Hard Choices in Artificial Intelligence

Kajol Sharma | Ayush Prabhu
Rprabhu.ayush@gmail.com | Kajolsharma7021@gmail.com
Keraleeya Samajam's Model College,
Khambalpada Road, Thakurli, Dombivli (East), Kanchangaon, Maharashtra

ABSTRACT

As AI systems are integrated into high-stakes social domains, researchers now examine how to design and operate them in a safe and ethical manner. However, the criteria for identifying and diagnosing safety risks in complex social contexts remain unclear and contested. In this paper, we examine the vagueness in debates about the safety and ethical behavior of AI systems. We show how this vagueness cannot be resolved through mathematical formalism alone, instead requiring deliberation about the politics of development as well as the context of deployment.

Keywords: AI ethics, AI safety, AI governance, AI regulation, Philosophy of artificial intelligence, Sociotechnical systems

1 Introduction

The rapid adoption of AI systems is reshaping many public, professional, and personal domains, providing opportunities for innovation while also generating new forms of harm. These harms are diverse, ranging from physical dangers related to new robotic systems, to economic losses related to welfare systems. In response, a broad spectrum of civil society initiatives has emerged to safeguard human domains from the effects of AI systems.

Debates about the sociotechnical gap have taken two forms. One is the proposal of normative principles to determine how the gap should be filled or who should do it. For example, the OECD Principles on Artificial Intelligence "promote artificial intelligence (AI) that is innovative and trustworthy and that respects human rights and democratic values," and are signed by governments.

This paper makes two key claims. First, AI development must be reconceived in terms of the multiple points of encounter between system capabilities and sociotechnical gaps. This requires a new vocabulary and framework to make sense of salient gaps in the context of technical design decisions, constituting a reciprocal relationship between system development and governance. Second, developers must take on new roles that

are sensitive to feedback about how to manage these gaps. This requires communicative channels so that stakeholders are empowered to help shape the criteria for design decisions.

2 Towards a Sociotechnical Lexicon for AI

At present, AI research lacks a robust sociotechnical lexicon. This would include the emerging problem space of AI Safety as well as newly-relevant questions of cybernetics in the context of present and future AI governance topics. In this section, we present a preliminary lexicon to reveal areas of overlap and divergence between these domains, enabling comparison between contemporary assumptions of AI development and possible alternative paradigms.

- **Agency** – the capacity of some agent (human or artificial) to act in order to achieve a particular outcome or result.
- **Intelligent Agent (IA)** – an autonomous entity which acts, directing its activity towards achieving goals.
- **Environment** – a domain in which an IA can perceive through sensors and act using actuators, in pursuit of a goal.
- **AI Model** – a mathematical representation of the environment, constructed through either simple rules, a model, or a combination thereof, the parameters of which may be learned from and updated with observed data.
- **Objective Function** – a mathematical representation capturing the goals of the IA.

3 The Problem of Vagueness

As AI systems are applied to more sensitive contexts and safety-critical infrastructure, normative indeterminacies are becoming more visible. Identifying the missing feedback in a given specification requires interrogating the functions of an AI system in a principled manner. This includes examining what task the AI system is trying to complete and how the system is meant to work in support of human contexts, as well as which normative standards would be appropriate to fulfill these needs. A classic example is the Sorites paradox: which grain of sand removed from a heap turns the heap into a non-heap? Such situations may yield existential uncertainty, which, if not resolvable through agreed-upon standards, may lead to arbitrary tradeoffs, compromise, or restrictions. We thus propose vagueness as a general descriptor for situations in which developers' attempts to model some domain via technical uncertainty fall short and give way to specific forms of indeterminacy. This exercise motivates the need for sustained engagement with the actual context of system development.

4 Epistemicism - Resolving Vagueness Through Model Uncertainty

Epistemicism claims bivalence as a basic condition for an object's existence. This is to say that for any given property of an object, there is in principle some sharp boundary by which the object either does or does not have that property. Illustrated through the Sorites paradox, epistemicists believe that there is an objective fact of the matter about the precise number of sand grains necessary to constitute a heap vs. non-heap, even though we may be ignorant of that cutoff point. The position thus holds that every object property or attribute must terminate at some boundary, no matter how inappreciable this boundary may be at present. This implies that acquiring more information may help reveal where the boundary actually is or could be drawn. Pure epistemicism is counterintuitive and is philosophically controversial in comparison with the claim that boundaries are semantic constructions.

- The machine's only objective is to maximize the realization of human preferences.
- The machine is initially uncertain about what those preferences are.
- The ultimate source of information about human preferences is human behavior.

