Hard Choices in Artificial Intelligence

Roel Dobbe[1], Thomas Krendl Gilbert[2,*], Yonatan Mintz[3,**]

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 harms caused by systems to stakeholders in complex social contexts and how to address these remains contested. In this paper, we examine the *normative indeterminacy* in debates about the safety and ethical behavior of AI systems. We show how dealing with indeterminacy across diverse stakeholders cannot be captured by mathematical uncertainty alone, instead requiring acknowledgment of the politics of development as well as the context of deployment. Drawing from two case studies, we formulate normative indeterminacy in terms of *sociotechnical* challenges, captured in four key dilemmas in the problematization, featurization, optimization, and integration stages of AI system development. The resulting framework of *Hard Choices in Artificial Intelligence* (HCAI) empowers developers to navigate these dilemmas by 1) cultivating distinct forms of sociotechnical judgment that correspond to each stage; 2) securing mechanisms for *dissent* that ensure safety issues are exhaustively addressed by providing stakeholders with continuous channels to advocate for their concerns. As such, the paper contributes to a timely debate about the status of AI development in democratic societies, arguing that deliberation should be the goal of AI Safety, not just the procedure by which it is ensured.

[*] Corresponding author. *Email address*: tg340@berkeley.edu *Postal address*: UC Berkeley Graduate Division, 424 Sproul Hall, Berkeley, CA 94720-5900

[**] All authors contributed equally to this work.

[1] AI Now Institute.

[2] UC Berkeley.

[3] Georgia Tech.

Keywords: Ethical AI, AI Safety, Philosophy of Artificial Intelligence, Sociotechnical Systems

1. Introduction

The development of artificial intelligence (AI) systems is at a crossroads. Even as the specter of opaque algorithms and mammoth data structures has raised questions about what norms and values must be "encoded" into systems to ensure they are safe, there is sharp confusion about what is at stake and what is even meant by "artificial intelligence". In contrast to earlier definitions [1, 2, 3], practitioners now embrace systems whose use of machine learning is simply meant to aid in predictive analytics and other decision tools, while the language used by policy-makers still compares systems to human capabilities [4]. Acknowledging this transition, Lucy Suchman has recently described AI as "a cover term for a range of technologies of data processing and techniques of data analysis based on the iterative adjustment of relevant parameters, according to some combination of internally and externally generated feedback" [5]. Yet anthropomorphic and "us vs. them" framings for safeguarding AI systems persist, stemming from the anxiety that advanced automated systems, including vehicles, "must at times make ethical choices autonomously, either via explicit pre-programmed instructions, a machine learning approach, or some combination of the two" [6]. Stuart Russell observes that "uncertainty about rewards [i.e. the objective of a system] has been almost completely neglected in AI" [7] and has called for a new "standard model" focused on *AI Safety*, defined broadly as the design of machines that provably follow human intentions even while acting independently [8].

Building on these discussions, this paper interprets the problem space of AI ethics in terms of the distinct forms of *normative indeterminacy* introduced by system development. Influenced by work in the philosophy of law [9, 10], we define normative indeterminacy as the lack of standards, specifications, and definitions that are needed to oversee the system's operation and evaluate it as successful with respect to the problem it is meant to solve. We hold that the unprecedented epistemic powers of contemporary and near-future AI systems mean that there is no fully determinate application of human semantics to the classifications, reward functions, held-out data, and edge cases those systems instantiate. While there may be a fact of the matter about where the limits of application lie, our knowledge of those limits is itself subject to some margin of error and lack of insight [11], compromising our capacity for identifying or predicting exhaustive safety conditions in advance. Thus the priority of developers is to track indeterminacies as they appear–to become attuned to these limits of application through new forms of judgment, rather than "precisifying" the model for its own sake. This requires a major transition away from model accuracy towards being confident in one's ability to anticipate system failures and to shift development priorities accordingly [12], which we argue is a more reliable and sure path to system safety than the alternative.

Attuning oneself to the indeterminate relationship between technical specification and social context means that AI development is an irreducibly *sociotechnical* endeavor [13]. Internalizing this perspective, which we return to again and again in the paper, is the first step in recognizing how AI development procedures are problematic and where working definitions of the system's functionalities and implications are most urgently needed. From this awareness it becomes possible to perceive the limits of purely formal language for addressing the complex nature of safety conditions and to discern a variety of techniques including stakeholder engagement, questioning constraint satisfaction criteria, solicitation of audits, as well as formal tools. With practice, these techniques refine one's capacity for judgment to include a spectrum of sociotechnical interventions that are more or less suited to the problem at hand. It then becomes possible to fully index a development decision with respect to the people and infrastructure the system in question will affect, in the form of a specific normative commitment.

However, there is more than one kind of normative commitment at stake in the development of AI systems whose chosen affordances also comprise a blueprint for the worlds they call into being. We claim that normative commitments manifest in three distinct ways: epistemic (how we know things as they are), semantic (how we capture the meaning of things in language or logic), and ontic (how we relate to things in the world itself). Each requires its own form of sociotechnical judgment, which must be reconciled not through meta-learning criteria (which would merely outsource indeterminate safety conditions to a higher level of abstraction) but by prioritizing channels for *dissent* within and across the development pipeline. Dissent is necessary in order to capture errors not anticipated by designers, thus ensuring these forms' inferential assumptions support rather than subvert each other sociotechnically and indexing commitments to the context of system operation [14] rather than the whims of practitioners. This means that safety criteria manifest ultimately in the organizational form of development itself, including democratic mechanisms for stakeholder representation and public accountability.

We emphasize that our concerns, while responding to more recent iterations of AI and computer systems, are not new. The research agenda of situated design [15] and Agre's call for a "critical technical practice" [16] comprise classic phenomenological critiques of "good old-fashioned" symbolic and expert systems, in particular the need to become critical about certain formal assumptions behind intelligence and to reassess problematic metaphors for perception and action [17]. Yet much technical research today has moved beyond these critiques. Reinforcement learning (RL), for example, reflects Dreyfus' exposition of intelligence as a learned, situated, dynamic activity developed from coping with one's surrounding environment and embodying different strategies for action. However, RL poses other challenges for humans, based on problems surrounding human-AI "value alignment" [18] and the "radicalization" observed from the integration of automated systems with social media [19]. The question is no longer what computers can or cannot do, but **what we should or should not allow computers to do for us**. Given this new context, we propose AI practitioners today will need distinct *technical critical practices* that guide how system objectives may be solicited from existing and emerging political orders and validated over time through sociotechnical standards for accountability.

We thus apply an insight to AI development that scholars in Science and Technology Studies (STS) have appreciated for over four decades: any and every technological system is inherently political, requiring normative deliberation and ongoing collective participation to ensure its safety for everyone affected by it [20].

1.1. Contributions

Our key contribution is the Hard Choices in Artificial Intelligence (HCAI) framework, delineated in Section 6. It comprises productive habits and appropriate metrics for development that are able to function as forms of warranted inquiry into the system's operation. We intend HCAI to serve as a diagnostic kit for identifying the root sources of indeterminacy in present and future AI systems, and close this section with practical advice regarding the material conditions under which indeterminacies arise and a discussion of the barriers to effective normative deliberation.

In addition, we make three other contributions to the AI ethics and safety landscape. In section 2 we interpret the extant literatures in these fields, as well as relevant discussions by critics and in the public sphere, in light of classic approaches to dealing with normative indeterminacy, crystallized in the philosophical concept of vagueness. In section 3 we demonstrate the sociotechnical indeterminacy of safety through two empirical case studies of the Amazon Rekognition API and Facebook's News Feed algorithm. In sections 4 and 5 we portray *indeterminacy of specification* as the critical source of normative indeterminacy in the development and implementation of any given AI system, and motivate the use of democratic channels for dissent as integral to understand, resolve, or protect this indeterminacy as appropriate for a given development pipeline. Section 7 concludes.

2. Canonical Approaches to the Problem of Vagueness

To get a grip on complex safety issues arising in AI systems, we first consider normative indeterminacy arising from various forms of *vagueness* of a system specification, later focusing our attention on safety. Vagueness is a central topic in metaphysics and the philosophy of logic and language that has important application in system engineering and artificial intelligence. It is about the indeterminacy of our relationship with the world, either in terms of the ways we are able to perceive it, the language we use to describe it, or the world itself. It is resolved through the drawing of boundaries–forms of classification, demonstration, analogy, and other rhetorical strategies that sort phenomena into particular qualities and quantities or draw distinctions of form and content [21]. A classic example is the Sorites paradox: which grain of sand removed from a heap turns the heap into a non-heap? Such situations yield *epistemic uncertainty*, which, if not resolvable through agreed upon standards, leads to situations of *normative indeterminacy*, in which arbitrary tradeoffs, compromise, or restrictions are necessary. With AI systems mediating increasingly sensitive and safety-critical processes and infrastructures, situations of normative indeterminacy are now surfacing at an unprecedented pace. Before we discuss two case studies in Section 3, we present three canonical interpretations of vagueness in the philosophical literature, and tie them to existing vital approaches and discursive practices in the AI systems literature. This exercise enables a deeper examination of the difference in interpretation of AI safety issues within the case studies to better understand the normative indeterminacy of safety.

