AI Development for the Public Interest: From Abstraction Traps to Sociotechnical Risks

McKane Andrus*, Sarah Dean†, Thomas Krendl Gilbert‡, Nathan Lambert† and Tom Zick§
Authors arranged alphabetically. *Partnership on AI, San Francisco, CA.
†Department of Electrical Engineering and Computer Sciences, University of California, Berkeley.
‡Center for Human-Compatible AI, University of California, Berkeley.
§Berkman Klein Center for Internet and Society, Harvard University.
mckane@partnershiponai.org, {dean_sarah, tg340, nol}@berkeley.edu, tzick@cyber.harvard.edu

***Abstract*—Despite interest in communicating ethical problems and social contexts within the undergraduate curriculum to advance Public Interest Technology (PIT) goals, interventions at the graduate level remain largely unexplored. This may be due to the conflicting ways through which distinct Artificial Intelligence (AI) research tracks conceive of their interface with social contexts. In this paper we track the historical emergence of sociotechnical inquiry in three distinct subfields of AI research: AI Safety, Fair Machine Learning (Fair ML) and Human-In-the-Loop (HIL) Autonomy. We show that for each subfield, perceptions of PIT stem from the particular dangers faced by past integration of technical systems within a normative social order. We further interrogate how these histories dictate the response of each subfield to conceptual traps, as defined in the Science and Technology Studies literature. Finally, through a comparative analysis of these currently siloed fields, we present a roadmap for a unified approach to sociotechnical graduate pedogogy in AI.**

I. Introduction

Recent years have seen an increasing public awareness of the profound implications of widespread artificial intelligence (AI) and large scale data collection. It is now common for both large tech companies and academic researchers to motivate their work on AI as interfacing with the "public interest," matching external scrutiny with new technical approaches to making systems fair, secure, or provably beneficial. However, developing systems in the public interest requires researchers and designers to confront what has been elsewhere referred to as the "sociotechnical gap," or the divide between the intended social outcomes of a system and what is actually achieved through technical methods [1].

Interventions in Computer Science (CS) education have made strides towards providing students with frameworks within which to evaluate technical systems in social contexts [2], [3]. These curricular modifications have drawn on fields like Law, Philosophy, and Science and Technology Studies (STS) to create both dedicated and integrated coursework promoting human contexts and ethics in CS [4]. However, as the majority of these courses are currently offered at the undergraduate level, graduate students may not reap the benefits of such reforms [4]. Given the role of graduate students as not only teachers, but drivers of cutting edge research and future decision makers in industry and academia, interventions aimed at them may play an outsized role in forwarding PIT goals.

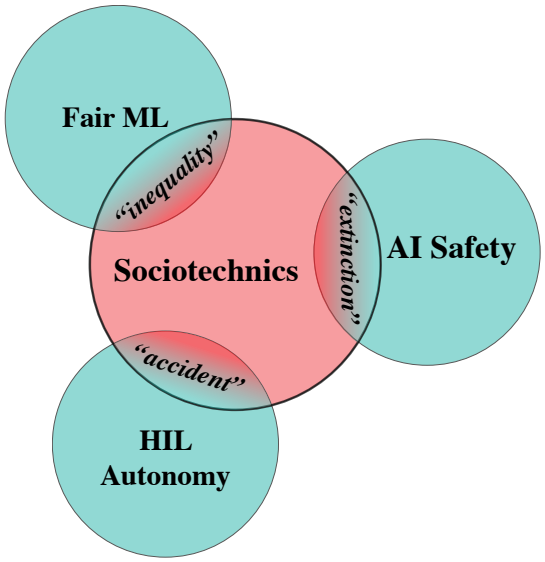

Fig. 1. Three contemporary areas of artificial intelligence research whose overlapping forms of sociotechnical inquiry remain problematically defined: AI Safety, Fair Machine Learning, and Human-in-the-Loop Autonomy.

It is challenging to pin down what it would mean to train a graduate AI researcher to address the sociotechnical gap. A key source of tension is the place of the sociotechnical in AI development: while practitioners claim to be working on technical solutions to social problems, theoretical and methodological formulations of the sociotechnical are inconsistent across prominent AI subfields, making it unclear if current initiatives in pedagogy and research are advancing, undermining, or neglecting the public interest.

