Global Catastrophic Risk INSTITUTE

March 4, 2022

RFI Response: National Artificial Intelligence Research and Development Strategic Plan— White House Office of Science and Technology Policy
87 FR 5876; Document Number 2022-02161

Dr. Alondra Nelson, Deputy Director of Science and Society of the Office of Science and Technology Policy (OSTP) and Performing the Duties of OSTP Director, the National Science and Technology Council's (NSTC) Select Committee on Artificial Intelligence (Select Committee), the NSTC Machine Learning and AI Subcommittee (MLAI-SC), the National AI Initiative Office (NAIIO), and the Networking and Information Technology Research and Development (NITRD) National Coordination Office (NCO):

Thank you for the invitation to submit comments in response to the Request For Information (RFI) to the National Artificial Intelligence Research and Development Strategic Plan. We, the Global Catastrophic Risk Institute (GCRI), are researchers with expertise on AI ethics and AI governance. We offer the following submission for your consideration.

We support the eight strategic aims described in the 2019 Update. As detailed below, we encourage specific changes to seven of the aims that support a more robust and inclusive approach to AI ethics and governance. These changes seek to ensure that the National AI R&D Strategic Plan supports broad and sustainable success and meets high social and ethical standards.

Our recommendations are as follows:

Strategy 1: Make long-term investments in AI research.

We recommend an emphasis on interdisciplinary research in which technological progress is oriented according to social and ethical values and in which technology governance is informed by a sound understanding of the nature of AI technology.

Investment in basic research is essential for improving the capability and reliability of AI systems. However, there has been a tendency for some research on AI systems to be focused on improving capabilities without substantial consideration of social and ethical

dimensions.[1] This risks the development of AI technology that can have inappropriate impacts on society. For example, machine learning research has often pursued larger neural network models to improve model accuracy without regard for the adverse energy resource and climate change consequences of larger models.[2] To align research on AI systems with social and ethical values, it is essential to have these values built into the core of the research, including the selection of which research directions to pursue.[3]

The ability of society to successfully govern AI technology additionally depends on basic research. AI technology has only recently risen to prominence as a societal issue, and the study of AI governance is likewise at an early stage. Research on AI governance has often focused on general concepts such as ethical principles, with less regard for how to operationalize them.[4] Given rapid ongoing changes in AI technology, an important challenge is to ensure that governance concepts are informed by a state-of-the-art understanding of the technology.[5] To formulate practical and technologically sound AI governance concepts, it is essential to invest in AI governance research in which the AI technology is not treated as a black box but instead is considered in detail.

Strategy 2: Develop effective methods for human-AI collaboration.

We recommend an emphasis on methods to ensure that human-AI collaboration is in the common public good and not just in the interests of the select few with control of advanced AI tools.

AI technology has enabled major advances in workplace productivity, bringing major economic benefits. However, this often comes at the expense of disadvantaged populations. For example, low-wage workers are forced to work unpredictable schedules that are optimized according to AI processing of last-minute data[6] and social

[1] Seth D. Baum, "On the promotion of safe and socially beneficial artificial intelligence", *AI & Society*, vol. 32, no. 4 (2017), pp. 543-551, https://doi.org/10.1007/s00146-016-0677-0.
[2] Lynn Kaack, Priya Donti, Emma Strubell, George Kamiya, Felix Creutzig, and David Rolnick, "Aligning artificial intelligence with climate change mitigation", 2021, https://hal.archives-ouvertes.fr/hal-03368037.
[3] Steven Umbrello and Ibo Van de Poel, "Mapping value sensitive design onto AI for social good principles", *AI and Ethics*, vol. 1, no. 3 (2021), pp. 283-296, https://doi.org/10.1007/s43681-021-00038-3.
[4] Jessica Morley, Libby Kinsey, Anat Elhalal, Francesca Garcia, Marta Ziosi, and Luciano Floridi, "Operationalising AI ethics: barriers, enablers and next steps", *AI & Society* (2021), https://doi.org/10.1007/s00146-021-01308-8.
[5] Wendell Wallach and Gary Marchant, "Toward the Agile and Comprehensive International Governance of AI and Robotics", *Proceedings of the IEEE*, vol. 107, no. 3 (2019), pp. 505-508, https://doi.org/10.1109/JPROC.2019.2899422.
[6] Cathy O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (Crown, 2016).

