DATA&
SOCIETY

POLICY BRIEF

AI Governance Needs Sociotechnical Expertise

Why the Humanities and Social Sciences Are Critical to Government Efforts

Serena Oduro
Tamara Kneese

Successful AI governance requires expertise in the *sociotechnical* nature of AI systems.[1] Because real-world uses of AI are always embedded within larger social institutions and power dynamics, technical assessments alone are insufficient to govern AI. Technical design, social practices and cultural norms, the context a system is integrated in, and who designed and operates it all impact the performance, failure, benefits, and harms of an AI system.[2] As the National Institute of Standards and Technology's (NIST) AI Risk Management Framework (RMF), which was designed to provide guidance to organizations developing, deploying, or using AI systems, rightly notes, "AI systems are inherently sociotechnical in nature, meaning they are influenced by societal dynamics and human behavior."[3] Sociotechnical research, "an approach to studying technologies within their social, political, economic, and cultural contexts,"[4] can improve system design,[5] reveal discrimination,[6] and advance accountability in ways that a purely technical approach to evaluating AI systems cannot.[7]

Putting this knowledge into practice will require the expertise of those in the humanities and social sciences who are equipped to conduct research on the interplay between AI systems and society. AI governance and risk management efforts that neglect to engage with experts focused on society, culture, politics, and the economy may produce AI systems that technically perform as intended but "fail" when deployed in the real world.[8] Early research has shown that when sociotechnical approaches are integrated into AI development and testing and in use-feedback, positive outcomes significantly increase for impacted communities, users, and AI developers.[9] Sociotechnical research and approaches have proven crucial to AI development and accountability — the key will be implementing AI governance practices that employ the expertise required to reap these benefits.

This policy brief explores the importance of integrating humanities and social science expertise into AI governance. We recommend this brief to any policymaker, researcher, or other stakeholder — in the US or abroad — working to advance AI safety tools, conduct qualitative work on algorithmic impact assessments or other AI assessments,[10] craft AI procurement standards, or engage in AI hiring efforts.[11] This brief is not meant to devalue technical interventions in AI governance; rather, we aim to highlight the full gamut of expertise needed to create an accountable AI ecosystem, including perspectives outside of STEM disciplines.

The Benefits Humanities and Social Science Can Bring to Our Current Moment

Advancing AI safety requires governance structures and policy solutions that recognize the sociotechnical nature of AI systems.[12] Notably, the field of safety engineering has long integrated a sociotechnical lens, recognizing that safety cannot be guaranteed when social components are ignored.[13] Analyzing fatalities caused by public testing of autonomous vehicles as a case study, for example, Inioluwa Deborah Raji and Roel Dobbe have described how a strictly technical engineering perspective often misses the true causes of system failures. Reports they describe by the National Transportation Safety Boards (NTSB) on two autonomous vehicle crashes concluded that in both cases, vehicles featured appropriate technical safety measures, and it was the fact that human operators rarely used those functions, or did so hesitantly, that caused the crashes[14] Safety is about more than sound design of technical

functions; it must also include practices like stakeholder engagement and a consideration of societal interactions at a broader scale to ensure that AI systems are safe from the perspective of the communities who will be impacted by them once they are deployed.[15]

Below, we outline some of the ways humanities and social science methods and expertise can help us to assess the performance and mitigate the harms of AI systems.

1. Generative AI Assessment

Historically, methods meant to identify and mitigate algorithmic harm have been predicated on specific use cases. Generative AI upsets this paradigm, with one multi-purpose model often deployed across a wide array of use cases. Where traditional machine learning approaches can systematically catalog the impacts of a system for a particular setting and set of impacted stakeholders, generative AI — especially the most powerful "foundation models" — are not tethered to a predefined set of uses and users.[16]

This expansiveness of generative AI means it is difficult to ensure comprehensive assessment and responsible evaluation through strictly technical methodologies. Humanities experts, social scientists, and user experience (UX) researchers and designers offer a tangible starting point for expanding the sociotechnical analysis of generative AI and addressing its technical and social dimensions simultaneously.[17] For example, University of Washington researchers applied the method of autoethnography to track their members' use of generative AI and associated self-reflections over several months, using this qualitative, ethnographic study to determine the technology's accessibility. Their analysis uncovered subtle ableism in some AI-generated results. To achieve a deeper understanding of the impacts of generative AI and articulate approaches to govern its use, many forms of disciplinary expertise must be assembled.

