CENTER FOR DEMOCRACY & TECHNOLOGY

Applying Sociotechnical Approaches to AI Governance in Practice

May 2024

Authored by
Miranda Bogen, *Director, CDT AI Governance Lab*
Amy Winecoff, *AI Governance Fellow*

Organizations of all kinds build, integrate, and deploy artificial intelligence (AI)-powered systems. Ensuring AI systems live up to their promise without causing undue harm requires recognizing and engaging with the broader contexts in which those systems are deployed. Taking a **sociotechnical approach** to building and governing AI systems that protect people's rights and safety is increasingly expected by policymakers and public stakeholders and requires thoughtful design of governance, safety, and technical development methods.

Social science expertise has long proven instrumental in how companies develop safe, well-functioning products in other technical domains. From the earliest days of the technology industry, sociotechnical experts — people with training and expertise in sociology, anthropology, political science, law, economics, psychology, public health, geography, social work, history, and other such fields — have been crucial to shaping the design of technology to be responsive to human needs and behavior.

Importantly, sociotechnical harms are not distinct from safety considerations. Experts in social science and the humanities can provide important perspective on effective governance of social systems; more effectively involve leadership, internal product teams, users, and affected communities in decision-making; and encourage the adoption of evaluation and measurement methods that incorporate a deeper understanding of human behavior and the social consequences of deploying technical systems in particular contexts. Failing to fully incorporate this perspective means that at minimum, the promise of technologies is unlikely to be realized — or worse, that they will lead to significant harm.

The importance of sociotechnical considerations in the domain of AI systems is often invoked, but rarely explained in a way that feels actionable for practitioners. To help make this important lens more tractable, this brief guide walks through a discussion of what constitutes a sociotechnical approach and offers wide-ranging examples of how such methods can be leveraged within existing AI design, development, and deployment processes. Drawing from these examples, we provide ten actionable recommendations for how AI teams and organizations can better integrate expertise at the intersection of technology and societal dynamics into their design, development, deployment, and governance.

What are Sociotechnical Approaches?

AI systems are not just technical artifacts — they are embedded in social structures, organizations, and societies. Applying a *sociotechnical lens* to AI governance means understanding how AI-powered systems might interact with one another, with people, with other processes, and within their context of deployment in unexpected ways. *Sociotechnical approaches* consider the human and institutional dimensions that affect how AI is used and the impact it will have. These approaches add a wealth of insight to teams developing AI-powered systems, helping technologists understand how users interact with products, how technologies affect social groups and economies, and how the impacts of technology can emerge over time as AI systems and people co-evolve.

By incorporating the input and perspective of experts who can bring this more fulsome, context-aware lens to the development and deployment of AI systems, AI-powered applications are more likely to be suited for their intended context of deployment and tuned to the needs of the intended community of users. In the case of more general purpose AI systems, the organizations creating and using these technologies will be better equipped to foresee unanticipated opportunities and issues, and spot evolving dynamics around how those tools are being incorporated into existing contexts — which will be critical to informing AI governance.

Sociotechnical approaches draw on a varied toolkit of research methodologies, including qualitative interview analysis, ethnographic research, and other qualitative techniques that can complement the quantitative methods familiar to most data scientists and AI engineers. In this way, sociotechnical experts are well-positioned to act as a bridge between the AI development teams and the communities of people disproportionately affected by AI systems, translating community insights into actionable plans.