media websites use AI recommender algorithms that increase user engagement by promoting politically extremist content.[7]

The 2016 National AI R&D Strategic Plan recognizes that “the challenge of understanding and designing human-AI ethics and value alignment into systems remains an open research area.” We agree that this is an important open research area. However, it is important to address: to whom specifically are AI systems aligned?[8] AI systems that are designed to support a single individual require fundamentally different designs than those that are designed to support society as a whole.[9] Only by including the common good of society in AI system design can human-AI collaboration advance the interests of all members of society instead of the select few.

Strategy 3: Understand and address the ethical, legal, and societal implications of AI.

We recommend a holistic and pluralistic approach to the ethical, legal, and societal (ELS) implications of AI, in particular, to address the full range of implications and to evaluate them in terms of a diversity of social and ethical perspectives.

It is vital that the ELS implications of AI are central to all work on AI and not tacked on as an afterthought. There are ELS implications in, among other things, (1) the selection of algorithms, determining which ethical concepts can be implemented in an AI system[10]; (2) the selection of the scale at which to implement the algorithm, determining the energy consumption of the AI system[11]; (3) the selection of training data, determining which problems and issues an AI system can be applied to[12] and the

[7] Joe Whittaker, Seán Looney, Alastair Reed, and Fabio Votta, "Recommender systems and the amplification of extremist content", *Internet Policy Review*, vol. 10, no. 2 (2021), https://doi.org/10.14763/2021.2.1565.

[8] Seth D. Baum, "Social choice ethics in artificial intelligence", *AI & Society*, vol. 35, no. 1 (2020), pp. 165-176, https://doi.org/10.1007/s00146-017-0760-1.

[9] Roger Lera-Leri, Filippo Bistaffa, Marc Serramia, Maite Lopez-Sanchez, and Juan Rodriguez-Aguilar, "Towards Pluralistic Value Alignment: Aggregating Value Systems through ℓ_p-Regression", Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022, May 9–13), https://filippobistaffa.github.io/papers/2022aamas.pdf.

[10] Suzanne Tolmeijer, Markus Kneer, Cristina Sarasua, Markus Christen, and Abraham Bernstein, "Implementations in machine ethics: A survey", *ACM Computing Surveys*, vol. 53, no. 6 (2020), article 132, https://doi.org/10.1145/3419633.

[11] Lynn Kaack, Priya Donti, Emma Strubell, George Kamiya, Felix Creutzig, and David Rolnick, "Aligning artificial intelligence with climate change mitigation", 2021, https://hal.archives-ouvertes.fr/hal-03368037.

[12] For example, an AI language processing system trained in a dominant language such as English will not function in other languages.

potential for biases in AI system outputs[13]; and (4) the deployment of AI systems, determining the specific societal and environmental impacts[14]. The breadth of ELS implications underscores the importance of embedding ELS at all points on the AI system lifecycle.

Work on ELS should additionally welcome a range of perspectives and support constructive and open debate. Recent work on AI ethics has emphasized the development of consensus-driven sets of ethics principles or guidelines. However, these guidelines come overwhelmingly from North American and European organizations, many of which had little or no public participation.[15] Marginalized populations, such as Indigenous peoples, may have differing perspectives on ELS issues.[16] Inclusion of diverse perspectives can help to overcome apparent gaps in existing ELS work, such as pertaining to the moral status of nonhumans.[17] Furthermore, active AI ELS debates remain unresolved, such as on the relative importance of near-term and long-term dimensions of AI.[18] Given the lack of universal consensus on AI ELS issues, it is important to support an inclusive and open-minded conversation about the issues.