However, this vision is inadequate for design situations in which human behavior is difficult to observe. Reasons for this could be empirical (sparse behavioral signals) or normative (concerns about surveillance or behavioral manipulation).

5 Ontic Incomparabilism - Respecting Value Pluralism

Meanwhile, ontic incomparabilism holds that there are fundamental limits to what our predicates or semantics can make of the world because there is no objective basis to prefer one definition of a concept to another. More concretely, even if we knew everything about the universe, there would still be no way to argue that a pile of sand "should be considered a heap" after exactly n+1 grains as opposed to after n grains. Ontic incomparabilism therefore claims that we cannot ever fully model the world by discovering additional criteria or accumulating sufficient information about it as its dynamics may be fundamentally unsuited to model specification.

Note that this position is distinct from views that the world is impossible for human minds to comprehend completely (as has been argued for specific physical phenomena, e.g. quantum mechanics) or that the world is impossible to describe accurately. Instead, the claim is that any finite number of descriptions or representations cannot exhaust the world's richness because its basic features are not readily discernible, and that there are in principle as many different ways of representing the world as there are agents capable of realizing their agency in that world. This means that modeling the world robustly would require securing the world's total cooperation with the boundaries being drawn over it.

Ontic Incomparabilism has found expression in terms of value pluralism, i.e. that there cannot or will never be an ultimate scheme for delineating human values because humans exist in the world in a way that cannot

be exhaustively represented. This transcends sociological fact (i.e. that people hold different beliefs about values, and value beliefs differently) to make an axiological, antimonist claim: values are indeterminately varied and incommensurable, and no ethical scheme could ever account for the range of values or concerns held by all humans for all time.

These conclusions have found support in the field of Computer Supported Cooperative Work (CSCW). Presenting them as a central challenge, Ackerman has described the inevitability of the "social-technical gap" of computer systems; the inherent divide between what we know we must support socially and what we can support technically. On this view, any system design requires fundamental political choices about how values of relevant stakeholders, including those indirectly affected by the system, result in some value hierarchy that may have undesirable consequences for how the benefits and harms of a system are distributed across society.

Correspondingly, the type of feedback most readily endorsed by ontic incomparabilists has been refusal, i.e. the explicit rejection of a system specification as unsuitable. This has been expressed recently through comparisons of facial recognition systems with plutonium.

6 A Framework of Commitments for AI Development

There are inherent sources of vagueness about what safety means, how it is formalized, and how it is enacted in an AI system. As a result, indeterminacies are encountered through possible design interventions that are technically comparable but normatively incommensurable. If left unaddressed or underconsidered, these may lead to harms, reinforcement of structural inequalities, or unresolved conflict across different stakeholders.

Thus, we analyzed a broad spectrum of technical, governance, and critical scholarship efforts to address the safety of AI systems, and how these fall into three canonical approaches to vagueness. For each lens, we determined the affordances and limitations of their associated cybernetic feedback modalities and the interventions that can be done with these to safeguard an AI system or improve the practices that design or govern it.

In this section, we integrate these lessons, arguing that designers should address hard choices by incorporating appropriate types of stakeholder feedback into the development and governance of the system. We also build on those lessons by explicating the role of democratic dissent as a critical additional form of cybernetic feedback in AI system development and governance, as motivated. Together, the facilitation of cybernetic feedback channels constitutes substantive commitments to the governance of the domain in which the system will operate. We thus delineate a set of commitments that would frame technical development as deliberative about the system's normativity. This recasts the traditionally linear "AI development pipeline" process as dynamic and reflexive, comprising cybernetic design principles for AI governance.

The resulting Hard Choices in AI (HCAI) Framework contains four cybernetic practices: sociotechnical specification, featurization, optimization, and integration. These activities and corresponding commitments will be introduced and discussed in the following subsections. We stress that this framework is a conceptual depiction of how to deliberate critically and constructively about normative indeterminacy. The framework may, however, help to identify concrete design approaches that can put commitments into action. In many instances, regulatory measures may form either an existing source of constraints and requirements in the development process or be informed by it. We do not advocate for particular law or policy interpretations, as these are just as contextual as design approaches, but see such translation work as a natural extension of this paper. Our framework naturally connects with and further concretizes the 'AI system lifecycle' as introduced in the OECD AI Principles.

6.1 Sociotechnical Commitments

Developers must diagnose situations of normative indeterminacy while remaining attentive to the fundamental limitations of technical logics to resolve them. This necessitates an "alertness" to all the factors responsible for the situation, including social, affective, corporeal, and political components.

