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

Dear 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,

We thank the OSTP, the National Science and Technology Council's (NSTC) Select Committee on Artificial Intelligence, the NSTC Machine Learning and AI Subcommittee, the National AI Initiative Office, and the Networking and Information Technology Research and Development National Coordination Office for the opportunity to submit comments in response to the update of the National Artificial Intelligence Research and Development Strategic Plan. We are professors and researchers with expertise in AI research and development, policy, and ethics, affiliated with centers at the University of California, Berkeley, including the Berkeley AI Research Lab; the Division of Computing, Data Science, and Society; the AI Policy Hub; the Center for Long-Term Cybersecurity and its AI Security Initiative; the CITRIS Policy Lab; the CITRIS and the Banatao Institute; the Center for Human-Compatible AI; the Human Rights Center at the UC Berkeley School of Law; as well as external technology and governance non-profit research organizations including the Future of Life Institute, the Digital Life Initiative at Cornell Tech, and The Future Society.

In this document, we affirm the continued importance of the eight strategic aims described in the 2019 Update. However, we advocate for modest changes to each aim that take into account the continued learning across the AI R&D landscape. Lastly, we advocate for the inclusion of a ninth strategy—one that draws attention to the need for research on transparency and documentation of AI systems and applications. We believe this strategy is a necessary addition to support responsible and sustainable advances in this technology. Our recommendations are intended to help ensure the National AI R&D Strategic Plan enables sustained technological innovation, supports broad inclusion, economic prosperity, and national security, and upholds essential democratic values.

We have included a one-sentence summary of our main recommendation for each strategy below.

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

We encourage a strengthened focus on multidisciplinary research that supports AI robustness, ethics, transparency, and security integrated with long-term investments in fundamental research.

We agree that long-term investments in fundamental research are needed to continue building on previous discoveries in AI. Specifically, we advocate that the sustained funding of R&D is an essential element that advances the trust in AI systems necessary to ensure they meet society's needs and adequately address requirements for robustness, ethics, transparency, and security. However, these research threads should not be seen as disparate, but as mutually reinforcing and essential to the development of AI.[1] AI advances are always socially driven. We should not seek advances at all costs, but in such a way that is safe, secure, responsible, and ethical. We note that research on general-purpose and scalable, multi-AI systems should be pursued cautiously and with these properties at the forefront, given the extreme potential risks from such systems.

For example, one of the subsections of research encouraged in the 2019 Update is the development of more capable and reliable robots. Indeed, it is not helpful or desirable simply to have more capable robots if they are not also reliable. In fact, the more capable the robot, the greater people may come to depend on or interact with it, implying the need for higher reliability and trustworthiness. This is particularly apparent in the context of domain-specific applications–a "reliable and safe" autonomous drone is unlikely to interact physically with humans as frequently as a self-driving car or Amazon warehouse robot.

Our point is consistent with the legal guidance of the National AI Initiative Act, in which Congress specifically states that the "United States government should use this Initiative to enable the benefits of trustworthy artificial intelligence while preventing the creation and use of artificial intelligence systems that behave in ways that cause harm."[2] The National Science Foundation (NSF) is additionally called upon to work on "research areas that will contribute to the development and deployment of trustworthy artificial intelligence systems."

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

We encourage greater focus on assessing the appropriateness of varying human-machine teaming arrangements and on understanding the associated human labor implications.

We emphasize the importance of trust and alignment in enabling human-AI collaboration. As described in the 2016 Plan and mentioned in the 2019 Update: "Appropriate trust of AI systems requires explainability, especially as the AI grows in scale and complexity. … This research area reflects the intersection of Strategies 2 and 3, as explainability, fairness, and transparency are key principles for AI systems to effectively collaborate with humans. Likewise, the challenge of understanding and designing human-AI ethics and value alignment into systems remains an open research area."

[1] Sheila Jasanoff, "Ordering Knowledge, Ordering Society," in Jasanoff, ed., *States of Knowledge*, pp. 13-45. http://sheilajasanoff.stsprogram.org/wp-content/uploads/Jasanoff_Ordering-KnowledgeOrdering-Society.pdf.

[2] National Defense Authorization Act for Fiscal year 2021. HR 6395. Division E - National Artificial Intelligence Initiative Act of 2020. https://www.congress.gov/116/crpt/hrpt617/CRPT-116hrpt617.pdf#page=1210.