In the context of developing and deploying AI systems, sociotechnical approaches include:

Ideation and Design

- **Co-designing technical solutions with the communities who stand to benefit from them.** For example, MIT researchers worked with youth from four countries to collaboratively imagine inclusive robotic agents.
- **Engaging meaningfully with anticipated system users, domain experts, and communities likely to be directly and indirectly impacted by an AI system** to inform its design from the bottom up.
- **Using qualitative ethnographic approaches to identify user and community needs** and consider how an AI-powered system would compare to non-AI interventions in addressing those needs.
- **Exposing AI system developers directly to the contexts where the technology they are building will be deployed** to help them tangibly understand the implications of technical choices involved in system design. For instance, a group of researchers and practitioners brought an individual who had been incarcerated due to a pre-trial detention algorithm to speak at a popular AI conference about the realities of pre-trial detention and the criminal legal system.
- **Analyzing the legal, social, and geopolitical context in which an AI-powered system may be deployed** to ensure it is designed to maximally respect relevant laws, norms, and human rights. In this vein, a large group of researchers at Google DeepMind and other organizations conducted an extensive exploration of societal risks of AI assistants — presumably similar to ones the company is considering building.
- **Mapping an organization's existing teams and processes to understand how the introduction of an AI system may lead to surprising failure modes.** For example, if an organization puts one team in charge of building foundation models, a different team to integrate them into products, and an interdisciplinary committee to review AI use cases, the organization will need to make sure appropriate information is being shared among the different teams to inform what safeguards ought to be integrated into the system given the context in which it will be deployed.

Building and Implementation

- **Interviewing individuals involved in the AI system development and deployment process to spot where assumptions about the system, its components, and its safeguards may be faulty** and identify opportunities to address these limitations before they lead to negative outcomes. As an example, a team developing an AI-powered sepsis detection model mapped how the technology would be integrated into clinical care and learned that nurses were critical to effective implementation of the system.
- **Considering how societal dynamics may have impacted the generation of data used for model training to inform data filtering and pre-processing efforts.** If those engaged in data labeling are not fairly compensated or supported, that could have meaningful effects on the accuracy and safety of the resulting models.

- **Qualitatively reviewing data labeling guidelines and reinforcement learning from human feedback (RLHF) processes to ensure they reflect appropriate nuance and account for regional and cultural differences and are responsive to usability challenges.** For instance, if labellers asked to identify household objects in photos are only given the option to select objects that commonly appear in high-income households, the resulting AI models trained on that data likely won't work for much of the world. If the population of people providing human feedback to a large-language model (LLM) only represents certain perspectives, the resulting model might be steered toward a narrow worldview that leads to downstream harms.
- **Ensuring that a system's target metrics, relevant evaluations, and guardrails have holistically and appropriately operationalized complex concepts** such as "quality," "safety," and "fairness."
- **Foreseeing how those reviewing or acting on the outputs of AI systems will interpret recommendations, and understanding what sort of human oversight and training will most effectively prevent automation bias or other harms.** For example, researchers found that being presented with results from AI risk assessment tools changed how people might make consequential decisions about granting government loans and setting bail terms.
- **Engaging with impacted communities throughout the development process to integrate ongoing input into implementation and evaluation decisions** that might not have been evident during the design stage, and translating between community members and technical decision-makers.

Deployment and Integration

- **Modeling how users, organizations, and societies might adapt to the introduction of new tools** in order to anticipate harmful design flaws and feedback loops and externalities, including by incorporating human factors analysis to minimize the risk of adverse events.
- **Building out methods to monitor not only immediate successes and acute harms of a system but also longer-term signals of benefits and indirect or cumulative harms**; for example, measuring not just whether an AI-powered hiring system rejects different applicants at disproportionate rates, but monitoring how the distribution of quality job opportunities changes over time. Such methods should be informed by a deep understanding of the context into which the system will be integrated.
- **Assessing how organizational processes and incentives may support or undermine efforts to conduct sound risk management of AI system development and deployment.** For instance, researchers exploring responsible AI efforts in industry identified that practitioners often struggle to secure attention to long-term outcomes when product teams work on short-term development timelines. To build effective governance processes, AI organizations must invest in necessary infrastructure, staff, and leadership involvement to ensure that new processes are equipped to effectively spot and remediate issues.