AI systems are not merely situated in some pre-existing sociotechnical environment. Rather, the development of the system itself creates novel situations that intervene on social life, reflected in the distinction between pre-existing, technical, and emergent bias. These require their own formal treatment.

Developers must also acquire practical reasoning to navigate across sociotechnical approaches to a problem and determine specifications accordingly. A specification that might make sense in one context may not make sense for another, either in terms of feature detection (e.g. facial vs. handwriting) or integration scale (municipal oversight vs. nationwide surveillance).

Developers must recognize the differences between these and internalize standards that guide the indeterminate application of abstract principles to the concrete needs and demands of the situation, in a manner responsive to stakeholder feedback. These comprise distinct forms of judgment: formulating the problem, evaluating system criteria, and articulating the performance thresholds that the system must meet in order to be safe.

6.2 Sociotechnical Specification - Engaging the "Stakes" and Forms of Agency

The HCAI Framework does not identify a clear start of AI development, but it does require the initial determination of how the problem is to be formulated and tackled, mechanisms for improving this determination through feedback and dissent, and what stakeholders are already implicated or should be involved in problem formulation. Moreover, not all normative dimensions can be foreseen upfront, as hard choices may surface in subsequent development considerations.

Aware of these historical, critical, and empirical complexities, we center the need for sociotechnical specification, i.e. the process of facilitating the different interests relevant in understanding a situation that

may benefit from a technological intervention. Developers must clarify what the system is actually for– whose agency it is intended to serve, who will administer it, and what mechanisms are necessary to ensure its operational integrity. The sociotechnical specification facilitates integral interventions to determine and resolve what safety means (semantic), how it is formalized (epistemic), and how it is enacted in a system (ontic). This facilitation cannot fall exclusively on the plate of designers or developers.

To appropriately surface parity throughout sociotechnical specification, the following challenges must be taken up: (1) negotiate a program of requirements and conditions on both process and outcomes; (2) determine roles and responsibilities across stakeholders; (3) agree on ethics and modes of inquiry, deliberation, and decision-making.

In sociotechnical specification, one needs to understand the context of integration. This includes the positions of different stakeholders with their reasoning and how these relate to each other. It requires an understanding or anticipation of the impacts on social behavior, broader societal implications, and how different solutions would sit within existing legal frameworks. This yields the following dilemma: The key hard choice for a successful AI system is to include sufficient perspectives and distribute decision-making power broadly enough in development to cultivate trust and reach a legitimate consensus, while resolving the situation in a set of requirements and a process with roles and responsibilities that are feasible. While we propose these diagnostic and procedural questions for AI system applications broadly (and prospectively for more computationally intensive systems in the future), here we focus our attention on contexts that are safety-critical by nature or play an important public infrastructural role. This includes systems that integrate on a global scale, interacting with a wide spectrum of local and cultural contexts.

Solidarity is necessary to resolve this hard choice by specifying warranted interventions for the system's subsequent development. The criterion for these interventions as warranted is twofold. First, indeterminacies that would necessarily prevent the system's successful operation must be resolved in advance.

Second, indeterminacies that do not threaten successful operation must be deferred for stakeholders to evaluate and interpret according to their own involvement and concerns. In this way, interventions will align abstract development commitments with specific possible design decisions, given the particularities of the situation and the most urgent needs of relevant stakeholders. Indeed, the three subspecies of hard choices described below do not comprise a linear, abstract checklist so much as forms of situational alertness to the possibility of parity throughout the iterative development process. Ideally, the initial problematization stage identifies all the strategies and modes of inquiry necessary to track and resolve indeterminacies. This includes an appropriate assignment of roles and responsibilities across all stakeholders.

6.3 Featurization - Epistemic Uncertainty

AI systems generally represent a predictive, causal, or rule-based model, or a combination thereof, that is then optimized and integrated in the decisionmaking capabilities of some human agent or automated control

system. As such, it has to answer the question 'what information it needs to "know" to make adequate decisions or predictions about its subjects and notions of safety?'. As the model represents an abstraction of the phenomenon about which it makes predictions, the chosen model parameterization and the data used to determine parameter values delimit the possible features and value hierarchies that may be encoded. If not anticipated and accounted for, this may deny stakeholders the opportunity to evaluate design alternatives and force potentially harmful and unsafe hard choices. In this way, featurization is an epistemic intervention on the indeterminacies that may be present or latent in the context that precedes or follows system operation. Featurization specifies the computational powers of the system: how the limits of what it can model determine its assumptions about people and the broader environment, and what kinds of objects or classes are recognizable to it.