In this paper, we go beyond coursework to analyze the historical and technical shifts behind the current conception of sociotechincal risks in prominent AI subfields. We look to existing research domains that grapple with the social and technical spheres at distinct levels of abstraction, and examine how their limitations and insights reflect nascent, if problematic, *forms of inquiry* into the sociotechnical. We assess current research in the socially-oriented subfields of AI highlighted in Fig. 1, namely AI Safety; Fairness in Machine Learning (Fair ML); and Human-in-the-Loop (HIL) Autonomy. AI Safety

focuses on value-alignment of future systems and cautions against developing AI systems fully integrated with and in control of society. Fair ML works to reduce bias in algorithms with potentially deleterious effects on individuals or groups. HIL Autonomy is a term we use to encompass emerging work on both human-robot interaction and cyber-physical systems. These research areas explore how to optimize interactions of autonomous systems with human intentions in the loop. By tracing the history of these subfields in a comparative fashion, we are able to characterize their distinct orientations towards sociotechnical challenges, highlighting both insights and blindspots. The goal of this analysis is not to capture each subfield's research agenda exhaustively, which is far beyond our scope. Instead, it is to highlight how these agendas claim relative access to some feature of the sociotechnical in the way they represent social problems as technically tractable. In doing so, they claim legitimacy and authority with respect to public problems.

We next refine this comparative history with lenses borrowed from the Science and Technology Studies (STS) literature, emphasizing the ways the subfields interface with sociotechnical risks. In particular, we portray how certain risks are deferred within each subfield's agenda. This deferral often takes the form of skillfully avoiding "abstraction traps" that have been recently highlighted by [5]. Such avoidance is important from a technical standpoint. However, true engagement with the sociotechnical requires reflexively revealing and resolving risks beyond the piecemeal formalisms that have defined each subfield's historical trajectory. We conclude with a brief sketch of pedagogical interventions towards this goal. Beyond classroom ethics curricula, we propose an agenda for clinical engagement with problems in the public interest as a part of graduate training in AI. This training would inculcate a direct appreciation of sociotechnical inquiry in parallel with the acquisition of specific technical skillsets. It would prepare practitioners to evaluate their own toolkits discursively, rather than just mathematically or computationally.

II. Emerging Sociotechnical Subfields of AI: Safety, Fairness, Resiliency

Recent technical work grappling with the societal implications of AI includes developing provably safe and beneficial artificial intelligence (AI Safety), mitigating classification harms for vulnerable populations through fair machine learning (Fair ML), and designing resilient autonomy in robotics and cyber-physical systems (HIL Autonomy). At present, these areas constitute heterogeneous technical subfields with substantive overlaps, but lack discursive engagement and cross-pollination.

Below, we outline the history and motivating concerns of each subfield and identify key developments and convenings. We highlight how the technical research agendas stem from distinctly sociotechnical concerns and will require interdisciplinary engagement to fully map out their stakes. By placing these subfields' agendas in the context of the local sociotechnical risks (respectively *extinction*, *inequity*, and *accident*), we argue that the representative technical formations (*forecasting* and *value alignment*, *fairness criteria* and *accountability*, *controllability* and *reachability*) iterate on those risks without reflexively interrogating or normatively addressing them. Throughout this work we use the term AI system to refer to a technical system with significant automated components (e.g. automated decision making systems or self driving cars). We note that AI is also a distinct field of academic study, further discussed in this section.

A. AI Safety

The field of artificial intelligence (AI) has often situated itself within a wider disciplinary context. Famously, at the foundational summits at Dartmouth and MIT, computer scientists, logicians, and psychologists came together to chart a course for artificial intelligence to arrive at human-like capabilities [6]. That course, which laid the groundwork for "good-old-fashioned" AI, had an incredible number of hiccups and was eventually overwritten. A similarly diverse group of disciplinary representatives moved the field away from symbolic and comprehensive logical reasoning to either more situated, interactionist understandings of cognition [7], [8] or to biologically-inspired, connectionist strategies of learning [9]. Following these interdisciplinary developments, there has been a growing concern about the capabilities of AI systems to endanger humans and society writ large [10], [11], [12]. Motivated both by longstanding concerns about the possibility of a "technological singularity" [13] as well as recent expansive applications of machine learning in critical infrastructure domains, many AI Safety promoters fear that AI researchers are approaching a level of capability that will expand beyond their control [13], [14].