Recommendations for AI Developers and Deployers

- **Integrate team members with sociotechnical expertise into product teams, AI safety teams, and responsible AI efforts.** The utility of these perspectives is higher the earlier they are incorporated, so make sure team members with social science and humanities skills are included in roadmapping, product ideation, and initiatives to define product scope and details. To maximize the value of such interdisciplinary teams, create opportunities for team members to proactively articulate and illustrate their capacity to contribute to the team's shared goals, and reward teams that take a holistic and inclusive approach to product development and risk management. When making product decisions, ensure that sociotechnical insights are elevated alongside technical considerations — and be prepared to change design and deployment plans based on sociotechnical input (up to and including not deploying a system).

- **Incorporate user experience and social science expertise in the design of algorithms and AI system components, not just user interfaces.** Psychologists, user experience designers, and user experience researchers have deep knowledge of people's experiences and behavior as they interact with technology, and are able to understand not only how users respond to interfaces but also underlying system components. When user experience professionals and psychologists design AI models and other system components to be responsive to user needs and constraints, AI systems will more likely address users' needs, and organizations will be better able to anticipate when modes of user interaction could inadvertently result in harm.

- **Empower user researchers and designers to engage in strategic areas of inquiry beyond product usability.** The impact of AI-powered products extends far beyond the points at which users directly interact with these tools, and in-house researchers can play an important role in spotting and addressing gaps and assumptions in product development — including understanding how dynamics within the product development pipeline might lead to downstream impacts. UX professionals can also help organizations build tools to help better surface failure modes.

- **Involve team members with sociotechnical expertise in discussions and decisions around product metrics and guardrails**. The choice of how to measure success and define guardrails of an AI system will profoundly shape decisionmaking about whether a product is ready to launch, the suitability of that tool in different contexts, and efforts to mitigate risks and harms. Holistically crafting key success metrics and guardrail metrics (sometimes called countermetrics) — including understanding what they do and don't capture about people's experiences with the technology — is critical to supporting sound decisions and risk management.

- **Lean on sociotechnical experts to apply mixed-methods approaches to grasp the implications of building and deploying AI-powered systems.** Quantitative metrics play a role in AI systems, but are insufficient on their own to understand the full impacts of AI.

Whereas quantitative approaches allow practitioners to answer very specific questions about data or a model (for example, whether a AI-powered diagnostic system functions effectively across a set of predefined populations), they typically do not provide insight into questions the practitioner did not ask (such as whether the relationship between doctors and patients would influence the adoption of the tool). And while qualitative methods rarely provide definitive answers to specific questions, they are often well-suited to helping practitioners explore the range of possible considerations relevant to a system. If practitioners focus only on narrowly-defined system properties like accuracy metrics, they could end up inadvertently deploying systems that are not sufficiently effective, or even pose safety risks. As the National Artificial Intelligence Advisory Committee recommends, practitioners ought to lean on interviews and ethnographic studies, analysis of logged data, case studies, sociological audits, and historical analysis to assess the ramifications of a particular technology or set of technologies. Importantly, social scientists can use a variety of methods to help evaluate AI systems within the intended context of deployment, which is necessary to truly foresee their impact.

- **Allocate sufficient resources and time for fulsome qualitative investigation and analysis. Include clear descriptions of the sociotechnical methods employed in the documentation of risk assessments, and articulate how these findings were incorporated into system design, impact assessment, and risk mitigation.** AI development is moving quickly, and it can be tempting to rely on automated measurements and evaluations to gauge quality and detect risks, but these complex systems are being integrated into every facet of people's lives and so deserve much deeper analysis. Organizations must allocate enough capacity and give researchers sufficient time, budget, and flexibility to collect and generate insights and influence decision-making about risks and benefits. In depth analysis is especially important for AI systems that are more likely to impact people's rights and safety — which sociotechnical experts play an important role in identifying. Robust involvement by sociotechnical stakeholders in the development and deployment process in these contexts can support an organization's justification of their risk management decisions. Insufficient involvement ought to be considered a negative signal about an organization's commitment to responsible development.

