projects" or "expanding the efficiency of the murderous American war machine?": AI professionals' views on working with the Department of Defense. Technical report, Center for Security and Emerging Technology, 2020.

[Amodei *et al.*, 2016] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. *CoRR*, abs/1606.06565, 2016.

[Anderson *et al.*, 2018] Janna Anderson, Lee Rainie, and Alex Luchsinger. Artificial intelligence and the future of humans. Technical report, Pew Research Center, 12 2018.

[Barocas *et al.*, 2019] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019. http://www.fairmlbook.org.

[Brundage *et al.*, 2018] Miles Brundage, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, Paul Scharre, Thomas Zeitzoff, Jacob Steinhardt, Carrick Flynn, Seán Ó hÉigeartaigh, Simon Beard, Haydn Belfield, Sebastian Farquhar, Clare Lyle, Rebecca Crootof, Owain Evans, Michael Page, Joanna Bryson, Roman Yampolskiy, and Dario Amodei. The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. Technical report, Future of Humanity Institute and University of Oxford and Centre for the Study of Existential Risk and University of Cambridge and Center for a New American Security and Electronic Frontier Foundation and OpenAI, 2018.

[Grace *et al.*, 2018] Katja Grace, John Salvatier, Allan Dafoe, Baobao Zhang, and Owain Evans. When will AI exceed human performance? Evidence from AI experts. *Journal of Artificial Intelligence Research*, 62:729–754, 2018.

[Gruetzemacher *et al.*, 2020] Ross Gruetzemacher, David Paradice, and Kang Bok Lee. Forecasting extreme labor displacement: A survey of AI practitioners. *Technological Forecasting and Social Change*, 161, 2020. 120323.

[Horowitz, 2018] Michael C Horowitz. Artificial intelligence, international competition, and the balance of power. *Texas National Security Review*, 1(3), 2018.

[Horowitz, 2019] Michael C. Horowitz. When speed kills: Lethal autonomous weapon systems, deterrence and stability. *Journal of Strategic Studies*, 42(6):764–788, aug 2019.

[Kanaan, 2020] Michael Kanaan. *T-Minus AI: Humanity's Countdown to Artificial Intelligence and the New Pursuit of Global Power*. BenBella Books, Dallas, TX, 2020.

[Krafft *et al.*, 2020] P. M. Krafft, Meg Young, Michael Katell, Karen Huang, and Ghislain Bugingo. Defining AI in policy versus practice. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, page 72–78, New York, NY, USA, 2020. Association for Computing Machinery.

[Noble, 2018] Safiya Umoja Noble. *Algorithms of Oppression*. NYU Press, New York, 2018.

[Russell, 2019] Stuart Russell. *Human Compatible: Artificial Intelligence and the Problem of Control*. Penguin, New York, 2019.

[Sandberg and Bostrom, 2011] Anders Sandberg and Nick Bostrom. Machine intelligence survey. Technical report, Future of Humanity Institute, Oxford University, 12 2011. Technical Report #2011-1.

[Scharre, 2018] Paul Scharre. *Army of None: Autonomous Weapons and the Future of War*. WW Norton & Company, New York, 2018.

[Solaiman *et al.*, 2019] Irene Solaiman, Jack Clark, and Miles Brundage. GPT-2: 1.5B release. OpenAI Blog. URL: https://perma.cc/PFA8-KTBP, 2019. Accessed: 23 Jul. 2020.

[Zhang and Dafoe, 2020] Baobao Zhang and Allan Dafoe. US public opinion on the governance of artificial intelligence. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, page 187–193, New York, NY, USA, 2020. Association for Computing Machinery.

[Zhang *et al.*, 2021a] Baobao Zhang, Markus Anderljung, Lauren Kahn, Noemi Dreksler, Michael C. Horowitz, and Allan Dafoe. Ethics and Governance of Artificial Intelligence: Evidence from a Survey of Machine Learning Researchers. *Journal of Artificial Intelligence Research*, 71, August 2021.

[Zhang *et al.*, 2021b] Daniel Zhang, Saurabh Mishra, Erik Brynjolfsson, John Etchemendy, Deep Ganguli, Barbara Grosz, Terah Lyons, James Manyika, Juan Carlos Niebles, Michael Sellitto, Shoham Yoav, Jack Clark, and Raymond Perrault. Artificial Intelligence Index Report 2021. Technical report, AI Index Steering Committee, Human-Centered AI Institute, Stanford University, Stanford, 2021.

[Zwetsloot and Dafoe, 2019] Remco Zwetsloot and Allan Dafoe. Thinking about risks from AI: Accidents, misuse and structure. Lawfare. URL: https://perma.cc/4J2N-2KYV, 2019. Accessed: 23 Sep. 2020.

[Zwetsloot *et al.*, 2021] Remco Zwetsloot, Baobao Zhang, Noemi Dreksler, Lauren Kahn, Markus Anderljung, Allan Dafoe, and Michael C. Horowitz. Skilled and mobile: Survey evidence of ai researchers' immigration preferences. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21), May 19–21, 2021, Virtual Event, USA*, 2021.