2.1. Epistemicism - resolving vagueness through model uncertainty

Epistemicism claims bivalence as a basic condition for an object's existence [22]. 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. While pure epistemicism is counterintuitive and is philosophically controversial in comparison with the claim that boundaries are semantic constructions [23], the essence of the position is simply that if distinct communities (or even the same person) claim the same property applies to the same object in different ways, then they are either ignorant about the property's actual boundary or are describing distinct objects.

Epistemicism has a powerful affinity with *metanormativism*, the notion that the criteria for rational decision-making are not fully known or confidently expressed because sufficient information about the environment, other agents, or oneself is absent. Because epistemicists believe that no comparable options are fundamentally "apples and oranges", as there must be some degree to which one is preferable over the other, metanormativism asserts the existence of a clear, positive value relation between available ethical actions: one must be unambiguously better, worse, or equal to the other for a given choice to be demonstrably rational. For example, William MacAskill has sought to articulate "second-order norms" that guide how one should act when multiple appealing moral doctrines are available [24]. MacAskill, whose work has been cited in support of technical work on AI value alignment and value learning [25, 26], has also proposed a "choice worthiness function" that would generate reward functions in an "appropriate" manner, where appropriateness is defined as "the degree to which the decision-maker ought to choose that option, in the sense of 'ought' that is relevant to decision-making under normative uncertainty" [27]. As such, metanormativism is a natural ally of expected utility theory and in particular the first axiom of the Von Neumann-Morgenstern utility theorem, specifying the completeness of an agent's well-defined preferences [28].

Distinct approaches to AI Safety have emerged to define the uncertain scale at which AI systems may cause social harm. At one end of this continuum is Existential Risk (hereafter referred to as x-risk), i.e. the effort to mathematically formalize control strategies that help avoid the creation of systems whose deployment would result in irreparable harm to human civilization. The x-risk literature has focused on the "value alignment problem" in order to ensure values programmed into an AI agent's reward function correspond with the values of relevant stakeholders (such as designers, users, or others affected by the agent's actions) [26]. But this position is also practically adopted by software engineers and tech enthusiasts for whom the uncertain specification of human preferences comprise an investment opportunity for new AI systems. The following quote from Mark Zuckerberg is illustrative: "I'm also curious about whether there is a fundamental mathematical law underlying human social relationships that governs the balance of who and what we all care about [...] I bet there is" [29]. The promise of such a function continues to provide guidance for designers about what decision procedures are acceptable or unacceptable for the system to follow, specifically when the goal state and risk scale are difficult to define [30, 31].

2.2. 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 [32]. 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 been interpreted to imply *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, anti-monist 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 [33]. Value pluralism is widely adopted by queer theorists who highlight how formal value specifications typically exclude certain sub-populations in favor of others [34]. For example, Kate Crawford has endorsed Mouffe's (1999) concept of "agonistic pluralism" [35] as a design ideal for engineers [36], while Hoffmann argues that abstract metrics of system fairness fail to address the hierarchical logic that produces advantaged and disadvantaged subjects and thereby disproportionately put safety harms on already vulnerable populations [37]. Mireille Hildebrandt has taken these perspectives to their logical extreme and advocates for "agonistic machine learning", suggesting that the human self should be treated as fundamentally incomputible [29]. On this view, any system design requires fundamental choices about how values of relevant stakeholders, including those indirectly affected by the system, result in some value hierarchy that has real consequences for how the benefits and harms of a system play out across society.

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 [13]. This frames the central danger in terms of software engineers who neglect certain value hierarchies, either by failing to interrogate the context of historical data or external cost biases through design choices that moralize existing structural inequalities [38]. The call to value pluralism, as such, is not opposed to pragmatism in the form of external mechanisms that regulate how our diverse commitments may be reconciled [39]. Rather, as designers compromise the public interest through incomplete system specifications that create external costs for society, they have merely reframed the central problems of modern political theory [40] and inherited the hallmarks of structural inequality. The history of social technology, from the modern census to the invention of writing, is saturated with ways in which forms of human identity were problematically obfuscated or delimited rather than protected or left undetermined [41]. This phenomenon underpins foundational concepts of twentieth century social theory [42] and deconstructionist critiques of Western philosophy as a "metaphysics of presence" [43].

2.3. Semantic indeterminism - declaring things fuzzy by nature

Finally, *semantic indeterminism* asserts that the extent to which we can determine the definition of a concept is the extent to which the members of a given community agree on that definition. Commonly associated with Wittgenstein [44], this position emphasizes the rules of language-games as defining how we refer to the world and the specific boundaries of a given community's concerns, social tastes, and modes of valuation. To again illustrate this via the Sorites: Persians, Romans, and even distinct Greek city-states may use alternative definitions of "heap" and thus confidently draw different cutoff points without ontological disagreement. Semantic indeterminism does not argue for a radical version of social constructivism according to which any claim to describe reality is arbitrary or fictional, e.g. the notion that claims about the objective world are impossible; rather, such claims simply cannot be interpreted outside the rules that particular language communities have adopted and refined over time.

Discursively, semantic indeterminism amounts to a belief in *fuzziness*, the notion that the lines between our ways of talking about “the world” are blurry and spread unevenly between distinct language communities or modes of expertise. Lessig’s famous modalities of regulation (laws vs. norms vs. markets vs. architecture) are an example of the multiple ways of resolving this fuzziness [45], as is “fuzzy logic” itself [46]. In the context of AI Safety, an exemplary discourse has formed within the Fairness, Accountability and Transparency in Computing Systems (FAccT) literature. FAccT research has harvested a multitude of definitions and tools aiming to address safety risks by diagnosing and reducing biases across various subgroups defined along lines of race, gender or social class (Narayanan 2018). While scholars have pointed out the critical and mathematical shortcomings of bias definitions and mitigation tools, these are still instrumented as means to resolve fuzziness in practice. Industry efforts have embraced bias tools as a means to engender trust in systems and make the world more socially equitable, while global efforts aim to codify algorithmic bias considerations into certifiable standards “to address and eliminate issues of negative bias in the creation of [...] algorithms” [47].

Still, the fuzzy tension between eliminating bias and winning social trust helps reveal the inconsistent determinations of what safety means throughout the entire lifecycle, including which norms should guide design and use decisions. As some argue, “it is important to acknowledge the semantic differences that “fairness” has inside and outside of ML communities, and the ways in which those differences have been used to abstract from and oversimplify social and historical contexts” [48]. For example, scholars have emphasized important semantic differences between “individual” and “social” fairness that could help clarify and procedurally reshape the way formal fairness criteria are reconciled with policy objectives [49].

We propose this fuzziness is best understood as a *sociotechnical* problem for AI development, i.e. that systems’ “core interface consists of the relations between a nonhuman system and a human system” [50], with various dimensions (e.g. users, citizens, operators, regulators), whose construction is hindered by limited knowledge, subject to error, of how key technical innovations bear on human contexts. Even carefully-designed formalisms that are sensitive to the implicit concerns of human agents are not guaranteed to learn the right preference structures in the right way without new forms of surveillance, control, and assigned roles for both humans and the systems themselves [51]. Such system setups are limited in three ways: (1) they can never formalize everything, and require subsequent developers to organize around them; (2) they attempt to resolve (and thereby confuse) content and procedure from the get-go, rather than treat the sociotechnical development of AI systems as a dynamic problem; (3) they are limited in addressing wider spectra of values across distinct peoples and cultures.

3. The Indeterminacy of Safety

In this section we examine the sociotechnical problem of "safe" AI empirically through case studies that distinctively highlight how indeterminacies are obfuscated on a societal scale. While safety has many definitions depending on context, we here specify it in terms of *protection* from harm or injury [52], *robustness* in the face of adverse conditions [53], and *resiliency* in response to stress or difficulty [54]. The indeterminacy implicit in these terms makes it difficult for safety to be deliberated in a meaningful and consistent way. Moreover, distinct sources of indeterminacy–in epistemics, semantics, or ontics–are responsible for specific critical examples of AI systems behaving in an "unsafe" manner and the specific forms of controversy they engender.

3.1. What Warrants Facial Detection? AWS Rekognition and the ACLU

Amazon Web Services (AWS) intended Rekognition to be a commercially available cloud based ML tool that helps enterprises with setting up and searching image based datasets for various facial recognition tasks. After Joy Buolamwini first pointed out gender and racial biases in commercial facial recognition software of IBM, Microsoft and Face++ [55], attorneys at the American

Civil Liberties Union (ACLU) tested Amazon's system's limitations by creating a data set made up of publicly available arrest photos, used these to train Rekognition, and then queried the system to find a match against photos of current members of Congress [56]. The ACLU reported not only that false identifications were found, but that of those members of Congress falsely identified as matching the arrest photo database, a disproportionate number (40 percent) were people of color. The ACLU concluded that Rekognition shows a bias in its predictions, making it unsafe to implement in high stakes contexts which disproportionately affect people of color (e.g. law enforcement). However, the AWS research team issued a rebuttal, commenting that Rekognition was used against its articulated design specifications, including a lower-than-recommended confidence threshold for a high-risk task, and that the dataset's construction did not account for possible inherent bias [57].