6.4 Optimization - Semantic Indeterminacy

The parameters of the system's internal model must be further determined by performing some form of optimization. This determines the input-output behavior of the model and how it will interact with human agents and other systems. Optimization extends across the design stage (e.g. training an algorithm) and implementation (e.g. finetuning parameters) and answers the question 'what criteria and specifications are considered to measure and determine whether a system is safe to integrate?'. Depending on the chosen representation, such optimization can either be performed mathematically, done manually through the use of heuristics and tuning, or some combination thereof. For mathematical optimization, the recruitment of historical and experimental data is needed to either (a) infer causal model parameters, (b) infer parameters of noncausal representations, or (c) iteratively adjust parameters based on feedback (as in reinforcement learning).

The objectives and constraints and the choice of parameters constitute a semantic intervention on how the identification of specific objects relates to the forms of meaning inherited by and active in the behavior of stakeholders themselves.To declare a system safe, it must go through a process of verifying and validating its functionality, both of itself as an artifact as well as integrated in the context of deployment. This is done with the help of engineers and domain experts who interface between the problem the system is meant to solve and the workings of the system itself.

6.5 Integration - Ontic Incomparabilism

Finally, as AI systems are rapidly introduced into new contexts, new forms of harm emerge that do not always meet standard definitions. In addition, the diversity of stakeholder expectations, as well as of environmental contexts, may challenge specifying safety for systems that are deployed across different jurisdictions. At a minimum, those developing and/or managing the system must specify mechanisms to identify, contest, and mitigate safety risks across all affected communities, as well as who is responsible for mitigating harms in the event of accidents. This can be done via general rules and use cases of safety hazards that identify terms

of consent, ensure interpretive understanding without coercion, and outline failsafe mechanisms and responsibilities.

Hence, such conditions should spell out both the technical mechanisms as well as the processes, organizational measures, responsibilities, and cultural norms required to prevent failures and minimize damage and harm in the event of accidents. Here we appropriate tradeoffs already identified by social theorists regarding the moral authority and political powers of social institutions.

7 Implications and Discussion

HCAI serves as a systematic depiction of the normative risks and sociotechnical gaps at stake in any AI system. But how should developers respond when examining particular proposed or existing systems? Here we present the normative implications of HCAI in terms of practical recommendations that go beyond existing governance and performance standards. We identify opportunities for policymakers, AI designers, and STS scholars to learn from each others' insights and adopt a cohesive approach to development decisions.

7.1 Expand the Boundary of Analysis to Include Relevant Sociotechnics - Systems, Organizations, and Institutions

Engineering and computer science disciplines have a long tradition of working with "control volumes," which are mathematical abstractions employed to render problems and their solutions in terms of technical terms. While often done in a more controlled context, the sociotechnical complexity and normative stakes of AI systems engaging in sensitive social and safety-critical domains require a more comprehensive lens. An algorithm or AI system alone cannot engage with its inherent normativity. In contrast, studies in systems safety have shown that safety is inherently an emergent property that "arises from the interactions among the system components." Such a systems lens also provides a more comprehensive starting point for controlling for safety, which is done by "imposing constraints on the behavior of and interactions among the components" of a system. This lens also explains how vulnerabilities of AI systems originate from across these components and system interactions, which corroborates insights from computer security that systems cannot be secured by addressing technical/mathematical vulnerabilities alone.However, reducing political reflection to the role of the "developer" is too narrow to adequately capture the implications for specification. Just like other actors, developers are embedded in a network and subject to power differentials themselves. Understanding how broader hierarchies of power both promote and constrain certain problem formulations is necessary to determine viable strategies for promoting system safeguards. Today, much AI research and development, system implementation and management, as well as computational and software infrastructure is in the hands of a small number of technology companies. As Gu¨rses and Van Hoboken argue, the move of tech companies to offer software engineering tools and data provision in service libraries and APIs has made the development of "values by design" an elusive task, and enabled new economic feedback loops that,

when implemented at scale, drive new forms of inequality across social groups. We believe that real success in safeguarding high-stakes systems will require forms of oversight and dissent that support machine politics and respond to emergent safety hazards through citizen deliberation, especially for AI systems developed and deployed by the private sector and state actors.