2. Auditing and Assessing Impacts

It is imperative that AI audits and impact assessments center humanities and social science expertise that enables the thorough investigation of AI-powered systems' impact on society. Technical approaches alone cannot fully capture AI's real-world benefits and harms, and can also frustrate efforts to promote transparency and engage the broader public. For example, it will be difficult to holistically assess algorithmic impacts in the public interest if developers audit their own AI systems and document only a narrow set of technical concerns.[18] Transparency and accountability are harder to achieve if assessments are only crafted by and decipherable to technical experts. Assessments should encourage consultation with historically marginalized communities, identify preventable harms, and standardize the information available for further research about which AI systems are used in which contexts and for which purposes.

Broader sociotechnical evaluations engage multidisciplinary stakeholders, who can identify the impacts of existing and emerging technologies more comprehensively. A key example comes from auditing in the social sciences, which have historically played a critical role in enforcing civil rights statutes. Social science audits have often employed participatory action research (PAR) methods, where researchers and community members work together to

develop and execute studies that accurately reflect the interests of impacted parties and aim to detect instances of discrimination.[19] One notable instance relates to studies on housing discrimination, where large-scale audits supported by the Department of Housing and Urban Development, in collaboration with researchers and community organizations, revealed discriminatory practices against Black prospective homebuyers and tenants. In today's landscape, housing approval processes often employ algorithmic systems, and similar participatory methods in AI assessment are needed to determine whether biases are being inadvertently perpetuated through technical means.

One participatory method that has shown promise in recent years is the adoption of crowdsourced AI audits, which often employ user-donated data that researchers analyze for instances of discrimination. This approach mirrors the principles of PAR methods in social science auditing, emphasizing the importance of community involvement and input for more meaningful evaluations.[20] By involving stakeholders from diverse backgrounds, enabling them to gain a deeper understanding of AI's impacts and effectiveness, and offer feedback accordingly, these methodologies are able to provide more robust findings.

3. Public Participation

Civil society organizations, academics, and policymakers have underscored the need for public participation, which "consists of measures that offer opportunities for people most likely to be affected by a given system to have influence into the system's design and deployment, including decision-making power."[21] Especially as AI systems are integrated in areas that impact civil and human rights, it is important that communities are meaningfully engaged with and empowered to shape whether and how these systems are integrated into their lives.[22]

Social science and humanities expertise and methods can help communities engage in the AI development and deployment process. Participatory methods in human-computer interaction (HCI) derive from ethnographic disciplines such as anthropology and sociology. Methods such as iterative testing, surveys, and focus groups can provide opportunities for participation in many forms and across the AI lifecycle, including in risk and harm identification, identifying design and measurement challenges, building trustworthiness methods, and establishing feedback loops with end-users and historically marginalized communities.[23]

Applied Sociotechnical Research: From Product Design to Real-World Impacts

While humanities and social science disciplines may seem separate from STEM and technology development, they have been essential contributors to breakthrough technologies well before AI. Anthropologists were famously part of the first cybernetics meetings in the 1940s, and the field of user experience (UX) research and design began with tech companies hiring ethnographers.[24] Since Xerox PARC's early, experimental practice of incorporating social scientists into the product development process in the late 1970s, ethnographers have been part of the tech workforce and contributed to R&D efforts.[25] In the 1980s, social science methods were often filtered through interaction design and human-centered design in attempts to make user interface research and design more socially responsible, and the term "user research" was used.[26] In the 1990s, as digital technologies became more widespread, more anthropologists entered tech companies.[27] Today, humanities and social science experts are employed at places like Microsoft Research, Apple, IBM, Intel, and the Institute for the Future; they are also working at newer companies like Google and Meta and in consultancies like Gemic, Stripe Partners, and IDEO. Their vital insights help technology companies anticipate design needs.