Belief in the prospect of an arbitrarily capable intelligent agent beyond designer control has raised the prospect of *extinction*, whether of humanity or all natural life, as a clear and present danger for AI development, serving as this field's distinct sociotechnical risk scenario. Regardless of the likelihood of such a scenario, the nascent field of AI Safety has arisen to preemptively confront these dangers. AI Safety takes a rather radical approach to the type of systems level thinking that we discuss, often viewing technical developments on a much longer and wider timescale—see, for example, the published work on AI arms races and potential research agendas to avert them [12], [15]. As such, a common feature of AI Safety research is *forecasting* future AI capabilities against various time horizons.

Despite recent high-profile endorsements from computer scientists and philosophers such as Stuart Russell [16] and Nick Bostrom [14], AI Safety is still a nascent research community. At present there is no independent conference for this field, although workshops and panels on AI Safety have become a regular fixture of larger AI venues such as NeurIPS, ICML, and AAAI, while specific AI-safety oriented research labs (e.g. CHAI, OpenAI) host invited technical presentations on a semiweekly or monthly basis. The field has also attracted interest from research centers and philanthropic organizations

dedicated to the study and mitigation of long-term existential risk, as well as industry leaders in AI.

Central to this work is a motivation to align AI systems with intended rather than specified rewards, as humans struggle to make explicit the rich normative context of their own goals and behaviors. Through this shift, AI Safety adjusts the framing of classical AI development towards "provably beneficial" rather than merely optimal systems. Under this framing, researchers focus largely on the problem of *value alignment*, i.e. whether or not an AI agent's programmed objective matches those of relevant humans or humanity as a whole [17]. For example, by understanding the problem of aligning an AI agent with a human collaborator as a problem of inverse reinforcement learning, researchers seek to solve this issue with a largely technical approach by borrowing core principles from economic game theory [18].

Considered as a whole, extended sociotechnical inquiry in AI Safety remains limited to catastrophic risk evaluation in cases where humanity's survival is at stake—a scale of concern that is not often found in engineering disciplines. Moreover, rigorous formal work often relies on intuition from mechanism design (e.g. an objectives-first approach, perfectly rational agents) whose assumptions inherit some of the formal limitations of and controversies surrounding prospect theory and social choice theory. Stemming from AI Safety, we see vigorous discussions surrounding AI Policy [19], ethics [20], and even reflexive interrogations as a practice in forecasting [21]. While lacking some qualities of sociotechnical inquiry, in particular a deeply reflexive methodology and historical orientation, we see potential to pivot these discussions away from narrowly-framed thought experiments about paperclip-maximizing robots [11] towards comparative investigations of the normative stakes of distinct AI-society interfaces.

B. Fairness in Machine Learning

The field of machine learning (ML) emerged in the late 1950s with the design of a self-improving program for playing checkers [22] and quickly found success with static tasks in pattern classification, including applications like handwriting recognition [23]. ML techniques work by detecting and exploiting statistical correlations in data, towards increasing some measure of performance. A prominent early machine learning algorithm was the perceptron [24], an example of supervised classification, perhaps the most prevalent form of ML. In this setting, a classifier (or model) is trained with labelled examples, and its performance is measured by its accuracy in labelling new instances. The perceptron spurred the development of deep learning techniques mid-century [25]; however, they soon fell out of favor, only having great success in recent decades in the form of neural networks via the increasing availability of computation and data. Many ML algorithms require large datasets for good performance, tying the field closely with "big data." However, optimizing predictive accuracy does not generally ensure beneficial outcomes when predictions are used to make decisions, a problem that becomes stark when individuals are harmed by the classification of an ML system.

The *inequality* resulting from system classifications is the central sociotechnical risk of concern to practitioners in this the subfield of Fair ML. A growing awareness of the possibility for bias in data-driven systems developed over the past fifteen years, starting in the data mining community [26] and echoing older concerns of bias in computer systems [27]. The resulting interest in ensuring "fairness" was further catalyzed by high profile civil society investigation (e.g. ProPublica's Machine Bias study, which highlighted racial inequalities in the use of ML in pretrial detention) and legal arguments that such systems could violate anti-discrimination law [28]. At the same time, researchers began to investigate model "explainability" in light of procedural concerns around the black box nature of deep neural networks. The research community around Fairness in ML began to crystallize with the ICML workshop on Fairness, Accountability, and Transparency in ML (FAT/ML), and has since grown into the ACM conference on Fairness, Accountability, and Transparency (FAccT) established in 2017.

