Stanford University
Human-Centered
Artificial Intelligence

Stanford | McCoy Family Center for
Ethics in Society

Report
May 2024

Workshop on Sociotechnical AI Safety

César Valenzuela, Jacqueline Harding

This workshop was co-hosted by the Stanford Institute for Human-Centered AI (HAI), the Stanford McCoy Family Center for Ethics in Society, and the MINT Lab at the Australian National University.

HOW CAN WE MAKE SURE THAT AI SAFETY DELIVERS ON ITS PROMISES to reduce present and future harms from advanced AI systems? On November 16 and 17, 2024, the Workshop on Sociotechnical AI Safety at Stanford (co-hosted by Stanford's McCoy Family Center for Ethics in Society, Stanford HAI, and the MINT Lab at Australian National University) aimed to make progress on this question, bringing together a diverse group of participants from industry and academia. The workshop put AI safety researchers in conversation with researchers whose work focuses on fairness, accountability, transparency, and ethics (FATE) in AI.

In his opening remarks, co-organizer Seth Lazar (ANU) emphasized the value of the sociotechnical approach to the assessment and mitigation of risks related to AI systems. For Lazar, the best hope for setting normative goals for AI, like "safety," is to integrate deep technical work with an equally robust understanding of how technology interacts with incentive structures and power relations in our societies. Sociotechnical approaches to safety afford precisely this understanding.

The presentations and discussions orbited around three main topics: (1) social inclusion; (2) the nuances and complexities of the conceptual landscape of the field; and (3) looking forward, identifying steps for the field to diversify and grow intellectually and, ultimately, develop better tools to identify and mitigate AI-related risks.

(1) Inclusion

A common theme in much of the discussion was inclusion. Shazeda Ahmed (UCLA) began the workshop by characterizing the epistemic community of AI safety, highlighting its close ties with effective altruist, longtermist, and rationalist movements. The ideological overlaps within the AI safety community, Ahmed argued, have allowed for effective field-building and dissemination of information. It has largely become an "epistemic culture," a community with its "own terminology, source texts, and knowledge claims."

Ahmed's project identified four ways in which the AI safety epistemic community maintains itself. First, there is online community building (via forums like LessWrong and its sister website, the AI Alignment Forum), including online career advising (through organizations like 80,000 Hours). These websites serve to record the community's collective knowledge at impressive speed; in particular, the Alignment Forum limits contributions to experts (as judged by moderators), functioning as a surrogate to a journal for AI safety research.

Second, Ahmed argued that AI forecasting plays an important role in the AI safety community (as exemplified by platforms like Metaculus). After all, the discipline of AI safety is largely predicated on the idea that significantly more capable AI is not only possible, but (increasingly) likely to emerge in the coming decades. So AI forecasting does not only play a social function; it also serves to motivate AI safety research and to recruit researchers — especially when the pace of AI's progress outstrips forecasts.

Third, and most importantly, the AI safety community produces AI safety research. Much of this research is produced by members of AI labs (such as Google DeepMind, OpenAI, or Anthropic) or research staff at nonprofits funded by the effective altruist community (such as Redwood Research or the Center for AI Safety), increasingly in collaboration with academia. This includes both conceptual and technical research; much of the technical research is published in computer science conferences and discussed by the wider AI community. Ahmed also presented work exploring what it is like to do alignment research from the perspective of researchers within the community, highlighting the importance of field-building contributions and the lack of consensus over measuring progress.

Fourth, high-value prize competitions serve to motivate engagement with AI safety research topics within the AI community at large. As Ahmed noted, prize competitions arguably play a less important role in maintaining the AI safety community than the other three factors, but they are nevertheless a distinctive feature of the AI safety community.

The audience discussion following Ahmed's talk drew attention to dynamics of the epistemic community that would be worth exploring further. For instance, although the AI safety community is often associated with a concern about long-term catastrophes, the accelerated pace in the development of AI capabilities has led people to focus on short-term risks that could emerge in the next five to 10 years. As Seth Lazar noted, the recent Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence adopted an existential risk approach that could help solve issues of factions within the community.