While humanities and social science experts may be involved in responsible AI, DEI (diversity, equity, and inclusion), sustainability, and ethics teams outside of or adjacent to product teams — corners of tech where social scientists have an outsized presence — they are often marginalized within corporations. Experts who work on AI ethics and responsible AI issues are often tasked with solving systemic problems without necessary support. Discussing the resulting burnout experienced by responsible AI practitioners, AI ethicist Emmanuel Goffi noted that the AI field is not "really open to the humanities," which leads developers to pursue a "quick technical fix" instead of "ethical thinking [that goes] deeper…and [applies] to how the whole organization functions."[28]

Relegating humanities and social science expertise to the periphery while pursuing a purely technical fix disregards the fact that AI systems require and benefit from a broader sociotechnical evaluation. Engaging sociotechnical expertise provides opportunities to create robust and iterative responses, and processes to inform and shape AI development, use, and oversight. Building off of capability evaluations (currently the dominant way to engage in safety evaluations), Google DeepMind researchers have proposed a sociotechnical safety evaluation of generative AI systems by incorporating two additional layers to safety evaluation: human interaction and systemic impacts.[29] These additional considerations are intended to better evaluate the wider context around AI's real-world performance — instead of relying solely on technical functions to assess safety.[30]

The Google DeepMind researchers' call for a sociotechnical approach to evaluating generative AI systems mirrors that of Genevieve Bell, an anthropologist who started Intel's first user experience group in 2005. Bell called for a more holistic approach to building and evaluating AI systems: "One of the first things that I realized is everyone is still thinking about AI like it is a computer science problem, and it is not. AI is going to move well beyond computers, and that will affect our lives."[31] The expertise exists — in academia, in companies, and elsewhere — but to reap the benefits, humanities and social sciences must be actively integrated into AI governance and risk management efforts and practices.

Recommendations for Building an AI Governance and Risk Management Ecosystem Equipped to Handle the Sociotechnical Nature of AI Systems

We offer the following recommendations for incorporating humanities and social science methods and expertise into government efforts. While we have focused our recommendations on the American federal government, they are relevant to all government efforts, including at the state and local level and internationally.

1. **Federal hiring efforts related to the National AI Talent Surge and the Office of Management and Budget's (OMB) memo on agency use of AI must invest in humanities and social science experts.**[32] Humanities and social science experts should be a focus of hiring efforts aimed at addressing the many issues with and potential opportunities of AI systems, as well as for roles facilitating meaningful public participation. As NIST notes in the AI RMF playbook, it is particularly important that "personnel with expertise in participatory practices" are involved in mediating between technical experts and the public to address AI risks.[33] Postings should include roles for experts with purely humanities and/or social science expertise who focus on sociotechnical research, instead of requiring applicants to also have technical expertise. Chief AI officers can also center the sociotechnical analysis of AI systems and prioritize humanities and social science methods and expertise in their agency work.

2. **During procurement processes, agencies should require vendors to incorporate sociotechnical research and evaluation methods.** Federal procurement is a key opportunity to ensure that vendors are engaging in "sound internal AI system assurances."[34] As this brief explores, a sociotechnical understanding and analysis of AI systems is crucial to determining their safety, efficacy, and impacts. By requiring vendors to demonstrate that they have engaged in a sociotechnical evaluation of their systems during a procurement opportunity, agencies will be better able to ensure that they are procuring well tested and aligned systems, and to enforce AI accountability practices throughout the industry.

3. **Humanities and social science methods and expertise should be included in the development of all guidelines and standards for AI assessment, research and development (R&D), and policy.** As this policy brief explains, understanding AI systems — from the harms they can proliferate to the opportunities they offer to innovate — cannot be fully realized without humanities and social science methods and expertise. It is essential that all efforts to advance the development of AI assessment tools, methods, and policies, from NIST's US AI Safety Institute to AI auditing to R&D programs, prioritize humanities and social science approaches alongside technical approaches.

Acknowledgements

The authors would like to thank the following individuals for reviewing this brief before publication: Miranda Bogen, Jenna Burell, Melissa Gregg, and Amy Winecoff. Thank you also to our Data & Society colleagues Jacob Metcalf and Briana Vecchione who offered valuable feedback and support.

Endnotes

1 Elham Tabassi, Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST Trustworthy and Responsible AI, National Institute of Standards and Technology, Gaithersburg, MD, [online], 2023, https://doi.org/10.6028/NIST.AI.100-1, https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=936225.

2 Ibid.

3 Ibid.

4 National Artificial Intelligence Advisory Committee, "National Artificial Intelligence Advisory Committee (NAIAC) Year 1 Report: Year 1," AI.Gov, May 2023.