First, it is unclear what *protection* means in development situations where private and public definitions are both at stake. For example, the ACLU's claim that facial recognition systems beneath a given accuracy threshold should not be used by law enforcement is rooted in the intuition that correcting misclassifications *a posteriori* is unacceptable, as it imposes Amazon's internal definitions of "harm" and "vulnerability" onto anyone that encounters the system. In contrast, Amazon's claim that the system does work according to design intentions and that the ACLU study used inappropriate settings makes sense in the context of system optimization, as the AWS team converged on the system's architectural parameters through agreements with the private contractors who intend to use it. It was not resolved whether the safety of Rekognition should be determined by its protection of *private* contracts (whose context is the online verification of edge cases by self-interested parties) or *public* assurances (whose context is the willingness to shield vulnerable communities from harm). Thus the legal contexts for AWS and the ACLU's definitions of protection are orthogonal, and the loci of perceived stakes are at two different points in the development pipeline.

Second, conditions of *robustness* must be specified according to the distinct expectations of designers and users, leading to inconsistent standards for platform governance. One can imagine AWS issuing a different response that included an apology to members of Congress, a request for the ACLU to expand its "testing" to other social domains, and a promise to improve Rekognition's accuracy going forward. However, this strategy might also become an object of public outcry; for example, Waymo regularly publishes safety reports on its vehicles but still faces the ire of Phoenix residents, who complain that "They didn't ask us if we wanted to be part of their beta test" [58]. While research and advocacy has led to improvements in commercially available software [59], many American cities have now banned the use of facial recognition by municipal agencies, citing surveillance concerns, local community interests, and social prejudice [60]. And under mounting pressure in the wake of the murder of George Floyd, multiple companies have stopped or paused selling their products to police forces. The question is whether cities and other social domains should be made ready for facial recognition tools (through e.g. concrete institutional reforms), or the tools should be made ready for cities to use them with confidence (via e.g. ongoing technical refinement). How one answers this question places the onus of sociotechnical robustness on either the *public officials* who administer the system or the *engineers* who build it, a problem defined as the "moral crumple zone" of moral and legal responsibility [61].

Third, a system's *resiliency* requires a metric of optimality, according to which abnormal dynamics can be discerned, diagnosed, and remedied. At a minimum, facial recognition assumes some definition of what a face is, and what the good, bad, and inaccurate ways there may be for identifying them. Recently, Luke Stark has argued that facial recognition tools are a form of racism that will incline any power structure towards discriminatory policies because the ability to rank facial features at scale will generate categories that can be used both to solder somatic attributes to personality characteristics and to legitimize political decisions [62]. Meanwhile, Wang and Kosinski claim to detect sexual orientation with the aid of deep neural networks and that the predictive power of AI models can be harnessed to discover patterns in facial features beneath human awareness [63]. The findings generated controversy [64], and a review of over a thousand studies on emotion expression found that efforts to “read out” people’s internal states from an analysis of facial movements alone, without considering context, are at best incomplete and at worst entirely lack validity [65]. However, Kosinski defended the study as an effort to “understand people, social processes, and behavior better through the lens of digital footprints” [66], highlighting the vague relationship between feature orderings in particular contexts and the general goals or ends that define human flourishing.

3.2. What Counts as a Community? Facebook News Feed and Sandy Hook

Facebook News Feed is a content generator that shows information about friends, upcoming events, birthdays, and targeted advertisements. Based on time on site and rate of click-through–the primary metrics for how Facebook both measures engagement and sells ad space–News Feed provides content optimized for individual users. The feature is highly controversial both for deriving profit from user attention (potentially eroding mental health and cognitive aptitude) and for its supposed spread of misinformation, particularly with respect to conspiracy theories such as the Sandy Hook massacre. This was an object of focus in a July 2018 interview between CEO Mark Zuckerberg and tech journalist Kara Swisher [67]. In the interview, Zuckerberg defended Facebook’s decision to permit Alex Jones–who had repeatedly insisted that the Sandy Hook massacre was a hoax–to continue to post to the site and therefore have his content aggregated by News Feed, as it was unclear if Jones was “intentionally getting it wrong”or “trying to organize harm against someone” in particular. In fact, Facebook did ban Alex Jones alongside six other provocateurs in early May 2019 [68], leaving unresolved what kind of aggregation metric, if any, is appropriate for handling misinformation in a more general sense.

The quote above illustrates how Zuckerberg views Facebook’s unprecedented epistemic power, i.e. the capacity to redistribute semantic boundaries however it prefers across national, ethnic, and economic categories, as sufficient grounds for its project to create a single macro-sociability standard for the entire world.

Here the criterion for *protection* remains torn between Facebook's private terms of service reflected in its application program interface (API), and its commitment to public security and enforcement. Facebook struggled to find an objective safety metric that reconciled freedom of speech with freedom from hate speech, leaving it forced to make judgment calls about the relationship between certain levels of online semantic contagion and the "offline" reality (i.e. that Sandy Hook and the Holocaust happened [69]) implicated in the platform's existence. It remains to be seen whether Facebook's new Oversight Board can make productive strides to resolve these indeterminacies. Early critiques indicate that it is a bold move to prevent government regulation [70] and cannot address all harms perpetrated and perpetuated over Facebook [71].

Facebook also lacked consistent standards for *robustness* between offended communities (who have a lower threshold for harm), and offending communities (who have a higher one), and content moderators (who must choose which thresholds are enforced). The interview with Swisher made these inconsistencies clear. Swisher challenged Zuckerberg's commitment to the ethos of "move fast and break things", arguing that all misinformation ought to be removed wholesale from the Facebook platform and evaluated before it is posted. Zuckerberg insisted that this content should only be removed in cases of immediate harm or when flagged by a sufficient number of users. Yet the notion of solving problems as they arise in ways that are sensitive to the (cultural) context places enormous onus on Facebook's own definitions of free speech and harm as applied to their users. In contrast, Zuckerberg's argument made more sense in the context of system optimization, as the best way of converging on value standards agreeable to Facebook users is through continuous confirmation of user engagement through verification metrics. Zuckerberg and Swisher thus favored two irreconcilable robustness standards for precise conditions of harm.

Most glaringly, Facebook lacked any kind of failsafe *resiliency* procedure for harm once it was committed, including how to minimize it and prevent future occurrences. Lacking any metric for what would constitute a well-functioning or "suboptimal" community, Zuckerberg could make no promises or reparative contracts that would right or even acknowledge specific wrongs that had been done between two communities (InfoWars vs. opponents), several communities, or some other unspecified number. This is due to Facebook's prioritization of a global standard for sociability beyond any notion of sovereignty, i.e. there is no "right reason" that could make the difference between rights and wrongs beyond its highly plastic terms of service. As quoted earlier in this paper, Zuckerberg's own commitment to a "fundamental mathematical law underlying human relationships" implies that metrics for sociotechnical resiliency would even constitute harm in their own right–if the company goal is to derive the general formula for social glue, then squabbles between sub-communities comprise mere learning opportunities for system optimization, not true crises of governance.

3.3. Epistemic vs. Ontic Indeterminacies

The cases of AWS Rekognition and Facebook News Feed illuminate distinct ways in which AI systems, even if they have data on the social patterns of billions of people or crystal clear terms of use, can fall under threat of philosophical vagueness. Both teams were completely confident in their own preferred criteria, whether an ironclad API or perfect information about the communities under observation. Rather, the problem was defining what counts as a community given indeterminate boundaries of content sharing, and which features of a face warrant detection independently of contractor agreements, both of which exist as political rather than narrowly technical problems. The arbitrary adjustments of raising the accuracy threshold for Rekognition and lowering the threshold for community harm via News Feed demonstrate the true purpose of both optimization standards: to compensate for a lack of cohesive and consistent safety metrics. In both instances, fuzziness was not appropriately resolved or acknowledged but shrugged off as a burden to be hefted by already-vulnerable communities.

As emphasized recently by Brian Cantwell Smith, truly intelligent systems need to *register* the world, i.e. "find it ontologically intelligible in such a way as to support our projects and practices" [72]. Our cases demonstrate that, beyond the old philosophical chestnuts of epistemicism and ontic incomparabilism, the sociotechnical implications of metanormativism and value pluralism are not contradictory–it is possible for judgments about the structure of human preferences simultaneously to be "precisified" with the aid of AI tools and to remain indeterminate in their communication and normative application. In order to take the task of registration seriously, developers must ask themselves: how can sources of indeterminacy be appropriately diagnosed? According to what criteria may claims be properly evaluated? And what would it mean to hold these sociotechnical determinations accountable?