We encourage additional investment in research on human-AI collaboration, including technology and policy strategies that may be pursued to support greater efficiency, effectiveness, and equity. We appreciate that the National AI R&D Strategic Plan outlines key areas of research, including where AI performs functions alongside humans, in instances where humans experience high cognitive load, and in lieu of humans where they have limited capabilities. Additional research is needed on human-machine teaming arrangements and the safeguards that must be in place to ensure that they function safely and without undue risk. This is an example of an area where closer coordination between DARPA, USD (R&E), and the National AI Initiative Office could support research advances, promote shared learning, and ensure maximum benefit from taxpayer dollars to support AI R&D, in both defense and civilian contexts. Furthermore, there should be support for efforts geared toward understanding the human labor impacts, including the toll on workers asked to interact with and rely on AI systems as well as workers involved in the development of AI systems such as data annotators[3] or UX and UI professionals.[4]

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

We encourage strengthened research and transparency in the integration of ethical, legal, and societal concerns throughout all stages of the AI lifecycle, as well as on the detection of malicious uses of AI including potential human rights abuses.

This strategy remains critical, and we underscore the importance of enabling more R&D resources that target the integration of ethical, legal, and societal concerns throughout all stages of the AI lifecycle, rather than simply after development or deployment. We also highlight the importance of research on varying interpretations of relevant, but contested terms such as "fairness" and "explainability" and their application in practice.[5,6,7] In addition to ethical, legal, and societal concerns, research related to the politics, justice, equity, and environmental implications of AI has flourished in recent years, but needs greater investment to ensure the insights from these fields can thoughtfully inform and be integrated from design through deployment and monitoring. This includes forming technology and governance oversight strategies that can be implemented throughout the AI system's lifecycle. Transparency will be critical here as value judgments will be incorporated into how technologists define and encode "ethical doctrine" (see p. 22 in AI R&D Strategic Plan). Additional research on how the human

[3] Milagros Miceli, Martin Schuessler, and Tianling Yang, "Between Subjectivity and Imposition: Power Dynamics in Data Annotation for Computer Vision," *Proc. ACM Hum.-Comput. Interact. 4*, CSCW2, Article 115 (October 2020), https://doi.org/10.1145/3415186.

[4] Richmond Wong, "Tactics of Soft Resistance in User Experience Professionals' Values Work," *Proceedings of the ACM on Human-Computer Interaction,* (October 2021): 1–28, https://doi.org/10.1145/3479499.

[5] Deirdre K. Mulligan, Joshua A. Kroll, Nitin Kohli, and Richmond Y. Wong, "This Thing Called Fairness: Disciplinary Confusion Realizing a Value in Technology." *Proc. ACM Hum.-Comput. Interact. 3, CSCW, Article 119* (November 2019), https://doi.org/10.1145/3359221.

[6] Jessica Newman, "Explainability won't save AI," *Brookings TechStream*. (May 19, 2021). www.brookings.edu/techstream/explainability-wont-save-ai/.

[7] Nicole Chi, Emma Lurie, and Deirdre K. Mulligan, (July 2021). "Reconfiguring Diversity and Inclusion for AI Ethics," AIES '21: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. https://dl.acm.org/doi/10.1145/3461702.3462622.

rights legal framework and norms can be used to guide ethical AI development and deployment is also needed.[8]

Lastly, an additional research challenge in this area that urgently requires greater investment is the detection of malicious uses of AI including the use of synthetic content for manipulation, harassment, financial, and political gain.[9]

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

We encourage strengthened research on how to manage and prevent safety and security challenges from increasing as AI systems become more advanced and multiply their capabilities, including the role of greater transparency and public awareness.

Technical solutions to prominent AI safety and security problems remain elusive and are a critical issue that requires federal R&D investments along with collaborative efforts among government, industry, academia, and civil society. It is imperative that the National Institute of Standards and Technology (NIST) adheres to the legal guidance in the National AI Initiative Act to support research on "safety and robustness of artificial intelligence systems, including assurance, verification, validation, security, control, and the ability for artificial intelligence systems to withstand unexpected inputs and adversarial attacks."[10] As stated in the 2019 Update, state-of-the-art AI systems today can still "be made to do the wrong thing, learn the wrong thing, or reveal the wrong thing, for example, through adversarial examples, data poisoning, and model inversion, respectively." This is particularly pressing for the application of AI technologies in critical infrastructure, defense, and safety-critical systems.