The audience was also interested in the intersection between the AI safety and FATE epistemic communities, while acknowledging that neither of these communities is a monolith and involves different methods, motivations, and concerns. Even among AI safety researchers, for example, there might be differences in how key concepts and tools are understood, or in which research areas deserve greater attention. In discussion, Ahmed also noted that the AI safety community's efforts on community building could provide a useful model for the FATE community to increase engagement.

As intellectual communities go, the AI safety community is remarkably well-organized, no doubt owing in part to overlaps in the ethical commitments and educational backgrounds of its members. There is a risk, though, that the (relative) homogeneity of the community excludes non "in-group" voices from conversations about AI safety. This point was made by Dylan Hadfield-Mennell (MIT), who argued, together with Simon Zhuang, that the very same formal arguments for taking AI safety seriously underline the necessity of broadening participation in AI's development and deployment.

We can reconstruct the argument as follows. Take some agent, who wants to optimize a utility function defined on some set of independent features. In practice, though, she can only optimize a proxy for her true utility function, defined on some proper subset of the features she is interested in (say, because it is difficult to measure some of the objectives). Under mild assumptions, Zhuang and Hadfield-Mennell show that letting an optimization process run using the proxy utility function will eventually result in states that are arbitrarily bad for the agent (from the perspective of her true utility function). Informally, the optimization process ends up "extracting" value from the features which are not represented by the proxy function in order to maximize value for those that are.

Hadfield-Mennell drew attention to a modified interpretation of the argument that leaves the formal result untouched: Rather than a single agent optimizing over some set of features that matter to her, suppose we have a set of agents, each of whom is interested in some independent feature(s). Then the very same argument implies that when we exclude any one of these agents from consideration by an optimization process, such as a recommender system (in the sense of ignoring the feature they care about when training the system), we will eventually end up in a situation which is actively worse for the excluded agents (that is, the point is not merely that they don't get the benefits of the optimization).

In other words, a foundational argument within technical AI safety provides an argument for broadening the AI safety community to include underrepresented viewpoints. Of course, as some audience members observed in discussion, there's an open question about how well the formal details of the argument apply to real systems. In particular, if one relaxes the assumption that features are independent, there might be ways to avoid the conclusion of the argument.

Inclusion strategies cannot remain at a local or even national level, though. In her talk "Safety and Geopolitics: A Critical Security Studies Lens," Marie-Therese Png (Oxford) employed ideas from critical security studies to argue that only a fully *global* approach to participation can make AI safer, drawing attention to the international supply chains that underlie much progress on the hardware side of AI's development. Png stressed the fact that the concentration of corporate and political power in a handful of entities enables the marginalization of stakeholders in the Global South, who are often left out in the assessment of AI systems. The concentration of power allows AI developers to rely on the extraction of cheap labor precisely from the Global South and obtain rents that are disproportionate from their activities. In that sense, Png emphasized the need for AI safety frameworks to go beyond technical aspects and account for relations of exploitation and extraction. More broadly, Png claimed that the "future of multilateralism is multistakeholderism," where not only the already powerful parties make decisions but all the relevant stakeholders are included in decision-making processes. Co-design and co-governance are crucial here, as what is even a risk is defined through the lens of power. Co-governance faces significant barriers, though, given existent dependency dynamics and the fact that greater inclusion may still coexist with persistent structural harms, which would undermine the former.

In the discussion, Png stressed the need to give affected communities the opportunities to express *their own* goals with respect to AI systems and drop the assumption that people's relation to and demand for AI is universal and homogeneous. Considerations about *qualification* were also brought into the debate: Who should be contributing to discussions on safety and security? For instance, "current user of the relevant technology" might not be a reliable criterion of qualification, as some communities might have to be consulted even if they do not have access to the technology yet. As Png recognized, it is unclear which stakeholders can be engaged in this debate from the many polities that are currently disempowered. This leads to related questions: How can we challenge qualification? How to ensure conversations do not go on for too long, delaying justice? Given structural barriers, how do we make productive engagement with communities?

What could inclusion in AI safety look like? Model fine-tuning, such as Reinforcement Learning from Human Feedback (RLHF), provides one potential point of intervention. RLHF involves three steps. First, researchers gather data about human preferences over the model's outputs. Second, this preference data is used to train a reward model; the reward model attempts to predict, for each new model output, the degree to which humans would approve of the output. Finally, this reward model is used to fine-tune the original model using reinforcement learning (specifically, a policy gradient method).