4. Shepherding Vagueness: Intuitive Comparability and Indeterminacy of System Specification

We turn to Ruth Chang's philosophical work [73, 74] to develop a sociotechnical framework for AI Safety. Beyond Epistemicism, Semantic indeterminism, and Ontic incomparabilism, Chang proposes a fourth position, *intuitive comparability*: while many human values are incommensurable, we are nevertheless able to articulate evaluative differences to make comparisons, even if two desires or concepts are not directly measurable against each other [75]. This allows people to make informed compromises between options based on practical deliberation regarding one's overarching goals. Consider an everyday example of choosing between a banana and a donut for breakfast. Your values to address nutrition (e.g. to lose weight on a diet) and ensure tastiness are qualitatively distinct. While these values are objectively incommensurable, you can evaluate options by interpreting them in the context of wanting to go on with your day or maintaining a particular lifestyle.

Intuitive comparability is relevant for what Chang calls *hard choices*: when different alternatives are on a par, "it may matter very much which you choose, but one alternative isn't better than the other [...] alternatives are in the same neighborhood of value [in terms of how much we care] while at the same time being very different in kind of value". As such, making hard choices requires normative reasoning: "when your given reasons are on a par, you have the normative power to *create* new will-based reasons for one option over another by putting your agency behind some feature of one of the options." In other words, even epistemicists would have to acknowledge hard choices in AI development because systems cannot reflect more deeply about their own goals in the face of unanticipated situations, while human designers can. In other words, a situation of normative uncertainty for the designer could well be one of normative indeterminacy for the system, which would fail to satisfy the first axiom of VNM utility in cases where the designer herself is able to satisfy it.

We propose intuitive comparability as a general analytic for situations in which developers' attempts to model or resolve vagueness with technical uncertainty fall short and give way to specific forms of indeterminacy. Put differently, hard choices are moments in which distinct technical interventions are feasible, yet would register the world in mutually exclusive ways (Section 6 will cover examples based on the case studies in Section 3). Abstract ethical principles, as developed with unprecedented enthusiasm in recent times [76], may orient developers towards the "neighborhoods of value" at stake in hard choices, but these do not address the fundamental tradeoffs that exist in adhering to various principles [77], let alone how these are valued across different stakeholders. As a result, navigating indeterminacies in AI development requires its own specialized research domain as well as the maintenance of a safety culture that is beholden both to the normative stakes of the application domain as well as the operational success criteria of the system.

Our essential claim is that matching safety principles with technical development procedures is fraught with hard choices, which are often implicitly or explicitly made by developers on behalf of various stakeholders. Indeterminacy is thus encountered as a local tension between technical interventions that are developmentally comparable but normatively incommensurable. It is not a problem to be avoided, but a byproduct of the ways that systems are both situated within and themselves re-situate the reproductive mechanisms of social order.

4.1. Implications of intuitive comparability for AI Safety

Good judgment in AI Safety is about orienting the epistemic powers of development towards the normative stakes of the domain. Developers must do this by reflexively acknowledging, communicating, and facilitating hard choices and their consequences in ways that reveal possible compromises or equilibria conditions to make systems normatively accountable to relevant parties. While necessarily agonistic, it is only through this process that the work of AI development can identify where indeterminacy in fact originates and what range of strategies could exhaustively address it. We thus endorse intuitive comparability not as a superior philosophical position, but as a diagnostic lens from which to investigate and propose what a cohesive sociotechnical approach to AI Safety would look like. This has several concrete implications.

Descriptively, intuitive comparability permits a richly pragmatic gloss of the AI development pipeline in terms of a basic *indeterminacy of system specification* whose various sub-forms–the system's learning criteria, the ways it secures consent from those who use it, the kinds of models it is permitted to learn–align with the canonical forms of vagueness described in Section 2. This presents two challenges. First, AI systems need to encode choices made by or on behalf of a diverse group of developers and affectees, including divergent values and interests [78]. Our goal is to build an analytical framework to draw attention to these moments and facilitate bridges between the ongoing technical contributions of AI Safety research and the core substantive insights of social sciences and the humanities. Second, the values that stakeholders care about are often complex and not readily translated into a solution that suits all needs, which can lead to situations of "moral overload" that require thinking outside of one's traditional design space [79]. As such, we extend and elaborate the argument of Hadfield-Menell and Hadfield that acknowledges the need to address misspecification between reward functions and wider social institutions [30]. A crucial corollary of the above is that every AI system is fundamentally political even before it can be conceived of as technical, a claim we further develop in Section 5 before proposing the diagnostic framework.

Diagnostically, intuitive comparability provides a foundation for compromise and adjudication between possible value regimes based on the dynamic redrawing of a system's formal boundaries and design parameters via qualitative feedback. The "hard choice" moments are those where different communities may clash and development criteria must be revisited. This matches the insight from participatory design that the design of (AI) systems restructures the context in which users or other affected stakeholders exist: "values emerge, whether you look for them or not" [80]. Iversen et al. argue this requires an "*a priori* commitment to cultivate the emergence and discovery of local expressions of values whilst being mindful of further expression of values during the course of the design process" [81]. Taking intuitive comparability as the compass, a development process can accommodate the pluralist perspectives and value commitments of marginal stakeholders and deal with the inherent emergence of value conflicts during later stages, based on a capacity for mediating different definitions (particularly epistemic and ontic) of safety. Consequently, abstract criteria for "appropriateness" may not be applicable to AI Safety until indeterminacies are addressed throughout the development pipeline by incorporating the hard-learned lessons of values and ethics in human-computer interaction design and STS [82, 83], which the technical AI Safety community has yet to absorb.

Like the cables that hold up a suspension bridge, AI Safety may be defined as the successful expression of and consilience between technology, norms, and politics in the development of a system, such that its operational success criteria are able to be affirmed by all stakeholders. This means that the relations that comprise the conceptual space of stakeholder preferences, values, and identities are distinctively tracked and confirmed without collapsing into each other. Just as the "stress point" of civil engineering is the identified and agreed-upon maximum strain the bridge can handle before buckling, the critical point for safe and human-compatible AI is the safeguarding of intuitive comparability, i.e. the capability of AI systems to support pluralist value hierarchies while preserving shared moral agency–the power to engage in, contribute to, and meaningfully contest the system's normative commitments.

5. Power and politics in technology development

Because the value hierarchy designed into a system will determine the space of actions available to it (as well as those that the system forecloses), it is crucial to acknowledge and account for the power and elevated status of design work [84]. This means recognizing developers' tendencies to prioritize certain actors and networks over others. Haraway [85], Harding [86], and other feminist scholars would argue that we cannot escape having some agenda: researchers are themselves situated in the social world they study.

Technology development is inherently political. [87] argue that conflicts should be expected and that "[i]t's not the IT designer's job to cover up or try to solve political conflicts that surface [..] it is their job to develop different design visions and assess their consequences for the affected parties." And design itself is political, both as a tool and a discipline. In exploring how technology and design can reinforce historical racial patterns, [41] criticizes the inherent forward movement implied in many modern design approaches, highlighting that they rarely "allow us to slow down and let ourselves breath in ways that might be useful," concluding: "If design is treated as inherently moving forward, that is, as the solution, have we even agreed upon the problem?" This perspective updates past insights, as [88] previously identified successful conditions for participatory processes, stating that these "are not likely to produce consensus, but they may reduce public mistrust and hostility toward political and administrative institutions in order to allow détente [..] in fact, détente is a more appropriate and realistic goal."

These perspectives corroborate a key ontological implication of AI development, also emphasized by Smith, beyond the descriptive and diagnostic criteria listed above: registration practices must refer not just to the world-as-registered but to the world as such, and “any system...must ground its deliberations in full-scale judgment *at every step of the inferential chain*, in order to ensure that its representations never take leave of accountability to the world” [72]. In other words, because different stages of AI development register the world in distinct ways, the relationship between system and world is inherently problematic. As we shall see, these stages correspond to the canonical forms of indeterminacy described earlier, and developers must cultivate new forms of political judgment, made possible through *dissent*, to render the entire system safe in ways that are both internally commensurable and externally accountable. [4]

6. A Framework of Sociotechnical Commitments for AI Development

Achieving safe AI by safeguarding stakeholders' access to hard choices requires expanding the development lifecycle towards including ongoing attunement to the vital indeterminacies and tradeoffs of particular social dynamics and implications of an AI system [89]. We delineate a set of commitments for treating AI development as *sociotechnical*, requiring technical development to be informed by and work in concert with deliberation about the system's normativity. These commitments recast the traditionally linear “AI development pipeline” process as a fundamentally dynamic and reflective practice.