8 Conclusion

Our framework is strongly influenced by the classic work of Philip Agre, which aimed to have AI practitioners and designers build better AI systems by requiring "a split identity - one foot planted in the craft work of design and the other foot planted in the reflexive work of critique". While we embrace the spirit of Agre's work, we also believe that the critical applications of today's AI systems require a new lens that can see beyond technical practices and reframes the inherently interdisciplinary practice of AI development as critical in its own right. Apart from reflexivity, such a critical practice includes the forms of feedback that the domain of application asks for. The technical work done by AI practitioners plays a necessary but not sufficient part in development. It must be compensated by efforts to facilitate stakeholders' ability to be "full and active participants," while "the tools and techniques for doing this are dependent on the situations within the workplace...steer[ing] toward understanding different, pluralistic perspectives of how we think and act". As such, we prioritize and label the centering of stakeholder safety concerns and hard choices to guide and inform AI development as cybernetic practices. We view this paper as a preliminary for what forms these practices might take in particular development domains and will pursue this effort in future work.

Our lodestar in this project is the intuition that clarifying the sociotechnical foundations of safety requirements will lay the groundwork for developers to take part in distinct dissent channels proactively, before the risks posed by AI systems become technically or politically insurmountable. We anticipate that cybernetic practices will need to be included within the training of engineers, data scientists, and designers as qualifications for the operation and management of advanced AI systems in the wild. Ultimately, the public itself must be educated about the assumptions, abilities, and limitations of these systems so that informed dissent will be made desirable and attainable as systems are being deployed. Deliberation is thus the goal of AI Safety, not just the procedure by which it is ensured. We endorse this approach due to the computationally underdetermined, semantically indeterminate, and politically obfuscated value hierarchies that will continue to define diverse social orders both now and in the future. Democratic dissent is necessary for such systems to safeguard the possibility of parity throughout their development and allow users to define the contours of their own values, AI's capacity for specification makes hard choices possible, but its inclination to misspecification makes them necessary.

Acknowledgements

We wish to thank Michael Dennis, Joan Greenbaum, Iason Gabriel, Rashida Richardson, Elizabeth Kaziunas, David Krueger, ´In˜igo Mart´ınez de Rituerto de Troya, Seda Gu¨rses, and Alva No¨e for their constructive feedback on earlier versions of this paper. T.K. Gilbert is funded by the Center for Human-Compatible AI, as well as a Newcombe Fellowship. R. Dobbe was partly supported by the AI Now Institute.

References

[1] A.J. Hawkins, ”Serious safety lapses led to Uber’s fatal self-driving crash, new documents suggest,” *The Verge*, November 6, 2019, https://www.theverge. com/2019/11/6/20951385/uber-self-driving-crash-death-reason-ntsb-dcouments.

[2] J. Henley, R. Booth, *Welfare surveillance system violates human rights, Dutch court rules*, the Guardian, http://www.theguardian.com/technology/
2020/feb/05/welfare-surveillance-system-violates-human-rights-dutch-court-rules. [3] K. Hill, *Wrongfully accused by an algorithm - the New York times*, New
York Times, https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.
html.

[4] A. Harmon, *As cameras track Detroit’s residents, a debate ensues over racial bias - the New York times*, New York Times, https://www.nytimes. com/2019/07/08/us/detroit-facial-recognition-cameras.html.

[5] K. Hao, *Facebook’s ad-serving algorithm discriminates by gender and race — MIT technology review*, MIT Technology Review, https://www.technologyreview.
com/2019/04/05/1175/facebook-algorithm-discriminates-ai-bias/.

[6] K. Hao, *He got Facebook hooked on AI. Now he can’t fix its misinformation addiction — MIT technology review*, MIT Technology Review, https://www.
technologyreview.com/2021/03/11/1020600/facebook-responsible-ai-misinformation/.

[7] M.S. Ackerman, *The intellectual challenge of CSCW: the gap between social requirements and technical feasibility*, Hum.-Comput. Interact., 15 (2–3) (2000), pp. 179-203.

[8] D. Schiff, J. Borenstein, J. Biddle, K. Laas, *AI ethics in the public, private, and NGO sectors: a review of a global document collection*, IEEE Trans. Technol. Soc., 2 (1) (2021), pp. 31-42, 10.1109/TTS.2021.3052127.

[9] L. Andersen, *Human rights in the age of artificial intelligence*, Tech. rep., Access now, Nov. 2018, https://www.accessnow.org/cms/assets/uploads/
2018/11/AI-and-Human-Rights.pdf.

[10] *Getting the future right – artificial intelligence and fundamental rights*, Tech. rep., European Agency for Fundamental Rights, Nov. 2020, https:// fra.europa.eu/en/publication/2020/artificial-intelligence-and-fundamental-rights. [11] B. Mittelstadt, *Principles alone cannot guarantee ethical AI*, Nat. Mach. Intell., 1 (2019), pp. 501-507, 10.1038/s42256-019-0114-4, https://www.
nature.com/articles/s42256-019-0114-4.