As will be clear, the effects of RLHF on model behavior depend in large part on the choices made during the preference data collection process (often called "value elicitation"). How are model outputs selected for value elicitation? Who are the people whose values are elicited? What factors are they prompted to consider in ranking model outputs (helpfulness, harmlessness, etc.)? If the preference data gathered is not representative of the population at large, then RLHF could — rather than ameliorating issues from pretraining — end up making models less useful for underrepresented groups, potentially even compounding biases acquired from pre-training data.

Engaging with this worry, Nahema Marchal and Iason Gabriel (Google DeepMind) presented STELA (Sociotechnical Language Agent Alignment), a framework for inclusive value elicitation. STELA breaks the value elicitation process down into four steps: (1) sample generation (decide which outputs participants should express preferences over; (2) norm elicitation (put participants in focus groups and have them express and deliberate their views on the model outputs); (3) ruleset development (distill the focus group discussion into a set of principles for the model to follow); and (4) ruleset review (have relevant experts, such as human rights experts, verify the ruleset developed).

The main feature that distinguishes STELA from prior work on value elicitation is its use of deliberative, in-person focus groups (rather than remote, independent crowdworkers). In particular, this means that it has greater control over participant selection, allowing for the possibility of greater inclusion. In the study Marchal presented, researchers selected participants from four demographic groups marginalized within the United States (women, African Americans, Southeast Asian-Americans and Latinas/os/xs). They compared the ruleset generated from these participants' discussions (a "community" ruleset) with the rulesets used in prior work on value elicitation (”developer rulesets”). Not only did they find that the community ruleset differed in important ways from the developer ruleset, but also that participants reported that the process helped them feel empowered.

In discussion, the audience drew attention to issues of democratic legitimacy, questioning the extent to which the study should be taken to ground the legitimacy of value elicitation, given that it was carried out by a leading AI developer (and that the process of ruleset development was still carried out by researchers, rather than community members). In particular, the audience wondered whether advocacy groups carried a particular form of legitimacy that justified or required their presence. The speakers acknowledged concerns about whether private companies should embrace this role of leading participatory processes or democratic projects at all; in particular, they emphasized the value of the situated knowledge that advocacy groups may bring to bear, while also highlighting its limitations. Likewise, the role of consent in legitimizing these and broader practices was also briefly discussed. Finally, the choice of categories was questioned, given that different marginalized groups and different majorities may have conflicting views about the categories chosen (e.g., should "white working class" be included?).

In a similar vein, Deep Ganguli (Anthropic) presented work with the Collective Intelligence Project to broaden inclusion during fine-tuning using Constitutional AI (CAI), a technique closely related to RLHF.

Rather than eliciting values from human participants, CAI involves constructing a list of principles (a ”constitution”). A language model then uses these principles to rank model outputs (in a similar way to how the human participants might be prompted to rank model outputs along one dimension or another); a reward model is trained from these rankings and used to tune the original model. Just as RLHF's effects will depend on the preference data used to train the reward model, CAI's effects will depend on the constitution used to generate the data for the reward model. This introduces representational issues; constitutions have previously been written by researchers at Anthropic, without input from the wider population.

To ameliorate this, the project Ganguli presented ("Collective Constitutional AI") canvassed a representative sample of the U.S. population (n=1000) to generate a new constitution. Not only could participants vote on existing rules, they could also propose novel rules themselves, which other participants could then vote on. The "public" constitution was generated by preserving those rules for which there existed sufficient consensus. Interestingly, similar to the project discussed above, researchers found significant differences between the public constitution and the developer constitution. To test whether these differences would affect downstream model behavior, they tuned a model using each of the two constitutions; they found that model performance was identical, but the "public" trained model exhibited less bias (as measured by the BBQ benchmark), although the differences were relatively small.

In discussion, concerns were raised about further applications and inclusion. For instance, how could we extend the public models to other languages, or other (non-U.S.) populations? The speaker noted that current models are trained predominantly on English-language data, which raises the question of how alignment interventions can universalize.