We coin this cycle the *Hard Choices in AI (HCAI) Framework.* While AI development can be unstructured, sometimes sticky, and often nonlinear, the circular flow of stages indicated in Figure 1 maximally reflects the sociotechnical nature of real world development. With this, we introduce a framework or certification analytic that AI developers and other parties can use to identify and track forms of indeterminacy, i.e. the hard choices that arise in or are made

[4] Today, most 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. While we do not discourage these actors from drawing inspiration from our proposal for development, we believe that success will require democratic structures and procedures.

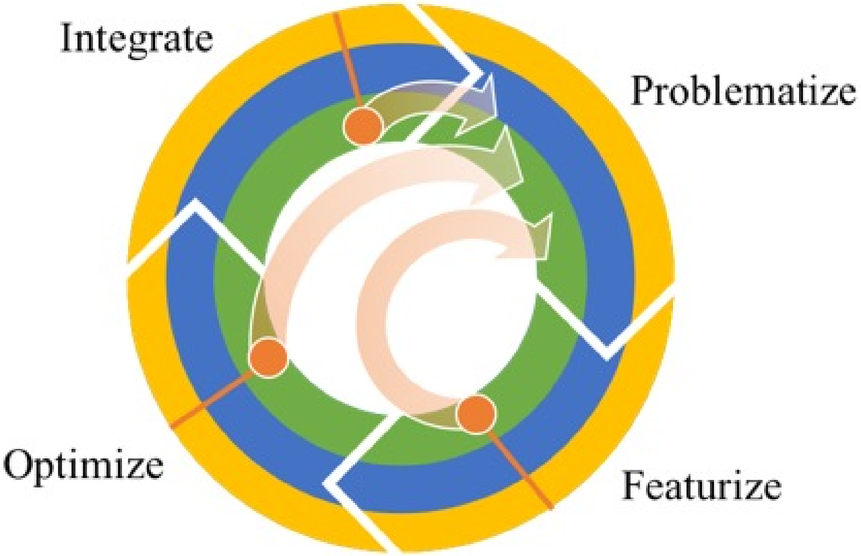

Figure 1: The Schematic Relationship between Development Commitments. The three concentric circles denote the formal, substantive and discursive commitments. Orange circles denote the occurrence of "hard choices", moments where normative indeterminacy arises, which require revisiting Problematization.

implicitly in AI development. It contains four predominant practices: problematization, featurization, optimization and integration. For each practice, there are three types of commitment: formal, substantive and discursive. These activities and commitments will be introduced and discussed in the following subsections.

We stress that this framework is diagnostic, and is not meant to resolve forms of indeterminacy but rather create a shared language to deliberate constructively about them. The framework may however help to identify concrete design approaches that can put its commitments in 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 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.

6.1. Sociotechnical Commitments

Formally, it is the core challenge for developers to diagnose situations of normative indeterminacy, i.e. material situations in which a finite set of possible specifications are each understood as technically viable yet normatively incommensurable between or within stakeholders. More importantly, developers need to understand the fundamental limitations and potential danger of technical logics to resolve them. This necessitates an "alertness" to all the factors responsible for the situation that one develops for, including social, affective, corporeal, and political components [90], requiring other forms of non-traditional expertise to weigh in and reframe or extend formalizations. Crucially, AI systems are not merely situated in some pre-existing sociotechnical environment; rather, the affordances of the system itself comprise novel situations that intervene on social life, reflected in [91]'s distinction between pre-existing, technical, and emergent bias, which require their own formal treatment [92].

Substantively, every stage of AI system development requires communication channels that allow stakeholders to assign local and individual meaning to different specifications and solutions. Following [93], we emphasize the need for dissent mechanisms that are able to track the intuitive comparability of different design options and their related value hierarchies by soliciting substantive feedback and concrete interests of affected groups. In contexts where a policy or solution is set by a majority or powerful player, "[s]uch dissent is needed not simply to keep the majority in check, but to ensure that decision-making is deliberative—undertaken in an experimental spirit—rather than simply imposed" and thus help "determine what problems are genuinely of public concern." These channels resituate AI development as an opportunity for communities to reimagine their own moral boundaries, helping both to ensure the expression of intuitively comparable human values and resolve situations of moral overload [79].

Discursively, the acquisition of practical dispositions that enable one to navigate across a spectrum of sociotechnical approaches to a problem and determine respective development specifications. 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 (outlined below in terms of sociotechnical virtues) that guide the indeterminate application of abstract principles to the concrete needs and demands of the situation. This comprises distinct forms of judgment: formulating the problem, evaluating system criteria, and articulating the needs that the system must address in order to be safe. Such ongoing engagement requires "reflexive inquiry [that] places all of its concepts and methods at risk [...] not as a threat to rationality but as a promise of a better way of doing things" [94].

To illustrate our sociotechnical commitments, we will refer to the AWS and News Feed case studies described earlier. We consider the relevant hard choices made throughout the development of both systems and illustrate their political impact. At distinct moments of formal specification, we ask: (1) how does indeterminacy manifest and what forms may intuitive comparability take? (2) In what concrete ways might dissent mechanisms address these issues? (3) What forms of judgment are needed to ensure developers are prepared to manage the indeterminacies their systems occlude or generate?

6.2. Problematization (defining 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. As depicted by the orange circles and arrows in Figure 1, this may require revisiting normative deliberation and update earlier specifications based on progressive insights. Aware of these historical, critical, and empirical complexities, we center the need for *problematization*, 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. This inherently asks for facilitation that cannot fall squarely on the plate of designers or developers.

To initiate and maintain intuitive comparability through problematization, the following challenges must be taken up: (1) Formal: negotiate a program of requirements and conditions on both process and outcomes; (2) Substantive: determine roles and responsibilities across stakeholders; (3) Discursive: agree on ethics and modes of inquiry, deliberation and decision-making. In problematization, 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:

Inclusion: What stakeholders are directly involved or indirectly affected by issues and solution directions considered? How is power and agency assigned along the process of development and integration? How are the boundaries of the AI system and its implications determined?

Resolution: What deliverables or outcomes are expected or envisioned for the project? What variables and criteria are needed to measure these outcomes? What ethical principles and decision-making process is needed to achieve resolution across different stakeholders? What conditions will allow both supportive and dissenting groups to express their concerns and contribute meaningfully to the development and integration of a resulting system?

The key dilemma 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 and/or serve an important public infrastructural function. 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 dilemma by specifying *warranted interventions* into the system's subsequent development, whose criteria aim both to resolve those indeterminacies that would necessarily prevent the system's successful operation while deferring others that must be left in the hands of stakeholders according to their involvement and loci of situational concern. 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. Here we follow Dewey's pragmatic formulation of epistemology: problematization is "the project of determining which modes and patterns of reflective inquiry best promote and enhance operational success" [95]. Indeed, the three species of hard choices described below do not comprise a linear, abstract checklist so much as forms of situational alertness to the possibility of intuitive comparability 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.

The cultivation of solidarity should not be understood as an automatic identification with the needs or expectations of stakeholders–which are necessarily distinct from those of designers–but as motivation to bolster the forms of agency and possibility which are already implicit in stakeholders' identities but can be made actionable and tractable through the work of development itself. Solidarity thus empowers developers to establish conditions under which stakeholders, in defining their own relationship with the system, are committing acts of self-determination rather than resignation [96]. Here, we endorse the vision for postcolonial computing by [97] which “acknowledge[s] stakeholders as active participants and partners rather than passive repositories of ‘lore’ to be mined”, and “acknowledge[s] and embrace[s] heterogeneity in design, rather than attempting to control or eliminate it.”

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 decision making 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 feature 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.

To harness the intuitive comparability at stake in featurization, the following challenges must be taken up: (1) Formal: make explicit and negotiate what can and cannot be modeled and inferred, crystallized in the underfeaturized/misfeaturized dilemma; (2) Substantive: engage stakeholders to challenge and inspire modeling assumptions to ensure application aligns with contextual expectations; (3) Discursive: validate the design with stakeholders to anticipate possible value conflicts that can arise due to the gap between model/system and world and plurality of values during deployment, preparing to revisit the modeling tools and methodology. 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 (e.g. faces) are recognizable to it. At a minimum, stakeholders must answer the following:

Underfeaturized: What possible model parameters do we choose not to include? What features will the model interpret as incomparable that may in fact be open to normative deliberation?

Misfeaturized: What features or actions do we choose to parameterize? What forms of dissent will be foreclosed by elements of computation, and for whom would this matter?

The danger lies in failing to adopt a mode of parameterization that is both computationally tractable and normatively defensible, which–given finite time and material resources as well as the vested interests of specific stakeholders–will err towards under- or overspecification in ways that developers cannot perfectly anticipate. The spirit of the dilemma is crystallized differently in distinct algorithmic learning procedures; for example, the division between model-based and model-free reinforcement learning essentially bears on what kind of control system is being designed and, respectively, whether this specification establishes a permissible space in which a given problem can be formulated and represented causally or merely defines permissible predictive signals (e.g. rewards, elements, qualities) within the environment. At least some corresponding social qualia will thus be made computationally commensurable despite being experienced as incomparable (as is often true for race, and was the root of the AWS Rekognition dispute) or are artificially bifurcated between articulated and unarticulated features despite their inherent comparability in lived experience (as Facebook News Feed was willing to do, in violation of community norms of information sharing and integrity).