Moreover, we agree that as AI systems continue to grow in capabilities, they will likely grow in complexity, making it ever harder for correct and desirable performance to be verified and validated.[11] AI safety and value alignment remain critical research challenges, especially for

[8] David Kaye, special rapporteur on the promotion and protection of the right to freedom of opinion and expression, "Report on artificial intelligence technologies and implications for freedom of expression and the information environment." *United Nations Office of the High Commissioner for Human Rights.* https://www.ohchr.org/EN/Issues/FreedomOpinion/Pages/ReportGA73.aspx.

[9] Shruti Agarwal, Hany Farid, Yuming Gu, Mingming He, Koki Nagano, and Hao Li. "Protecting world leaders against deep fakes," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. (2019) https://openaccess.thecvf.com/content_CVPRW_2019/papers/Media%20Forensics/Agarwal_Protecting_World_Leaders_Against_Deep_Fakes_CVPRW_2019_paper.pdf.

[10] National Defense Authorization Act for Fiscal year 2021. HR 6395. Division E - National Artificial Intelligence Initiative Act of 2020. https://www.congress.gov/116/crpt/hrpt617/CRPT-116hrpt617.pdf#page=1210.

[11] Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, von Arx S, Bernstein MS, Bohg J, Bosselut A, Brunskill E, Brynjolfsson E, Buch S, Card D, Castellon R, Chatterji N, Chen A, Creel K, Davis JQ, Demszky D, Donahue C, Doumbouya M, Durmus E, Ermon S, Etchemendy J, Ethayarajh K, Fei-Fei L, Finn C, Gale T, Gillespie L, Goel K, Goodman N, Grossman S, Guha N, Hashimoto T, Henderson P, Hewitt J, Ho DE, Hong J, Hsu K, Huang J, Icard T, Jain S, Jurafsky D, Kalluri P, Karamcheti S, Keeling G, Khani F, Khattab O, Kohd PW, Krass M, Krishna R, Kuditipudi R, Kumar A, Ladhak F, Lee M, Lee T, Leskovec J, Levent I, Li XL, Li X, Ma T, Malik A, Manning CD, Mirchandani S, Mitchell E, Munyikwa Z, Nair S, Narayan A, Narayanan D, Newman B, Nie A, Niebles JC, Nilforoshan H, Nyarko J, Ogut G, Orr L, Papadimitriou I, Park JS, Piech C, Portelance E, Potts C, Raghunathan A, Reich R, Ren H, Rong F, Roohani Y, Ruiz C, Ryan J, Ré C, Sadigh D, Sagawa S, Santhanam K, Shih A, Srinivasan K, Tamkin A, Taori R, Thomas AW, Tramèr F, Wang RE, Wang W, Wu B, Wu J, Wu Y, Xie SM, Yasunaga M, You J, Zaharia M, Zhang M, Zhang T, Zhang X, Zhang Y, Zheng L, Zhou K, and Liang P, "On the Opportunities and Risks of Foundation Models," *arXiv,* (2021), https://arxiv.org/abs/2108.07258.

multi-purpose or general-purpose AI systems,[12,13] as stated in both the 2016 Plan and 2019 Update. We expect these challenges will increase in the near future, as AI systems become more advanced and multiply their capabilities, with both greater beneficial opportunities and risks in case of misuse or failures of safety or security controls.

We believe that greater transparency and public awareness are needed to support AI safety and security. End-users should have an understanding of the safety and security of AI systems and supporting accountability mechanisms, including clear steps for redress. Research is necessary on how to do this effectively. We also advocate for studying the kinds of vulnerabilities and failures that are likely to arise from real-world threat scenarios, and from software vulnerabilities in the AI supply chain.

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

We encourage research on how to reduce energy and carbon footprints for AI development and operation, and the role of public training and testing environments in that reduction.