5 Madeleine Clare Elish and Elizabeth Anne Watkins, *Repairing Innovation: A Study of Integrating AI in Clinical Care*, Data & Society Research Institute, 2020, https://datasociety.net/wp-content/uploads/2020/09/Repairing-Innovation-DataSociety-20200930-1.pdf.

6 Safiya Umoja Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: New York University Press. 2018.

7 Emanuel Moss, Elizabeth Watkins, Ranjit Singh, Madeleine Clare Elish, and Jacob Metcalf, *Assembling Accountability: Algorithmic Impact Assessment for the Public Interest*, June 29, 2021, http://dx.doi.org/10.2139/ssrn.3877437.

8 National Artificial Intelligence Advisory Committee, "National Artificial Intelligence Advisory Committee (NAIAC) Year 1 Report: Year 1," AI.Gov, May 2023.

9 Deirdre K. Mulligan and Helen Nissenbaum, 'The Concept of Handoff as a Model for Ethical Analysis and Design,' in Markus D. Dubber, Frank Pasquale, and Sunit Das (eds), *The Oxford Handbook of Ethics of AI* (2020; online edn, Oxford Academic, 9 July 2020), https://doi.org/10.1093/oxfordhb/9780190067397.013.15; Safiya Umoja Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: New York University Press. 2018; Karen Levy, *Data Driven: Truckers, Technology, and the New Workplace Surveillance*, Princeton: Princeton University Press, 2023.

10 In OMB's M-memo 5(c)(iv)(A)(1) agencies have to document "[t]he intended purpose for the AI and its expected benefit, supported by specific metrics or qualitative analysis. Metrics should be quantifiable measures of positive outcomes for the agency's mission — for example to reduce costs, wait time for customers, or risk to human life — that can be measured using performance measurement or program evaluation methods after the AI is deployed to demonstrate the value of using AI.34 Where quantification is not feasible, qualitative analysis should demonstrate an expected positive outcome, such as for improvements to customer experience, and it should demonstrate that AI is better suited to accomplish the relevant task as compared to alternative strategies." Office of Management and Budget, *Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence*, March 28, 2024, https://www.whitehouse.gov/wp-content/uploads/2024/03/M-24-10-Advancing-Governance-Innovation-and-Risk-Management-for-Agency-Use-of-Artificial-Intelligence.pdf.

11 "Join the National AI Talent Surge," AI.gov, Accessed April 24, 2024. https://ai.gov/apply/., https://www.whitehouse.gov/ostp/news-updates/2024/01/29/a-call-to-service-for-ai-talent-in-the-federal-government

12 Seth Lazar and Alondra Nelson, "AI Safety On Whose Terms?" *Science* 381,138-138, 2023, DOI:10.1126/science.adi8982.

13 Shalaleh Rismani, Renee Shelby, Andrew Smart, Edgar Jatho, Joshua Kroll, AJung Moon, and Negar Rostamzadeh, "From Plane Crashes to Algorithmic Harm: Applicability of Safety Engineering Frameworks for Responsible ML," April 19, 2023, https://doi.org/10.1145/3544548.3581407.

14 Inioluwa Deborah Raji and Roel Dobbe, "Concrete Problems in AI Safety, Revisited" arXiv preprint, Dec 18 2023, arXiv:2401.10899.

15 Ibid.

16 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," arXiv preprint, July 12 2022, arXiv:2108.07258.

17 Laura Weidinger, Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, Iason Gabriel, Verena Rieser, and William Isaac, "Sociotechnical Safety Evaluation of Generative AI Systems" arXiv, Oct 21 2023, http://arxiv.org/abs/2310.11986.

18 Serena Oduro, Ranjit Singh, and Jacob Metcalf,"Response to the National Telecommunications and Information Administration's Request for Comment on AI Accountability Policy," Data & Society Research Institute, June 12, 2023, https://datasociety.net/wp-content/uploads/2023/06/NTIA_Comment_June_2023R1.pdf.

19 Briana Vecchione, Karen Levy, and Solon Barocas, "Algorithmic Auditing and Social Justice: Lessons from the History of Audit Studies," In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pp. 1-9. 2021.

20 Rumman Chowdhury and Jutta Williams, "Introducing Twitter's First Algorithmic Bias Bounty Challenge," July 30, 2021. https://blog.twitter.com/engineering/en_us/topics/insights/2021/algorithmic-bias-bounty-challenge; Mitchell Clark, "After Accusations, Twitter Will Pay Hackers to Find Biases in Its Automatic Image Crops," *The Verge*, July 30, 2021. https://www.theverge.com/2021/7/30/22602553/twitter-image-cropping-algorithm-bias-competition-bounty-def-con.