Formally, the dilemma manifests in choosing a model capacious enough to represent the nature of the environment in a manner necessary to keep the system and its affected stakeholders safe, but constrained enough that its training would not be intractable, violate safety boundaries [98], or assume too much about private information that risks compromising stakeholders' self-determination. Imposing modeling constraints necessarily creates technical bias, which may take away space for stakeholders to express or protect their own specific values in terms of the phenomena permitted or excluded by the model's system boundaries [92]. There is already some technical work acknowledging this as a formal dilemma with no optimal solution in the context of reinforcement learning [99, 100]; the deeper sociotechnical point is that the criterion for these constraints, which entail a choice of the moment at which a model must remain technically ignorant or intentionally suboptimal, must be specified in terms of a commitment to the self-determination of stakeholders rather than a desire for maximum technical proficiency.

In the context of News Feed, the dilemma rested in the choice between some hard-coded limitations into what sorts of content are visible, vs. permitting the algorithm to extract whatever signals it needs in order to maximize its predictive accuracy for purposes of content aggregation. Zuckerberg, as we saw, unilaterally favored the latter approach, but had to interpret his user base as ignorant about the shape of their own value commitments to justify this ("there are things that different people get wrong"). Facebook may have continued its model-free epistemic agenda if the ensuing media controversy, acting as an ersatz mode of stakeholder dissent, had not compelled it to change tack and specify News Feed's model to exclude misinformation peddlers. Meanwhile, we have already noted that AWS did not address the distinctive politics of facial feature detection in its public response to the ACLU. They could have done so accordingly: either transition to a more automated decision procedure that would secure feature detection against direct human oversight while increasing accuracy for the congressional dataset at stake, or adopt a new governance structure to mediate the ethical and legal applicability of the tool and ratify the specific environmental conditions the system is allowed to represent.

All such instances require *context discernment*, the disqualification of specific features and modeling actions that, while technically proficient, are judged to be sociotechnically inappropriate within the problem space at hand. Here we draw from [101]: "The task of the craftsman is not to *generate* the meaning, but rather to *cultivate* in himself the skill for *discerning* the meanings that are *already there*." Featurization needs to anticipate how the model would be integrated in and interacting with the context of deployment, how else it could be (mis)used, what bias issues may arise during training and how to account for and protect vulnerable affected groups, and how chosen objective functions may generate externalities, as well as who is likely to bear their cost. In the event no consensus is reached and dissent persists, the option of not designing the system should be preserved [102].

6.4. Optimization (semantic indeterminacy)

The parameters of the system and predictive features used 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 (most often) enacted through a combination thereof. For mathematical optimization, the recruitment of historical and experimental data is needed to either (a) infer causal model parameters (e.g. for system identification, an inference practice common in control engineering [103]), (b) infer parameters of noncausal representations, and/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.

Therefore the following challenges must be taken up: (1) Formal: assess the extent and limitations with which the optimization criteria and procedure can translate and respect specifications, crystallized in the validation/verification tradeoff; (2) Substantive: codify a validation procedure for empirical criteria that conforms to stakeholders’ specific concerns, addressing specifications not covered through mathematical optimization; (3) Discursive: adjudicate and modify verification and validation strategies over time as indeterminacies of featurization and integration continue to be highlighted. 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. Here, the minimum requirements for safe outcomes are impartial assessments of the following questions:

Verification: Does the system meet its specifications (was the right system built)? Are the needs of prospective users being met? Is the system able to predict or determine what it was meant to?

Validation: How does the system perform in its empirical context (was the system built right)? Does the system behave safely and reliably in interaction with other systems, human operators and other human agents? Is there risk of strategic behavior, manipulation, or unwarranted surveillance? Are there emergent biases, overlooked specifications, or other externalities?

This dilemma poses several concrete challenges for development. First, systems that are mostly optimized in a design or laboratory environment fall inherently short as their data cannot fully capture the context of integration. In the development of safety-critical systems, this design issue is acknowledged by the need to minimize any remaining errors in practice (through feedback control [104]) and putting in place failsafe procedures and organizational measures as well as promoting a safety culture. Second, accounting for interactions with other systems and human agents is not to be taken lightly and is heavily undervalued in current AI literature [105]. For example, the overspecification of environments through simulation (as now popular in the development of autonomous vehicles) may backfire if the optimization scheme overfits the model for features that are not reflective of the context of integration. Third, a lack of validation and safeguarding systems in practice can result in disparate impacts [106] and failures. This is especially pertinent for underrepresented (and undersampled) groups that are often not properly represented on AI design teams [107]. For systems that are "optimized in the wild" with reinforcement and online learning techniques, these considerations are even more acute, although recent efforts have proposed hybrid methods that can switch from learning to safety-control to prevent disasters [108]. This technical point, which mirrors the well known bias-variance tradeoff, becomes *sociotechnical* at the moment when the choice of optimization procedure is interpreted from the standpoint of jurisprudence applicable to the domain.

These challenges are present in both our case studies. The designers of AWS facial detection created the system with common commercial tasks in mind (e.g. sentiment analysis), and determined their own confidence levels through extensive internal verification. However, the ACLU bypassed the Rekognition API and applied detection to a task that it was not explicitly meant to perform, i.e. matching faces of politicians to those of recent arrestees. This revealed indeterminate criteria in the way the system was validated. Likewise, Facebook relied on user engagement as a verification metric for how certain kinds of content should be weighted for specific user groups. However, quantitative shifts in weight could never fully compensate for the qualitative exposure to harmful content, depending on the under-specified semantics of "harm" that Facebook relied upon to track their own optimization assumptions. In both cases, appropriate validation criteria remain out of reach, pending sustained user feedback whose bearing for overall development would require the force of law or regulatory oversight.

The cultivation of *stewardship* is needed to reconcile the technical problematics of value alignment with optimization procedures capable of providing qualitative assurances to the particular sociotechnical stakes of the domain, whether physical, psychological, social, or environmental. System engineers must internalize an understanding of how the finitude of their teams' tools and procedures bears on the urgency felt by stakeholders towards objects of sociotechnical concern, compelling attention to how sparse team resources should be allocated and complemented, rather than to abstract notions of accuracy or efficiency. Only in this way can under- or mis-featurization risks be managed and mitigated without perverting intended stakeholders' semantic and moral commitments. The team must decide: what internal verification strategies might we need in order to safeguard the validations already endorsed by legal inquiry? Here "quality management" must be elevated to the contestation and adjudication of how (possibly pluralist) values are operationalized without compromising intuitive comparability.

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 [109]. This dimension serves as a decisive *ontic intervention* of what kind(s) of agency stakeholders possess as far as the system is concerned.

To preserve intuitive comparability at integration, the following challenges have to be taken up: (1) Formal: assess what kind(s) of agency all affected stakeholders have if the system fails, crystallized in the exit/voice dilemma; (2) Substantive: establish open feedback channels by which stakeholders express their values and concerns on their terms; (3) Discursive: in order to establish channels as trustworthy, justify these through ongoing public communication to stakeholders and updates to the design and/or governance of the system. Resolving these challenges requires representative input and mitigation of issues for the following:

Exit: Are stakeholders able to withdraw fully from using or participating in the system? Is there any risk in doing so? Are there competing products, platforms or systems they can use? Have assurances been given about user data, optimization, and certification after someone withdraws?

Voice: Can stakeholders articulate proposals in a way that makes certain concerns a matter of public interest? Are clear proposal channels provided for stakeholders, and are they given the opportunity to contribute regularly? Are the proposals highlighted frequently considered and tested, e.g. through system safety? Are stakeholders kept informed and regularly updated?

To the extent that proposed value hierarchies remain indeterminate beyond commitments supported through featurization and optimization, sociotechnical integration challenges systems to handle the multiple objectives, values, and priorities of diverse stakeholders. At stake here are the unexpressed moral relationships of subpopulations not originally considered part of the potential user base, who must bear the "cost function" of specification, as well as other forms of agency (animal, environmental, cybernetic) alien to yet implicated in system specification and creation. At a minimum, system administrators must determine whether the user will interpret the system agreement as primarily economic (in which case the user acts as a *consumer*) or political (in which case the user acts as a *citizen*). More Exit implies a market setting, while more Voice suggests a primarily political context.

For instance, the user agreements of Rekognition may be understood either in terms of private contracts (in which case data is treated as a commodity and alternative platforms are implied to exist) or public assurances (in which case data is inalienable and Rekognition is interpreted as a public utility or service). If the former takes precedence, Rekognition's agreements with contractors might be regulated with private certification that attests the features and uses of data it forbids; if the latter is more important, deployment increasingly depends on a public accreditation that guarantees legal protections will take priority in all use cases regardless of features or data. By comparison, Facebook's global reach and ambition to create an unprecedented single community of persons around the world is reflected in a vaguely-specified API and oversight mechanism, making it difficult to challenge its technical conflation of individual consent with scale of content sharing, which has bearing on user's own notions of legitimacy.