By shifting the focus to fairness properties of learned models, Fair ML adjusts the framing of the ML pipeline away from a single metric of performance. There are broadly two approaches: individual fairness, which is concerned with similar people receiving similar treatment [29], and group fairness which focuses on group parity in acceptance or error rates [30]. The details of defining and choosing among these *fairness criteria* amount to normative judgements about which biases must be mitigated, with some criteria being impossible to satisfy simultaneously. Much technical work in this area focuses on algorithmic methods for achieving fairness criteria through either pre-processing on the input data [31], in-processing on the model parameters during training [32], or post-processing on model outputs [33].

The Fair ML community is oriented towards the sociotechnical, engaging actively with critiques from STS perspectives. FAccT is a strong locus of interdisciplinary thought within computer science, and the addition of *transparency* and *accountability* to the title opens the door to a wider range of interventions. Building upon model-focused concepts like explainability, blendings of technical and legal concepts of recourse [34] and contestability [35] widen the frame to explicitly consider the reaction of individuals to their classification. Similarly, there have been multiple calls to re-center stakeholders to understand how explanations are interpreted and if they are even serving their intended purpose [36], [37]. The community is increasingly open to discussing scenarios in which technical intervention, like the police use of facial recognition, is not desired. This encompasses both technical resistance [38] and procedural approaches to delineating the valid uses of data [39] and models [40].

C. Human-in-the-Loop Autonomy

As many of the earliest robotic systems were remotely operated by technicians, the field of robotics has always had prob-

lems of human-robot interaction (HRI) at its core [41]. Early work was closely related to the study of human factors, an interdisciplinary endeavor drawing on engineering psychology, ergonomics, and accident analysis [42]. With advancements in robotic capabilities and increasing autonomy, the interaction paradigm grew beyond just teleoperation to *supervisory control*. HRI emerged as a distinct multidisciplinary field in the 1990s with the establishment of the IEEE International Symposium on Robot & Human Interactive Communication. Modern work in this area includes modeling interaction from the perspective of the autonomous agent (i.e. robot) rather than just the human overseer. By incorporating principles from the social sciences and cognitive psychology, HRI uses predictions and models of human behavior to optimize and plan. This work mitigates the sociotechnical risk of *accidents* – defined specifically as states in which physical difficulties or mishaps occur. Such physical risks are mitigated by making models robust to these potentially-dangerous conditions.

Digital technology has advanced to the point that many systems are endowed with autonomy beyond the traditional notion of a robotic agent, including traffic signal networks at the power grid. We thus consider the subfield of *HIL Autonomy* to be the cutting edge research that incorporates human behaviors into robotics and cyber-physical systems. This subfield proceeds in two directions: 1) innovations in physical interactions via sensing and behavior prediction; 2) designing for system resiliency in the context of complicated or unstable environments. These boundaries are blurring in the face of increasingly computational methods and the prospective market penetration of new technologies. For example, the design of automated vehicles (AVs) poses challenges along many fronts. For more fluent and adaptable behaviors like merging, algorithmic HRI attempts to formalize models for one-on-one interactions. At the same time, AVs pose the risk physical harm, so further lines of work integrate these human models to ensure safety despite the possibility of difficult-to-predict actions. Finally, population-level effects (e.g. AV routing on traffic throughput and induced demand) require deeper investigation into interaction with the social layer.

The emerging subfield of HIL Autonomy uses ideas from classical *control theory* while trying to quantify and capture the risk and uncertainty of working with humans [43], [44]. It thus inherits some of the culture around verifying safety and robustness through a combination of mathematical tools and physical redundancy, due to a history of safety-critical applications in domains like aerospace. Technical work in this area typically entails including the human as part of an under-actuated dynamical system [45], [46], such as a un-modeled disturbance. Through this lens, human-induced uncertainty is mitigated by predicting behavior in a structured manner, maintaining the safety of the system through mathematical robustness guarantees [47]. To make this concrete, a lane-change maneuver in an AV might include both an aggressive driving plan that takes likely human behaviors into account as well as a *reachability* safety criterion which could be activated via feedback if observed human behavior falls outside of the expected distribution. At a higher level of planning, the lane change maneuver may only be directed if it is expected to be advantageous for global traffic patterns.