21 Michele Gilman, *Democratizing AI: Principles for Meaningful Public Participation*. Data & Society. 2023. https://datasociety.net/library/democratizing-ai-principles-for-meaningful-public-participation/.

22 The White House, "Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People,"2022. https://www.whitehouse.gov/ostp/ai-bill-of-rights/.

23 Serena Oduro, Jacob Metcalf, Ranjit Singh, Briana Vecchione, and Meg Young, "Response to RFI Related to NIST's Assignments Under the AI Executive Order," Data & Society Research Institute, 2024, https://datasociety.net/wp-content/uploads/2023/06/NTIA_Comment_June_2023R1.pdf.

24 Claus Pias, *Cybernetics: The Macy Conferences* 1946-1953. Chicago: University of Chicago Press, 2016; Ronald Kline, "How Disunity Matters to the History of Cybernetics in the Human Sciences in the United States," 1940–80 History of the Human Sciences, 33(1), 12-35, 2020, https://doi.org/10.1177/0952695119872111; Lucy Suchman, "Consuming Anthropology," In *Interdisciplinarity: Reconfigurations of the Social and Natural Sciences*, edited by A. Barry and G. Born, 141–60. Routledge, 2013; Melissa Cefkin (Ed.), *Ethnography and the Corporate Encounter: Reflections on Research in and of Corporations* (Vol. 5), 2010. Berghahn Books; Mariette L. Baba, De-Anthropologizing Ethnography: A Historical Perspective on the Commodification of Ethnography as a Business Service, In *Handbook of Anthropology in Business*, pp. 43-68. 2014. Walnut Creek: Left Coast Press.

25 Lucy Suchman, *Human-Machine Reconfigurations: Plans and Situated Actions*. 2nd ed. of *Learning in Doing: Social, Cognitive and Computational Perspectives*. Cambridge: Cambridge University Press, 2006.

26 Ben Shneiderman, "The Future of Interactive Systems and the Emergence of Direct Manipulation." *Behaviour & Information Technology* 1, no. 3 (1982): 237–56. doi:10.1080/01449298208914450; Norman, Donald. *The Design of Everyday Things*. New York: Basic Books, 1988.

27 Melissa Cefkin (Ed.), Ethnography and the Corporate Encounter: Reflections on Research in and of Corporations (Vol. 5), 2010. Berghahn Books

28 Melissa Heikkilä, "Responsible AI Has a Burnout Problem," *MIT Technology Review*, November 11, 2022, https://www.technologyreview.com/2022/10/28/1062332/responsible-ai-has-a-burnout-problem.

29 Laura Weidinger, Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Juan Mateos-Garcia, Stevie Bergman et al. "Sociotechnical Safety Evaluation of Generative AI Systems." *arXiv preprint arXiv*:2310.11986, 2023.

30 Ibid.

31 Genevieve Bell and Jim Euchner "Creating a New Engineering Discipline for the Age of AI, Research-Technology Management," 65:2, 11–17, 2022, DOI: 10.1080/08956308.2022.2021715

32 This recommendation mirrors the NTIA's recommendation in their report on AI accountability: "We recommend an investment in federal personnel with appropriate sociotechnical expertise to conduct and review AI evaluations and other AI accountability inputs." The National AI Talent Surge, established in President Biden's "Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," aims to quickly expand the number of AI experts in the federal government. OMB's M-memo "Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence" also encourages agencies to recruit AI talent.

33 US Department of Commerce and NIST, "Risk Management Playbook (Govern 5.1)," accessed April 24, 2024, https://airc.nist.gov/AI_RMF_Knowledge_Base/Playbook/Govern#Govern%205.1.

34 National Telecommunications and Information Administration, "AI Accountability Policy Report," March 27, 2024, https://www.ntia.gov/issues/artificial-intelligence/ai-accountability-policy-report

Data & Society is an independent nonprofit research institute that advances new frames for understanding the implications of data-centric and automated technology. We conduct research and build the field of actors to ensure that knowledge guides debate, decision-making, and technical choices.

www.datasociety.net

@datasociety

Layout by Hector Sandoval

May 2024