Deployment administrators and their regulating authorities must cultivate *public accountability* to deal with these challenges, ensuring both Voice and Exit remain possible for stakeholders such that some criterion of trustworthiness is maintained: anyone can leave if they want, but enough people choose to remain because they believe in their ability to express concerns as needed. Trustworthiness lies in maintaining potential stakeholders' belief in their ability to exert different kinds of agency as they see fit, either within the system (by dissenting to its current mode of operation) or outside it (by choosing it through active use). This sociotechnical balance must hold regardless of the specific commitment being made; for example, AWS may specify some channel by which vulnerable groups can opt out of a publicly-operated Rekognition use context (preserving Exit), or supply private contractors with a default user agreement that must be relayed to anyone whose data will be used by the system (preserving Voice). Either way, administrators must justify their decision to model people either as consumers (a customer, client, or operator treated more or less as a black box) or as citizens (a subject with guaranteed rights, among them the right to dissent to relevant forms of political power) in the context of the terms for system integration. Ultimately, the inability to have meaningful exit or voice can motivate collective action to aim to reshape power relationships [110], a phenomena that has recently manifested when pushing back against harmful AI systems [111].

7. Conclusion: Towards Technical Critical Practices

Our aims are strongly motivated by the classic work of Philip Agre on *critical technical practice*, 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 participation and decision-making authority of various forms of expertise and situated knowledge 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” [15]. As such, we prioritize and label the centering of stakeholder safety concerns and hard choices to guide and inform AI development as *technical critical practices.* We view this paper as a preliminary for what these practices might look like in particular development domains, and intend to 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 technical critical practices will need to be integrated into 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, as a pathway to system development, is necessary for such systems to safeguard the possibility of intuitive comparability and allow users to define the contours of their own values. To paraphrase Reinhold Niebuhr [112], AI's capacity for specification makes hard choices possible, but its inclination to misspecification makes them necessary.

Acknowledgements

We are grateful to participants in the NeurIPS 2019 Workshop on AI for Social Good for valuable discussions and suggestions. We also wish to thank Michael Dennis, Joan Greenbaum, Iason Gabriel, Rashida Richardson, Elizabeth Kaziunas, David Krueger, and Seda Gürses for their constructive feedback on previous versions of this paper. One of the authors is funded by the Center for Human-Compatible AI.

References

[1] J. McCarthy, M. L. Minsky, N. Rochester, C. E. Shannon, A proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955, AI magazine 27 (4) (2006) 12–12.

[2] M. Minsky, Steps toward artificial intelligence, Proceedings of the IRE 49 (1) (1961) 8–30.

[3] J. Weizenbaum, Computer power and human reason: From judgment to calculation.

[4] P. Krafft, M. Young, M. Katell, K. Huang, G. Bugingo, Defining ai in policy versus practice, in: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 2020, pp. 72–78.

[5] L. Suchman, Keynote address: Ai at the edgelands: Data analytics in states of in/security (2020).

[6] N. J. Goodall, Machine ethics and automated vehicles, in: Road vehicle automation, Springer, 2014, pp. 93–102.

[7] S. Russell, Provably beneficial artificial intelligence, Exponential Life, The Next Step.

[8] S. Russell, Human compatible: Artificial intelligence and the problem of control, Penguin, 2019.

[9] T. R. Allan, Dworkin and dicey: The rule of law as integrity, Oxford J. Legal Stud. 8 (1988) 266.

[10] L. B. Solum, On the indeterminacy crisis: Critiquing critical dogma, The University of Chicago Law Review 54 (2) (1987) 462–503.

[11] T. Richard, R. Tuck, Free riding, Harvard University Press, 2009.

[12] R. A. Stephans, System safety for the 21st century: The updated and revised edition of system safety 2000, John Wiley & Sons, 2012.

[13] M. S. Ackerman, The intellectual challenge of cscw: the gap between social requirements and technical feasibility, Human–Computer Interaction 15 (2-3) (2000) 179–203.

[14] B. Friedman, P. H. Kahn, A. Borning, Value sensitive design and information systems, The handbook of information and computer ethics (2008) 69–101.

[15] J. Greenbaum, M. Kyng, Design at work: Cooperative design of computer systems, L. Erlbaum Associates Inc., 1992.

[16] P. Agre, P. E. Agre, Computation and human experience, Cambridge University Press, 1997.

[17] H. L. Dreyfus, Skillful coping: Essays on the phenomenology of everyday perception and action, OUP Oxford, 2014.

[18] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, D. Mané, Concrete problems in ai safety, arXiv preprint arXiv:1606.06565.

[19] M. Fisher, A. Taub, How youtube radicalized brazil, The New York Times International Edition 3.

[20] L. Winner, Do artifacts have politics?, Daedalus (1980) 121–136.

[21] T. Williamson, Vagueness, Routledge, 2002.

[22] S. Schiffer, The epistemic theory of vagueness, Philosophical Perspectives 13 (1999) 481–503.

[23] M. Gómez-Torrente, Two problems for an epistemicist view of vagueness, Philosophical issues 8 (1997) 237–245.

[24] W. MacAskill, Practical ethics given moral uncertainty, Utilitas 31 (3) (2019) 231–245.

[25] N. Soares, B. Fallenstein, Aligning superintelligence with human interests: A technical research agenda, Machine Intelligence Research Institute (MIRI) technical report 8.

[26] N. Soares, The value learning problem, Machine Intelligence Research Institute, Berkley.

[27] W. MacAskill, Normative uncertainty as a voting problem, Mind 125 (500) (2016) 967–1004.

[28] J. Von Neumann, O. Morgenstern, Theory of games and economic behavior (commemorative edition), Princeton university press, 2007.

[29] M. Hildebrandt, Privacy as protection of the incomputable self: From agnostic to agonistic machine learning, Theoretical Inquiries in Law 20 (1) (2019) 83–121.

[30] D. Hadfield-Menell, G. K. Hadfield, Incomplete contracting and ai alignment, in: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, 2019, pp. 417–422.

[31] G. Irving, A. Askell, Ai safety needs social scientists, Distill 4 (2) (2019) e14.

[32] E. Barnes, J. R. G. Williams, A theory of metaphysical indeterminacy.

[33] W. MacAskill, The infectiousness of nihilism, Ethics 123 (3) (2013) 508–520.

[34] O. Keyes, Counting the countless: Why data science is a profound threat for queer people, Real Life 2.

[35] C. Mouffe, Deliberative democracy or agonistic pluralism?, Social research (1999) 745–758.

[36] K. Crawford, Can an algorithm be agonistic? ten scenes from life in calculated publics, Science, Technology, & Human Values 41 (1) (2016) 77–92.

[37] A. L. Hoffmann, Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse, Information, Communication & Society 22 (7) (2019) 900–915.

[38] V. Eubanks, Automating inequality: How high-tech tools profile, police, and punish the poor, St. Martin's Press, 2018.

[39] W. James, The will to believe: And other essays in popular philosophy, Longmans, Green, and Company, 1896.

[40] J. Dewey, Public & its problems.

[41] R. Benjamin, Race after technology: Abolitionist tools for the new jim code, Social Forces.

[42] B. Krais, Gender and symbolic violence: Female oppression in the light of pierre bourdieu's theory of social practice, Bourdieu: critical perspectives (1993) 156–177.

[43] M. Heidegger, J. Macquarrie, E. Robinson, Being and time.

[44] L. Wittgenstein, Philosophical Investigations, Basil Blackwell, Oxford, 1953.

[45] L. Lessig, Code: And other laws of cyberspace, ReadHowYouWant. com, 2009.

[46] G. Gerla, Comments on some theories of fuzzy computation, International Journal of General Systems 45 (4) (2016) 372–392.

[47] I. A. B. W. Group, Proceedings of the ieee algorithmic bias working group.

[48] S. Rea, A survey of fair and responsible machine learning and artificial intelligence: Implications of consumer financial services, Available at SSRN 3527034.

[49] S. Corbett-Davies, S. Goel, The measure and mismeasure of fairness: A critical review of fair machine learning, arXiv preprint arXiv:1808.00023.

[50] E. Trist, The evolution of socio-technical systems: A conceptual framework and an action research program, Ontario Ministry of Labour, 1981.

[51] P. Eckersley, Impossibility and uncertainty theorems in ai value alignment (or why your agi should not have a utility function), arXiv preprint arXiv:1901.00064.

[52] L. Evans, Traffic safety, 2004.

[53] L. Gu, R. Yang, C.-H. Tho, M. Makowskit, O. Faruquet, Y. L. Y. Li, Optimisation and robustness for crashworthiness of side impact, International Journal of Vehicle Design 26 (4) (2001) 348–360.