One issue raised was about the role of disagreement in aggregation; if we only accept rules for which there is sufficient consensus, do we risk prioritizing majority viewpoints? Similarly, what should we do in the case of disagreement? Should we generate multiple constitutions and tune different models, or attempt to aggregate conflicting sets of preferences in a single constitution? One audience member put this worry in terms of the model's "liberal bias," which might be reinforced by having consensus as a desideratum. The speaker welcomed this criticism.

A further important idea is how to replicate the real-world political engagement of citizens with their constitution (through litigation, civics, disobedience, etc.), which brought up the need for feedback mechanisms during and after the tuning process.

Both of the previous projects involved safety interventions during fine-tuning. Nathan Lambert (AI2), though, presented work complicating the effectiveness of these interventions, focusing on RLHF. RLHF makes several in principle assumptions about the representation and aggregation of preferences, and also in practice assumptions (which Lambert dubbed "presumptions") about the process of tuning models using RL. To name just a few: RLHF assumes that human preferences are context-independent and stable across time, that the preferences of different users are well-modeled through cardinal aggregation, and that maximizing reward will lead to better downstream model behavior. Many of these assumptions go unchallenged by researchers; as Lambert pointed out, the proprietary nature of most reward models inhibits their investigation. So a lack of transparency from model developers undermines scrutiny of RLHF.

Discussion focused on more philosophical questions (what is meant by a "preference" as operationalized by RLHF?) as well as more practical ones (are there techniques for aggregating preferences along different dimensions?). There was widespread agreement that RLHF has limitations as a technique for aligning model behavior with human interests; many participants felt that it intervened in the model development process too late.

(2) Complicating the Conceptual Landscape

A key term in discussions about AI safety is “alignment.” It usually refers to the compatibility between human interests and the functioning of AI technology, which in some scenarios could pursue its own interests at the expense of ours. However, as various speakers stressed, there is no consensus on the definition of the term itself or on the right path toward alignment. In her talk “Integrating Transdisciplinary Insights Towards a Transgranular Entity Alignment Framework,” Shiri Dori-Hacohen (UConn) argued that current research on alignment is often carried out from a techno-deterministic and reductive perspective that overlooks two important facts. First, problems of misalignment also emerge from non-AI related aspects of technological systems, especially in the case of social media platforms where AI isn’t involved or plays a minor role (e.g., WhatsApp). Second, research on alignment wrongly presupposes that there is a *single* set of human desires and needs to which AI can be said to be “misaligned.” Instead, that set of desires and needs is politically and philosophically contested. Further, issues of alignment emerge in every aspect of our life and planet and cannot be restricted to a single sphere or domain.

In response, Dori-Hacohen drew on systems engineering, biology, and social sciences to offer a “transgranular entity alignment framework” that considers inter-entity, intra-entity, and cross-granularity interactions to assess alignment relationships between any two given entities. The framework, which offers an alignment score from 0 to 1 (0 being fully misaligned; 1 being fully aligned), hypothesizes that alignment modeling can be done between entities at many granularity levels: from macromolecular entities to the biosphere, including social systems. In the discussion, Dori-Hacohen clarified that the framework is fully *descriptive*—that is, it does not assume that either alignment or misalignment is good or bad and acknowledges the latter as a pervasive feature of the world.

What other frameworks can we use to think about alignment? In his paper “Toward Normative Alignment of AI Systems,” Mark Riedl (Georgia Tech) proposed the notion of “normative alignment” (in contrast to *value* alignment): alignment understood as conformity of AI systems to the norms of our communities. According to Riedl, the more common idea of *value* alignment runs into the difficulties of encoding the values at stake, which is impossible to do at a level of enough specificity such that no agent could circumvent. The focus on community norms allows for the possibility of AI behaving acceptably across different social contexts, as there would not be a *single* set of standards to which AI would be designed to conform—thus, there would be no need to encode rules. This is closer to how people behave in social life, adapting their behavior across a variety of situations. In that sense, one way to train large language models toward normative alignment is the use of *stories*, as they usually are powerful demonstrations of norms, encode social and cultural norms, and offer models of idealized behavior.

As the audience pointed out, several questions remain. Just like we need to ask who determines the relevant values when we talk about value alignment, the normative alignment approach would raise the same problem: *Who* determines these norms? Riedl specified that the relevant norms are those of the group that interacts with the AI system, but this still leaves open the question about which subgroup gets to determine such norms. Relatedly, how do we make sure we incorporate the norms that serve the interests of minority groups? Riedl acknowledged how this is related to data issues to which there is still no technical solution.