The trend toward larger and more complex AI models, requiring larger training datasets and significant computing resources, has increased in recent years. This trend typically benefits already powerful companies and institutions, and comes with a significant and often-unsustainable environmental cost.[14,15] More research is needed to better understand how to reduce energy and carbon footprints for AI development and operation, and the role of public training and testing environments in that reduction.[16]

Shared public datasets and secure environments for AI training and testing are an important way to ensure that progress in AI meets the needs of a diverse spectrum of AI interests and applications and can support the public good. Public datasets and environments for AI training and testing can also offer secure software sandboxes, regulatory sandboxes, and testing servers. By creating shared datasets and secure environments for cross-institutional testing, a greater diversity of innovators, entrepreneurs, SMEs in various sectors, and researchers from varying epistemological approaches may be supported.

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

[12] Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt, "Unsolved Problems in ML Safety," *arXiv*, https://arxiv.org/abs/2109.13916.

[13] Stuart Russell, *Human Compatible: Artificial Intelligence and the Problem of Control.* (New York: Viking, 2020).

[14] Emma Strubell, Ananya Ganesh, Andrew McCallum, "Energy and Policy Considerations for Deep Learning in NLP," In the 57th Annual Meeting of the Association for Computational Linguistics (ACL). Florence, Italy. (July 2019). arxiv.org/abs/1906.02243.

[15] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell, "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.* (March 2021). dl.acm.org/doi/abs/10.1145/3442188.3445922.

[16] David Patterson, Joseph Gonzalez, Urs Hölzle, Quoc Hung Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeffrey Dean. (2022), "The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink," *TechRxiv*. Preprint. https://doi.org/10.36227/techrxiv.19139645.v2.

We encourage research that investigates how standards, benchmarks, and testing requirements for a broad set of quality controls will inform evolving AI development and deployment, and how to encourage adoption.

Ongoing efforts to measure and evaluate AI technologies through standards and benchmarks, for example by the National Institute of Standards and Technology (NIST), the International Organization for Standardization (ISO), and the Institute of Electrical and Electronics Engineers (IEEE), are extremely valuable. However, as noted in the 2019 Update, we agree that benchmarks, metrics, and testing requirements for a broad set of quality controls are still lacking and require greater research investment. Specifically, we argue that while benchmark datasets are important, they should be alive – in the sense that they need to be enhanced by new data and connected to domain problems with human committees and evaluators, not just metric numbers, which should serve as supporting information.

We also caution that establishing standards and benchmarks can lead to lock-in and path dependencies for AI system development and deployment that will be difficult to circumvent. If these processes are too rigid and resource intensive, they may lead to workarounds, lack of compliance, and other harmful spillover effects. It is therefore of great importance that standards and benchmarks are robust yet flexible enough to adapt to changing norms and needs, including how to draw upon a human rights legal framework, which puts humans' civil, political, economic, and social wellbeing—as opposed to institutional benefits—at the center of development.[17] We encourage funding allocation to support research in this space, especially the utility of the NIST AI Risk Management Framework (NIST AI RMF) to better ensure its long-term success. While NIST's AI RMF is voluntary, it would benefit from research on how AI governance testbeds may be used to evaluate its effectiveness at shaping AI development and deployment.

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

We emphasize the need to not only broaden participation in computing and engineering fields, but also to provide educational opportunities to train computer scientists and engineers to be fluent in social and ethical impact, and in professional responsibility.

As noted in the 2019 Update, we agree that multidisciplinary teams are essential to a thriving AI R&D workforce, and that "it is imperative to broaden the participation among groups traditionally underrepresented in computing and related fields." We emphasize that many different definitions of "underrepresented" may be in scope here–the inclusion of foreign

[17] Brandie Nonnecke and Phil Dawson, "Human rights implications of algorithmic impact assessments: Priority considerations to guide effective development and use". Harvard Carr Center Discussion Paper Series. (Oct. 21, 2021), https://carrcenter.hks.harvard.edu/publications/human-rights-implications-algorithmic-impact-assessments-priority-considerations.

researchers, ethnic minorities, women, representatives of the LGBTQ+ community, the differently abled, and other groups historically marginalized within the disciplinary culture of computer science and engineering. The integration of feedback from diverse stakeholder groups at multiple points of AI development is a recognized path to system reliability and safety.[18] In addition to providing education in computational thinking at all levels across disciplines, we emphasize the need for educational materials and opportunities to help train computer and information scientists and engineers to be fluent in social and ethical impact, and in professional responsibility.[19,20]

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

We encourage increased focus on international cooperation and coordination on AI research as well as support for research partnerships that include civil society and impacted communities.