[54] J. C. Whitson, J. E. Ramirez-Marquez, Resiliency as a component importance measure in network reliability, Reliability Engineering & System Safety 94 (10) (2009) 1685–1693.

[55] J. A. Buolamwini, Gender shades: intersectional phenotypic and demographic evaluation of face datasets and gender classifiers, Ph.D. thesis, Massachusetts Institute of Technology (2017).

[56] J. Snow, Amazon's face recognition falsely matched 28 members of congress with mugshots, American Civil Liberties Union 28.

[57] M. Wood, Thoughts on machine learning accuracy, AWS News Blog.

[58] S. Romero, Wielding rocks and knives, arizonans attack self-driving cars, The New York Times 31.

[59] I. D. Raji, J. Buolamwini, Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products, in: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, 2019, pp. 429–435.

[60] D. Lee, San francisco is first us city to ban facial recognition, BBC.

[61] M. C. Elish, Moral crumple zones: Cautionary tales in human-robot interaction, Engaging Science, Technology, and Society 5 (2019) 40–60.

[62] L. Stark, Facial recognition is the plutonium of ai, XRDS: Crossroads, The ACM Magazine for Students 25 (3) (2019) 50–55.

[63] Y. Wang, M. Kosinski, Deep neural networks are more accurate than humans at detecting sexual orientation from facial images., Journal of personality and social psychology 114 (2) (2018) 246.

[64] H. Murphy, Why stanford researchers tried to create a 'gaydar'machine, The New York Times 9.

[65] L. F. Barrett, R. Adolphs, S. Marsella, A. M. Martinez, S. D. Pollak, Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements, Psychological science in the public interest 20 (1) (2019) 1–68.

[66] B. Resnick, This psychologist's "gaydar" research makes us uncomfortable. that's the point., Vox.

[67] K. Swisher, Zuckerberg: The Recode interview (Jul. 2018). URL https://www.vox.com/2018/7/18/17575156/mark-zuckerberg-interview-facebook-recode-kara-swisher

[68] M. Isaac, K. Roose, Facebook bars alex jones, louis farrakhan and others from its services, The New York Times.

[69] D. K. Mulligan, D. S. Griffin, Rescripting search to respect the right to truth.

[70] J. Naughton, Facebook's 'oversight board' is proof that it wants to be regulated – by itself, The Guardian.

[71] D. Ghosh, Facebook's oversight board is not enough, Harvard Business Review.

[72] B. C. Smith, The promise of artificial intelligence: reckoning and judgment, Mit Press, 2019.

[73] R. Chang, Incommensurability, incomparability, and practical reason, Harvard University Press, 1997.

[74] R. Chang, The possibility of parity, Ethics 112 (4) (2002) 659–688.

[75] R. Chang, Hard choices, Journal of the American Philosophical Association.

[76] B. D. Mittelstadt, P. Allo, M. Taddeo, S. Wachter, L. Floridi, The ethics of algorithms: Mapping the debate, Big Data & Society 3 (2) (2016) 2053951716679679.

[77] B. Mittelstadt, Principles alone cannot guarantee ethical ai, Nature Machine Intelligence (2019) 1–7.

[78] I. Van de Poel, Conflicting values in design for values, Handbook of ethics, values, and technological design: Sources, theory, values and application domains (2015) 89–116.

[79] J. Van den Hoven, G.-J. Lokhorst, I. Van de Poel, Engineering and the problem of moral overload, Science and engineering ethics 18 (1) (2012) 143–155.

[80] J. Halloran, E. Hornecker, M. Stringer, E. Harris, G. Fitzpatrick, The value of values: Resourcing co-design of ubiquitous computing, CoDesign 5 (4) (2009) 245–273.

[81] O. S. Iversen, K. Halskov, T. W. Leong, Rekindling values in participatory design, in: Proceedings of the 11th biennial participatory design conference, 2010, pp. 91–100.

[82] K. Shilton, Values and ethics in human-computer interaction, Foundations and Trends® in Human–Computer Interaction 12 (2).

[83] R. Dobbe, M. G. Ames, Translation Tutorial: Values, Engagement and Reflection in Automated Decision Systems (Jan. 2019).
URL https://medium.com/@roeldobbe/up-next-for-fat-from-ethical-values-to-ethical-practices-ebbed9f6adee

[84] L. C. Irani, M. S. Silberman, Stories we tell about labor: Turkopticon and the trouble with" design", in: Proceedings of the 2016 CHI conference on human factors in computing systems, 2016, pp. 4573–4586.

[85] D. Haraway, Situated knowledges: The science question in feminism and the privilege of partial perspective, Feminist studies 14 (3) (1988) 575–599.

[86] S. G. Harding, The science question in feminism, Cornell University Press, 1986.

[87] K. Bødker, F. Kensing, J. Simonsen, Participatory IT design: designing for business and workplace realities, MIT press, 2009.

[88] D. Nelkin, M. Pollak, Public-participation in technological decisions-reality or grand illusion, Technology Review 81 (8) (1979) 54–64.

[89] D. S. Mackay, What does mr. dewey mean by an" indeterminate situation"?, The Journal of Philosophy 39 (6) (1942) 141–148.

[90] S. Amrute, Of techno-ethics and techno-affects, Feminist Review 123 (1) (2019) 56–73.

[91] B. Friedman, H. Nissenbaum, Bias in computer systems, ACM Transactions on Information Systems (TOIS) 14 (3) (1996) 330–347.

[92] R. Dobbe, S. Dean, T. Gilbert, N. Kohli, A broader view on bias in automated decision-making: Reflecting on epistemology and dynamics, arXiv preprint arXiv:1807.00553.

[93] E. Anderson, The epistemology of democracy, Episteme: A journal of social epistemology 3 (1) (2006) 8–22.

[94] P. Agre, Toward a critical technical practice: Lessons learned in trying to reform ai in bowker, G., Star, S., Turner, W., and Gasser, L., eds, Social Science, Technical Systems and Cooperative Work: Beyond the Great Divide, Erlbaum.

[95] W. Blattner, What heidegger and dewey could learn from each other, Philosophical Topics 36 (1) (2008) 57–77.

[96] R. M. Unger, The critical legal studies movement, Harvard law review (1983) 561–675.

[97] L. Irani, J. Vertesi, P. Dourish, K. Philip, R. E. Grinter, Postcolonial computing: a lens on design and development, in: Proceedings of the SIGCHI conference on human factors in computing systems, 2010, pp. 1311–1320.

[98] J. Achiam, D. Held, A. Tamar, P. Abbeel, Constrained policy optimization, arXiv preprint arXiv:1705.10528.

[99] R. Choudhury, G. Swamy, D. Hadfield-Menell, A. D. Dragan, On the utility of model learning in hri, in: 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI), IEEE, 2019, pp. 317–325.

[100] L. Yu, T. Yu, C. Finn, S. Ermon, Meta-inverse reinforcement learning with probabilistic context variables, in: Advances in Neural Information Processing Systems, 2019, pp. 11772–11783.

[101] H. Dreyfus, S. D. Kelly, All things shining: Reading the Western classics to find meaning in a secular age, Simon and Schuster, 2011.

[102] E. P. Baumer, M. S. Silberman, When the implication is not to design (technology), in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2011, pp. 2271–2274.

[103] R. Guo, L. Cheng, J. Li, P. R. Hahn, H. Liu, A survey of learning causality with data: Problems and methods, arXiv preprint arXiv:1809.09337.

[104] K. J. Åström, R. M. Murray, Feedback systems: an introduction for scientists and engineers, Princeton university press, 2010.

[105] R. Parasuraman, V. Riley, Humans and automation: Use, misuse, disuse, abuse, Human factors 39 (2) (1997) 230–253.

[106] S. Barocas, A. D. Selbst, Big data's disparate impact, Calif. L. Rev. 104 (2016) 671.

[107] S. M. West, M. Whittaker, K. Crawford, Discriminating systems: Gender, race and power in ai, AI Now Institute (2019) 1–33.

[108] J. F. Fisac, A. K. Akametalu, M. N. Zeilinger, S. Kaynama, J. Gillula, C. J. Tomlin, A general safety framework for learning-based control in uncertain robotic systems, IEEE Transactions on Automatic Control 64 (7) (2018) 2737–2752.

[109] T. Flew, The citizen's voice: Albert hirschman's exit, voice and loyalty and its contribution to media citizenship debates, Media, Culture & Society 31 (6) (2009) 977–994.

[110] A. O. Hirschman, Exit, voice, and loyalty: Responses to decline in firms, organizations, and states, Vol. 25, Harvard university press, 1970.

[111] K. Crawford, R. Dobbe, T. Dryer, G. Fried, B. Green, E. Kaziunas, A. Kak, V. Mathur, E. McElroy, A. N. Sánchez, et al., Ai now 2019 report, New York, NY: AI Now Institute.

[112] R. Niebuhr, The essential Reinhold Niebuhr: Selected essays and addresses, Yale University Press, 1986.