Perhaps more fundamental than defining "alignment" is the *identification of the risks* that we deem significant enough to talk about AI safety in the first place. This is often a challenging task. In her paper "System Safety for Responsible ML Development: Translating and Expanding System Theoretic Process Analysis," Shalaleh Rismani (Mila) suggested that a useful tool in this respect may be the System Theoretic Process Analysis (STPA). STPA is a hazard analysis framework that starts precisely from the identification of harms and losses and the assessment of interactions between systems. It then moves to create control structures, identify unsafe control actions, and identify loss scenarios. Such features facilitate the design and implementation of accountability methods and protocols. STPA is useful also because it: (1) allows us to map sociotechnical systems, which helps us build on and expand previous analyses of the capabilities of those systems; (2) is particularly suited to capture the evolving capabilities of machine learning systems by virtue of iteratively conducting analysis of changing components; and (3) provides means of understanding causal scenarios both theoretically *and* empirically.

The discussion of Rismani's presentation brought up the challenges that practitioners are already facing when using safety engineering and STPA, as well as concerns about the limits of STPA and possible applications. For instance, could STPA address safety issues that are linked, not to harms and losses caused by mistakes in the use of ML systems but rather caused *intentionally*? Rismani emphasized that perhaps a *security* framework is typically more apt to deal with this type of cases, although those types of harms and losses can still be accounted for to some extent by STPA as control failures. Crucially, though, STPA relies on the assumption that we have some control of the system in question and that we *already know* the losses we care about. In that sense, the approach could potentially obscure some losses or face limitations when it comes to addressing the challenges of models where both inputs and outputs are quite open-ended, like chatbots.

In turn, as yet another alternative approach to risk assessment, Tegan Maharaj (Toronto) argued for the use of *deep risk mapping*, a general framework for identifying risks using AI. Broadly speaking, deep risk mapping integrates deep learning within an agent-based modeling framework. Maharaj illustrated the general deep risk mapping framework with a case study on contact tracing during the COVID-19 pandemic, showing how agent-based modeling could be used to simulate interactions among a population, allowing researchers to compare the effectiveness of different protocols for contact tracing. Maharaj emphasized that deep risk mapping often identifies underestimated risks, such as compounded harms for marginalized groups or feedback loops in climate change. In particular, it is important that deep risk mapping draws on domain-specific expertise (as in the case study, which involved close collaboration between epidemiologists and ML researchers); in order for the simulations it employs to bear sufficient resemblance to the actual world to enable successful prediction, the assumptions in the model must be guided by knowledge in the domain in which it is applied.

Discussion focused on the general role of formal modeling in mitigating the risks of AI. In particular, participants worried that sometimes formalization leads to premature operationalization; the concept of "intersectionality," it was pointed out, is often formalized in a very "conceptually thin" way, leading to an inadequate understanding of intersectional harms. That said, there was agreement that there is a role for agent-based simulation to play in modeling sociotechnical risks from AI.

(3) Looking Forward

The richness and variety of voices in the workshop enabled the identification of pressing topics of research that require further research or remain underexplored in the AI safety literature, such as issues around social impact evaluations or democratization.

In her paper "Evaluating the Social Impact of Generative AI Systems in Systems and Society," Irene Solaiman (Hugging Face) presented a framework for social impact evaluations of generative AI systems across modalities. The framework was informed by a workshop series that convened 30 experts across industry, academia, civil society, and government. It responds to the current context of social impact evaluations: growing initiatives for AI regulation, on the one hand, and the lack of standardized evaluations and of coverage across risks or demographics, on the other.