International cooperation and coordination on AI is increasingly critical and we advocate maintaining and expanding this emphasis. For example, further research is needed to advance opportunities for collaboration with allies to improve information sharing, reduce potential "race-to-the-bottom" dynamics, and design Track I, 1.5, and II diplomacy mechanisms.

In addition to partnerships with academia and industry that generate technological breakthroughs in AI, we also recommend the inclusion of partnerships with civil society and impacted communities to ensure applications of AI achieve their aims and do not cause unexpected or disproportionate harm.

[New] Strategy 9: Support transparency and documentation of AI systems and applications.

We encourage support for research that identifies effective mechanisms for transparency and documentation of AI systems and applications.

We argue an additional strategic aim is warranted and therefore propose a ninth strategy to support transparency and documentation of AI systems and applications. The need for ongoing research on the transparency and effective explainability of AI systems is already discussed in both Strategy 3 and Strategy 4. However, research into how to document and share the characteristics of AI systems is a current gap in the R&D Plan. While there has been critical

[18] Roel Dobbe, Thomas Krendl Gilbert, and Yonatan Mintz, "Hard choices in artificial intelligence," *Artificial Intelligence* 300 (2021): 103555. https://doi.org/10.1016/j.artint.2021.103555.

[19] Barbara J. Grosz, David Gray Grant, Kate Vredenburgh, Jeff Behrends, Lily Hu, Alison Simmons, and Jim Waldo, "Embedded EthiCS: integrating ethics across CS education," *Communications of the ACM*, *62*, no. 8, (Oct. 29, 2019): 54-61. https://doi.org/10.1145/3330794.

[20] Amy J. Ko, Alannah Oleson, Neil Ryan, Yim Register, Benjamin Xie, Mina Tari, Matthew Davidson, Stefania Druga, and Dastyni Loksa, "It is time for more critical CS education," *Communications of the ACM*, *63*, vol. 11 (2020): 31-33. https://doi.org/10.1145/3424000.

research in this space in recent years,[21,22,23,24] there is ongoing need for research on how best to carry out and facilitate standardized descriptions of features of AI systems. Some of the types of descriptions that may be relevant are characteristics about the AI system, its performance metrics, and its outcomes including expected behaviors, limitations, evaluation across varying conditions and populations, information about which datasets and training environments have been used and why, as well as human-interpretable logging of a system's activity, metadata, and impacts. Further research is also needed to explore processes that support these activities, which include verification of the characteristics over time, internal reviews, and reporting mechanisms. Improving classification and documentation of AI systems and applications should be a research priority because the current lack of standardization contributes to the dearth of trust in AI development, preventing increased discovery and adoption.[25,26,27] Moreover, this is an area that would benefit from federal investment because industry is unlikely to address this on its own and because it may facilitate greater coordination and communication between organizations, disciplines, and sectors.

We understand that the National AI R&D Strategic Plan is, by design, solely concerned with addressing the research and development priorities associated with advancing AI technologies, and does not describe or recommend policy or regulatory actions related to the governance or deployment of AI. The call for increased focus on transparency and documentation of AI systems is oriented toward supporting research and development. Without institutionalized mechanisms for sharing the types of tools being built and used for different purposes, it is more challenging to share knowledge and learn from the experiences of others.

The 2019 Executive Order on Maintaining American Leadership in Artificial Intelligence called on federal agencies to improve their data and model inventory documentation to enable discovery and usability, and the 2019 Update emphasized in Strategy 5 that, "development and adoption of best practices and standards in documenting dataset and model provenance will enhance trustworthiness and responsible use of AI technologies." However, Strategy 5 is primarily focused on improving access to datasets and training environments rather than documenting the characteristics and uses of AI systems. Adding a new strategy to support transparency and documentation of AI systems and applications will not only accelerate research

[21] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, Timnit Gebru, "Model Cards for Model Reporting," *FAT* '19: Proceedings of the Conference on Fairness, Accountability, and Transparency* (2019), https://dl.acm.org/doi/10.1145/3287560.3287596.

[22] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. "Datasheets for Datasets," arxiv.org/abs/1803.09010.

[23] Bin Yu and Karl Kumbier, "Veridical data science," *PNAS.* 117, no. 8 (Feb. 25, 2020): 3920-3929. https://doi.org/10.1073/pnas.1901326117.