The extent to which HIL Autonomy engages with the sociotechnical is thus far limited. Human-centered research focuses on localized one-to-one interactions, while research considering more global interactions remains largely in the realm of the technical. However, the critical "alt.HRI" track at the ACM/IEEE International Conference on Human-Robot Interaction indicates an emerging interest in how robotic systems interact with society more broadly. In such venues, questions are raised surrounding how robots interact with social constructions of race [48], [49] and issues of robot-community integration are being studied in settings ranging from healthcare [50] to gardening [51]. There is also work which considers the incorporation of social values into cyber-physical systems, e.g. fair electricity pricing for smart grids [52]. While our identification of this emerging subfield is perhaps more speculative than the previous two, the physical realization of AI technologies will remain a crucial site of sociotechnical inquiry.

III. Sociotechnical Integration

While the subfields of AI Safety, Fair ML, and HIL Autonomy each consider problems at the interface of technology and human or social factors, there are differences which arise in part to their disparate histories. One difference is in time-scales. AI Safety is primarily concerned with long term outcomes of mis-aligned AI development, while Fair ML focuses on practical implementations of individual models and algorithms with imperfect datasets. HIL Autonomy bisects the two, with both longer term considerations of how numerous autonomous agents will re-define how humans interact in the environment and short term focus on maintaining safety, e.g. in the presence of unexpected adverse road conditions. Another difference arises from how the subfields position themselves at different levels of abstraction. HIL Autonomy is physically grounded, with a history closely tied with embodied interaction with humans and the social layer, while Fair ML is socially grounded, and has strong instincts for sociotechnical dialog and historical situatedness. On the other hand, AI Safety positions itself at the highest level of generality, relegating machine learning to the status of a tool and interpreting robotics as an application of formal guarantees.

For these subfields to place their sociotechnical inquiry on firmer foundations, it will be necessary to establish more reflexive relationships with their inherited assumptions about risk. Each subfield interprets itself as filling well-defined sociotechnical gaps, i.e. that there is a discernible divide between social problems and technical agendas. But in fact, the way these subfields have defined and worked on those gaps is itself problematic, piecemeal, and lacking in definition, i.e. it is normatively indeterminate. Reflexive inquiry is needed not to fill those gaps, but to define and interpret them more richly, so that their salience and urgency can be evaluated.

At a minimum, researchers and practitioners must learn to see behind their own technical abstractions to the social reality they assume, recognize that this reality may have been problematically defined, and learn to inquire into these definitions directly, perhaps with the aid of new transdisciplinary tools. We now provide a high-level summary of this agenda, moving from a comparison of common *technical traps* to more indeterminate conceptions of sociotechnical risk.

A. Grappling with Shortcomings in Framing

Each of the subfields discussed in the previous section seeks to expand the technical framing of their parent field to include human and social factors. In [5], the *framing trap* is introduced as the failure to model the full system of interest (e.g. with respect to a notion of fairness or safety). Technical researchers are at risk of falling into this trap whenever they draw a *bounding box* around the system that they study. Often, the consequences of this trap manifest as the *portability trap* [5], which occurs when technical solutions designed for one domain or environment are misapplied to another context. Technical researchers are at risk of falling into this trap whenever they mistakenly view a bounding box as appropriate to a new context.

The subfields of AI Safety, Fair ML, and HIL Autonomy can be viewed as attempts to avoid the framing trap. In the fields of AI, ML, and robotics, the workflow often entails *featurization* by defining data or inputs/outputs, *optimization* by fitting a model or designing a control policy, and then *integration* into the larger system. Researchers in the emerging sub-fields are beginning to understand the downsides of this unidirectional workflow, and the necessity of interrogating the modelling choices made at each step. For example, AI Safety questions the way that features are used to define optimization objectives in light of potentially catastrophic effects of integration, while Fair ML questions the inequalities arising from model optimization.