Strategy 4: Ensure the safety and security of AI systems.

We recommend an emphasis on a dynamic research and development program to ensure that AI systems remain safe and secure as the technology and its usage evolve over time.

An essential aspect of AI safety and security is that as AI systems become more capable and become used more widely, the safety and security challenges become

[13] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan, "A survey on bias and fairness in machine learning", *ACM Computing Surveys*, vol. 54, no. 6 (2021), article 115, https://doi.org/10.1145/3457607.

[14] For example, decisions of when to deploy autonomous weapons; Hendrik Huelss, "Deciding on Appropriate Use of Force: Human-machine Interaction in Weapons Systems and Emerging Norms", Global Policy vol. 10, no. 3 (2019), pp. 354-358, https://doi.org/10.1111/1758-5899.12692.

[15] Daniel Schiff, Jason Borenstein, Justin Biddle, and Kelly Laas, "AI ethics in the public, private, and NGO sectors: a review of a global document collection", *IEEE Transactions on Technology and Society*, vol. 2, no. 1 (2021), pp. 31-42, https://doi.org/10.1109/TTS.2021.3052127.

[16] Jason Edward Lewis, Angie Abdilla, Noelani Arista, Kaipulaumakaniolono Baker, Scott Benesiinaabandan, Michelle Brown, Melanie Cheung et al., "Indigenous protocol and artificial intelligence position paper", Indigenous Protocol and Artificial Intelligence Working Group and the Canadian Institute for Advanced Research, Honolulu, HI (2020), https://spectrum.library.concordia.ca/id/eprint/986506.

[17] Andrea Owe and Seth D. Baum, "Moral consideration of nonhumans in the ethics of artificial intelligence", *AI and Ethics*, vol. 1, no. 4 (2021), pp. 517-528, https://doi.org/10.1007/s43681-021-00065-0.

[18] Charlotte Stix and Matthijs M. Maas, "Bridging the gap: the case for an 'Incompletely Theorized Agreement on AI policy", *AI and Ethics*, vol. 1, no. 3 (2021), pp. 261-271, https://doi.org/10.1007/s43681-020-00037-w.

more important. Prior to the deep learning revolution, AI technology had limited usage[19] and likewise little need for safety and security. By now, AI technology is used widely across economic sectors and other areas of human society. Barring a new AI winter, i.e. another years-long drop in AI research if AI progress fails to live up to expectations, AI technology will only grow in its importance. As it does, the potential for AI systems to cause harm is likely to increase. In risk terms, harm from AI systems could become more frequent, due to their wider usage, and more severe, due to their usage in more high-stakes applications. Prospects for catastrophic harm may further increase via the use of AI in crucial sectors such as agriculture[20] and via increasingly general-purpose AI systems that are becoming widely used across sectors.[21]

Given the growing stakes, it is vital for AI safety and security to keep up with the changing technology. This is not a trivial challenge. Some aspects of AI safety and security may remain viable even as the technology becomes more advanced,[22] whereas other aspects may need to be customized for more advanced systems.[23] Research and development on AI safety and security must be forward-looking in order to keep society safe in the face of more capable and more widely deployed AI systems. Doing so will further be of economic and strategic benefit because safer and more secure AI technology would permit the technology to be deployed more widely, especially in more sensitive settings.

Strategy 5: Develop shared public datasets and environments for AI training and testing.

We have no comments on Strategy 5.

Strategy 6: Measure and evaluate AI technologies through standards and benchmarks.

[19] Terrence J. Sejnowski, *The Deep Learning Revolution* (MIT Press, 2018).

[20] Victor, Galaz, Miguel A. Centeno, Peter W. Callahan, Amar Causevic, Thayer Patterson, Irina Brass, Seth Baum et al., "Artificial intelligence, systemic risks, and sustainability", *Technology in Society*, vol. 67 (2021), article 101741, https://doi.org/10.1016/j.techsoc.2021.101741.