[24] W. James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu, "Definitions, methods, and applications in interpretable machine learning," *PNAS,* 116, no. 44, (Oct. 29, 2019): 22071-22080. https://doi.org/10.1073/pnas.1900654116.

[25] "OECD Framework for Classification of AI Systems: a tool for effective AI policies," OECD Digital Economy Papers. (Feb. 2022.) https://www.oecd-ilibrary.org/science-and-technology/oecd-framework-for-the-classification-of-ai-systems_cb6d9eca-en.

[26] Catherine Aiken, "Classifying AI Systems," CSET Data Brief. (Nov. 2021.) https://cset.georgetown.edu/publication/classifying-ai-systems/.

[27] Thomas Krendl Gilbert, Sarah Dean, Tom Zick, and Nathan Lambert. (Feb. 2022), "Choices, Risks, and Reward Reports: Charting Public Policy for Reinforcement Learning Systems," *Center for Long-Term Cybersecurity White Paper Series.* https://cltc.berkeley.edu/reward-reports/.

in this critical area, but also advance the aims of the other eight strategies by contributing to knowledge of the AI landscape.

Contact

Thank you for the opportunity to comment on the National Artificial Intelligence Research and Development Strategic Plan. If you need additional information or would like to discuss further, please contact Jessica Newman at jessica.newman@berkeley.edu.

Our best,

Anthony M. Barrett, Ph.D., PMP, Visiting Scholar, AI Security Initiative, Center for Long-Term Cybersecurity, UC Berkeley

Ann Cleaveland, Executive Director, Center for Long-Term Cybersecurity, UC Berkeley

Camille Crittenden, Ph.D., Executive Director, CITRIS and the Banatao Institute, UC Berkeley; Co-Founder, CITRIS Policy Lab and EDGE in Tech Initiative at UC

Samuel Curtis, AI Policy Researcher & Project Manager, The Future Society

Jordan Famularo, Ph.D., Postdoctoral Scholar, Center for Long-Term Cybersecurity, UC Berkeley

Hany Farid, Ph.D., Professor, Electrical Engineering and Computer Sciences and the School of Information, UC Berkeley

Thomas Krendl Gilbert, Ph.D., Research Affiliate, Center for Human-Compatible AI, UC Berkeley; Digital Life Initiative, Cornell Tech

Ken Goldberg, Ph.D., Professor, Industrial Engineering and Operations Research William S. Floyd Jr. Distinguished Chair in Engineering, UC Berkeley

Carlos Ignacio Gutierrez, AI Policy Researcher, Future of Life Institute

Dan Hendrycks, Ph.D. Candidate, Berkeley AI Research Lab, UC Berkeley

Niki Iliadis, Senior AI Policy Researcher, The Future Society

Alexa Koenig, J.D., Ph.D., Executive Director, Human Rights Center, UC Berkeley School of Law; Co-Founder, Human Rights Investigations Lab

Yolanda Lannquist, Head of Research & Advisory, The Future Society

Richard Mallah, Director of AI Projects, Future of Life Institute

Nicolas Miailhe, Founder & President, The Future Society

Nicolas Moës, Head of Operations & AI Policy Researcher, The Future Society

Jessica Newman, Director, AI Security Initiative, Center for Long-Term Cybersecurity, UC Berkeley; Co-Director, AI Policy Hub

Brandie Nonnecke, Ph.D., Director, CITRIS Policy Lab, CITRIS and the Banatao Institute, UC Berkeley; Co-Director, AI Policy Hub

Ifejesu Ogunleye, Researcher, AI Security Initiative, Center for Long-Term Cybersecurity, UC Berkeley

Andrew W. Reddie, Ph.D., Assistant Professor of Practice, School of Information, UC Berkeley

Stuart Russell, Ph.D., Professor Computer Science and Smith-Zadeh Professor in Engineering, UC Berkeley

Charis Thompson, Ph.D., Chancellor's Professor and Associate Dean, Computing, Data Science, and Society, UC Berkeley

Richmond Y. Wong, Ph.D., Postdoctoral Scholar, Center for Long-Term Cybersecurity, UC Berkeley

Bin Yu, Ph.D., Chancellor's Distinguished Professor, Departments of Statistics and Electrical Engineering and Computer Sciences, Class of 1936 Second Chair, L&S, UC Berkeley

Rebecca Wexler, J.D., Assistant Professor, School of Law, UC Berkeley