- **Mobilize contextual and sociotechnical experts to develop constructive and inclusive opportunities for public input and co-design.** The development of AI-powered systems should be deeply informed by users, stakeholders, and impacted communities, but these processes can backfire without thoughtful design and facilitation, leading participants to feel organizations have taken advantage of their input without truly incorporating their feedback or respecting their time. Team members with social science expertise can help ensure appropriate stakeholders are included, spot and reduce barriers to meaningful participation (e.g., digital divide, childcare needs, and language and accessibility barriers), recognize and address power dynamics that can undermine the goals of engagement, and make sure information is presented in a clear and accessible manner. Such experts are also critical to help translate between internal teams and external stakeholders, to maximize mutual understanding, and make insights and suggestions as actionable as possible.

- **Enable team members with social science expertise to share lessons learned, both within and across organizations, and ensure leaders recognize their contribution.** Particularly in organizations oriented around technology functions, team members with social science expertise can find themselves siloed and their contributions undervalued. To fully benefit from sociotechnical insights, organizations should facilitate insight sharing across products and teams, and share findings and resulting decisions publicly to the greatest extent possible in order to inform and integrate external expertise. Even if sociotechnical approaches lead to significant roadmap changes, organizational leaders should reward team members whenever their expertise helps organizations make more robust and holistic decisions.

- **Recognize that sociotechnical expertise can be both a source of constructive input to identify beneficial applications of technology as well as an important source of critique when assumptions are faulty.** Social science researchers may be accustomed to identifying gaps in analysis and issue-spotting, but they are also critical partners in recognizing opportunities where thoughtful technical interventions can help support people and communities. In particular, empowering sociotechnical experts to facilitate the co-design of AI-powered tools together with the communities those tools stand to serve or impact can support overall value creation, while minimizing risks and harms. At the same time, organizations should value and act on insights that might question fundamental assumptions that have shaped a particular tool or system.

- **Engage contextual experts in monitoring the impact of AI systems that have been deployed** to surface surprising patterns, unanticipated impacts or interaction effects, and harms that were underappreciated or prove to be more serious than anticipated in order to update risk management and governance frameworks.

Conclusion

AI is deeply intertwined with social systems, organizations, institutions, and culture. Sociotechnical approaches to AI system development and deployment are important to contend with the socially-embedded nature of AI to ensure that these systems are safe and effective and that their risks have been appropriately managed. People with expertise in sociology, anthropology, political science, law, economics, and psychology already exist in a wide range of technical and non-technical roles in AI companies but tend to be underused in AI system development efforts. Instead, they are often relegated to siloed roles in AI ethics or governance, compliance, or pre-deployment user interface testing where they have limited input to early design and prototyping, with limited authority to substantively modify product roadmaps.

By following these recommendations, companies can more meaningfully engage existing sociotechnical expertise throughout the design, development, evaluation, and deployment process, and deepen their capacity to integrate sociotechnical considerations into AI governance efforts more broadly. Embracing these approaches will help practitioners both

produce less harmful technologies and realize more benefits of their systems for the social good. And as policymakers, regulators, and the public expect those developing AI systems to foresee the impact of the technologies they are building, deeper integration of these sorts of sociotechnical approaches into the core efforts of AI development must become the default. Only through holistic and context-sensitive efforts can practitioners effectively protect people's safety and rights in the face of all-too-rapid deployment of AI-powered technologies.

Find more from CDT's AI Governance Lab at *cdt.org*

*The **Center for Democracy & Technology** (CDT) is the leading nonpartisan, nonprofit organization fighting to advance civil rights and civil liberties in the digital age. We shape technology policy, governance, and design with a focus on equity and democratic values. Established in 1994, CDT has been a trusted advocate for digital rights since the earliest days of the internet. The organization is headquartered in Washington, D.C. and has a Europe Office in Brussels, Belgium.*