The framework divides impact in two overarching categories: what can be evaluated in a technical system and its components (which includes evaluations of bases systems, or systems that have no predetermined application), and what can be evaluated among people and society. For the latter, the framework assesses trustworthiness and autonomy; inequality, marginalization, and violence; concentration of authority; labor and creativity; and ecosystem and environment. Further, social impact is studied across seven different categories: bias, stereotypes, and representational harms; cultural values and sensitive content; disparate performance; privacy and data protection; financial costs; environmental costs; and data and content moderation labor costs. For each of these categories, the framework identifies what to evaluate and the limitations in such evaluations. As an example of the shortcoming of such limitations, Solaiman highlighted that *environmental* evaluations are often limited to the carbon emissions of training, testing, and deploying these systems, yet the energy costs of manufacturing hardware remain underexplored. Moreover, as Solaiman stressed, the environmental impact of *manufacturing* goes beyond the impact of carbon emissions and includes effects on natural resources like water.

The audience raised some concerns about the limits of evaluations. For instance, the fact that corporations in charge of evaluations may just want to build more profitable models potentially shapes how those evaluations are carried out. There is also an abundance of evaluation models, and it is unclear which matter more. Further, how to ensure evaluations can have teeth and make an impact? To this last question, Solaiman noted that the impact of evaluations depends on the topic area (e.g., if the evaluation suggests a model is likely to promote terrorist actions, its impact is more likely to be greater). More generally, Solaiman stressed the general limitations of evaluations: Evaluations just provide a signal for how the model should be used, not the solution for every issue raised by the model.

Regarding the democratization of AI, the audience once again emphasized a notable omission from much work on AI safety: community advocacy groups, who may bring significant knowledge to the table. How do we enable engagement between these groups and AI developers? As Rishi Bommasani (Stanford) suggested in his presentation "Transparency for Foundation Models: A Lost Cause or a Valiant Flight?," ensuring greater transparency may be key moving forward: In order to identify the most appropriate points for public engagement and intervention, we need to have greater insight into the full development pipeline. Bommasani discussed the Foundational Model Transparency Index, an in-progress effort to directly rate companies for their transparency and help improve transparency over time. The index specifies 100 fine-grained indicators that codify transparency for foundation models, examining the upstream resources used to build a foundation model (e.g., data, labor, compute), the details about the model itself (e.g., size, capabilities, risks, mitigation), and the downstream impact (e.g., distribution channels, usage policies, affected geographies). Findings from the index suggest that there is widespread lack of transparency throughout the pipeline. The greatest opacity exists with respect to the downstream use, as no developer discloses downstream impact of its flagship model. Recommendations from the index include that developers increase transparency for both existing and future foundation models by working closely with deployers, regulators, and downstream developers.

In the discussion, Bommasani explained how transparency is essential for at least one theory of safety: The scrutiny afforded through transparency can be a powerful tool to guarantee safety. Participants also emphasized the need to decide *collectively* which measures of transparency we want to deploy, and for which purposes, while making sure developers' goals align with ours in this respect. Relatedly, questions on social externalities and accountability mechanisms led Bommasani to highlight the fine line between transparency and related values like accountability and responsibility.

(4) Conclusions

Where should we go from here? Making progress on AI safety requires addressing many difficult questions, both technical and non-technical. Moreover, we need answers urgently; given the speed with which AI is being deployed, we can't delay developmental and political interventions. Ultimately, the question that matters most is: Given the societal impact that AI will have (and is already having), what do we want our collective future to look like? Throughout the workshop, it was affirmed that the best chance we have at designing and building that future (and at finding interventions that truly make AI "safer") is to widen the coalition working on these issues.

Jacqueline Harding is a PhD student in Symbolic Systems within Stanford's Philosophy Department. Her work sits at the intersection of (the philosophy of) cognitive science and machine learning.

César Valenzuela is a PhD Candidate in the Philosophy Department at Stanford. Their work is at the intersection of democratic theory and applied ethics.

Stanford University's Institute for Human-Centered Artificial Intelligence (HAI) applies rigorous analysis and research to pressing policy questions on artificial intelligence. A pillar of HAI is to inform policymakers, industry leaders, and civil society by disseminating scholarship to a wide audience. HAI is a nonpartisan research institute, representing a range of voices. The views expressed in this policy brief reflect the views of the authors. For further information, please contact **HAI-Policy@stanford.edu.**

Stanford University
Human-Centered
Artificial Intelligence

Stanford HAI: 353 Jane Stanford Way, Stanford CA 94305-5008

T 650.725.4537 **F** 650.123.4567 **E** HAI-Policy@stanford.edu **hai.stanford.edu**