[21] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein et al., "On the opportunities and risks of foundation models", Stanford Institute for Human-Centered Artificial Intelligence (2021), https://arxiv.org/abs/2108.07258.

[22] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané, "Concrete problems in AI safety", (2016), https://arxiv.org/abs/1606.06565.

[23] Stuart Russell, *Human Compatible: Artificial Intelligence and the Problem of Control* (Viking, 2019).

We recommend support for programs that can facilitate the ongoing development and adoption of standards across the AI industry, including via the development of regimes for the certification of AI systems.

A variety of AI standards and frameworks are in the process of being developed, including the National Institute of Standards and Technology (NIST) AI Risk Management Framework[24] and the Institute of Electrical and Electronics Engineers (IEEE) Standards Association P2863 Recommended Practice for Organizational Governance of Artificial Intelligence.[25] These standards and frameworks will provide voluntary guidance for developers and deployers of AI systems. As these and other frameworks are completed, two challenges will be faced. One is to ensure that the standards remain relevant and appropriate in the face of ongoing changes in AI technology and its various applications. The other is to facilitate the adoption of the standards by AI developers and deployers.

One valuable approach to facilitating adoption of standards is via certification regimes.[26] Certification regimes serve to address information asymmetries between insiders within an organization and outsiders who wish to know if the organization is complying with relevant standards. Certification is already widely used in many sectors, such as in the US EnergyStar program for consumer appliances, the ISO 9001 program for supply chains, and the LEED program for building design. AI certification regimes have been developed or proposed by, among others, the European Commission,[27] IEEE,[28] and the government of Malta.[29] When implemented effectively, certification regimes can incentivize adoption of standards. Certification regimes are also flexible in terms of being public or private (or both), voluntary or mandatory, and geographically local or global. These attributes make certification regimes an important element of AI standards adoption.

Strategy 7: Better understand the national AI R&D workforce needs.

[24] https://www.nist.gov/itl/ai-risk-management-framework

[25] https://sagroups.ieee.org/2863

[26] Peter Cihon, Moritz J. Kleinaltenkamp, Jonas Schuett, and Seth D. Baum, "AI certification: Advancing ethical practice by reducing information asymmetries", *IEEE Transactions on Technology and Society*, vol. 2, no. 4 (2021), pp. 200-209, https://doi.org/10.1109/TTS.2021.3077595.

[27] European Commission, “White paper on artificial intelligence: A European approach to excellence and trust,” (2020), https://ec.europa.eu/info/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en.

[28] IEEE, “The ethics certification program for autonomous and intelligent systems (ECPAIS)”, https://standards.ieee.org/industry-connections/ecpais.html.

[29] Malta Digital Innovation Authority, “AI-ITA blueprint guidelines”, (2019), https://mdia.gov.mt/wp-content/uploads/2019/10/AI-ITA-Blueprint-Guidelines-03OCT19.pdf.

We recommend an emphasis on a multidisciplinary and multitalented workforce that is capable of developing and applying AI technology for the common good.

As AI technology grows in its ethical, legal, and societal (ELS) significance, it is vital for the computer scientists and engineers who design and deploy AI systems to be conversant in ELS topics. Historically, AI technology had little societal impact; AI computer scientists and engineers were likewise focused narrowly on how to increase the capabilities of AI systems with little regard for ELS implications.[30] The emergence of AI as a class of technology with significant ELS implications has created a need for cultivating an understanding of ELS among AI computer scientists and engineers. Some progress on this front has been made,[31] but this remains a major weakness of the AI workforce.

Concurrently, there is a need for ELS experts who are conversant in AI technology. Computer scientists and engineers play a vital role in AI technology design and in relating design details to AI governance. However, it is inappropriate to expect computer scientists and engineers to have a comparable depth of knowledge about ELS as people who are professionally trained in ELS fields such as moral philosophy, law, and the social sciences. In order for ELS experts to successfully contribute to AI governance, it is vital for them to have some knowledge (the more the better) of how AI technology works. This is not typically taught in ELS university programs. It is therefore important to support dedicated programs. The Technology, Management, and Policy Consortium[32] provides a valuable set of benchmarks for how such programs can be designed and executed.

Strategy 8: Expand public-private partnerships to accelerate advances in AI.

We recommend that public-private partnerships be pursued both domestically and internationally to ensure that the AI industry as a whole is oriented toward the common good.

In AI, as is the case in many industries, there is often tension between the private interest and the public good. For example, as noted above, companies running social media websites sometimes use AI recommender algorithms that increase user

[30] John Bohannon, "Fears of an AI pioneer", *Science*, vol. 349, no. 6245 (2015), p.252, https://doi.org/10.1126/science.349.6245.252.
[31] For example, the AAAI/ACM Conference on AI, Ethics, and Society, https://www.aies-conference.com.
[32] https://tmpconsortium.org

engagement by promoting politically extremist content[33] and AI developers sometimes build larger neural network models to improve model accuracy without regard for the adverse energy resource and climate change consequences of larger models.[34] Additionally, in the international context, competition between military adversaries can result in both sides pursuing unsafe AI technology.[35] These situations are known as collective action problems; they constitute an important class of challenges in AI governance.[36]

Public-private partnerships can play a valuable role in addressing AI collective action problems. Domestically, these partnerships can help support private firms as they orient their activities toward the common good. Internationally, the partnerships can facilitate the cooperation needed to solve AI collective action problems at the global scale. The US already participates in international public-private partnerships for mutual benefit with other countries.[37] Additionally, some international organizations already work to bring together public and private AI stakeholders; these include the Global Partnership on Artificial Intelligence (GPAI)[38] and the OECD Artificial Intelligence Policy Observatory.[39] The US should participate in these organizations as part of its program on public-private partnerships. The US should further seek to include diverse participants from the international community, including to support global justice in AI technology.[40] Additionally, where appropriate, including rival powers such as China in these forums would further facilitate the resolution of AI collective action problems. If successful, these partnerships could support a “race to the top” dynamic in which competition between companies and countries advances benefits for the national and global common good.[41]

---

[33] Joe Whittaker, Seán Looney, Alastair Reed, and Fabio Votta, "Recommender systems and the amplification of extremist content", *Internet Policy Review*, vol. 10, no. 2 (2021), https://doi.org/10.14763/2021.2.1565.

[34] Lynn Kaack, Priya Donti, Emma Strubell, George Kamiya, Felix Creutzig, and David Rolnick, "Aligning artificial intelligence with climate change mitigation", 2021, https://hal.archives-ouvertes.fr/hal-03368037.

[35] Richard Danzig, "Managing Loss of Control as Many Militaries Pursue Technological Superiority", Center for New American Security (2018), https://www.cnas.org/publications/reports/technology-roulette.

[36] Robert de Neufville and Seth D. Baum, "Collective action on artificial intelligence: A primer and review", *Technology in Society*, vol. 66 (2021), article 101649, https://doi.org/10.1016/j.techsoc.2021.101649.

[37] For example, the US-Israel Binational Industrial Research and Development Foundation (BIRD), established in 1977, aims to provide funding to joint projects of mutual benefit to both the US and Israel; see https://www.birdf.com.

[38] https://gpai.ai

[39] https://oecd.ai/en

[40] Eugenio V. Garcia, "The International Governance of AI: Where is the Global South?" (2021), https://www.researchgate.net/publication/348848134.

[41] Will Hunt, "The Flight to Safety-Critical AI: Lessons in AI Safety from the Aviation Industry", Center for Long-Term Cybersecurity, University of California, Berkeley (2020), https://cltc.berkeley.edu/2020/08/11/new-report-the-flight-to-safety-critical-ai-lessons-in-ai-safety-from-the-aviation-industry.