Still, sometimes the frame is not opened wide enough. For example, by failing to account for the larger system in which risk assessments are used, approaches to Fair ML may mistakenly treat loaning decisions the same way they treat pretrial detention, despite salient differences between the financial and criminal justice systems. By adopting more rigorously a *heterogeneous engineering* approach [5], researchers and practitioners can explicitly determine which properties are not tied to the technical objects under design but to their social contexts. For example, the aerospace industry is an engineering domain with considerable heterogeneity—an awareness of the regulatory context, from the flight deck procedures to air traffic control, is necessary for the development of flight technologies.

B. Abstraction Traps in AI Research

To motivate a stronger cross-disciplinary discourse among and outside of these subfields, we now make further use of the framework of abstraction traps provided by [5] to point systematically to shortcomings and highlight potential new areas of inquiry. Alongside the framing and portability traps, we discuss: the formalism trap, the ripple effect trap, and the solutionism trap.

The *formalism trap* occurs when mathematical formalisms fail to capture important parts of the human context. For example, the fairness of a system is often judged by procedural rather than technical elements, and the perceived reliability may depend more on predictability rather than formally verified safety. All of the discussed subfields are posed to fall into the formalism trap, which requires a deeper engagement with sociotechnical complexities to avoid. Ultimately, the validity and desirability of specific metrics arising from mathematical abstractions will be determined through intimate reference to social context rather than technical parsimony. If systems are not flexible enough to allow for public input, the validity can be compromised.

The *ripple effect trap* occurs when there is a failure to understand how technology affects the social system into which it is inserted. AI Safety considers ripple effects to some extent, but in a narrowly formal manner. For example, while automated vehicles are known to affect traffic, road, and even infrastructure design, most technical research has focused on incorporating these as features to be modeled rather than questioning the status of AVs as the dominant form of future mobility. Engagement across the entire sociotechnical stack requires understanding social phenomena like the "reinforcement politics" of dominant groups using technology to remain in power and "reactivity" like gaming and adversarial behavior. If a system encourages people to behave in an adversarial manner, it may call for utilizing richer design principles to promote cooperation, rather than merely throwing more advanced AI methods at the assumed dynamics.

Finally, the *solutionism trap* occurs when designers mistakenly believe that technical solutions alone can solve complex sociological and political problems. For example, while the legal community has encouraged technical fields to build systems that are reliably safe and fair, these interventions must be specified in terms of norms that can be appropriately internalized by practitioners. The General Data Protection Regulation has had a mixed reception—while it did articulate normative landmarks for subfields to pay attention to, some of its requirements (e.g. consent as a legal basis for data processing) were highly underspecified. This specification vacuum empowered prominent private actors to advance their own standards in a way that is ethically questionable but politically effective, achieving market buy-in from enough other actors before the law can catch up [53], [54]. Technical practitioners will need the ability to stand up and contest would-be standards publicly, rather than relying on the law to interpret systems before their sociotechnical scope has been appropriately modeled. To avoid the solutionism trap, it is important to maintain a robust culture of questioning which problems should be addressed, and why these problems and not others: in the form of humility or a "first, do no harm" perspective.

C. From Avoiding Traps to Anticipating Risks

An important initial step for grappling with abstraction traps is for technical practitioners in fields of AI Safety, Fair ML, and HIL Autonomy to consider them explicitly when attempting to solve and formulate problems. In following with [5], we find it most helpful to consider the traps in reverse order: is it worth designing a technical solution? Can we adequately reason about how the technology will affect its social context? Can the desired properties of the system be captured by mathematical abstractions? Are the technical tools appropriate to the context? And are all relevant actors included in the framing? By considering these questions, researchers and practitioners will be encouraged to grapple with the plural temporalities defined by *ongoing sociotechnical engagement* through the validation of assumptions behind featurization, optimization, and integration.

While researchers in AI Safety, Fair ML, and HIL Autonomy are well positioned to begin asking these questions, it is only a first step. There is an inherent vulnerability in applying computational decision heuristics to vital social domains. Autonomous AI systems introduce possibilities of catastrophic failure and normative incommensurability to contexts that were previously accessible only to human judgment and which we may never be able to exhaustively specify or completely understand. Beyond the mere *avoidance of conceptual traps*, practitioners must learn to *anticipate sociotechnical risks* as integral to the endeavor of building AI systems that interface with social reality.

The distinct intuitive approaches to risk taken by each of the examined subfields (*extinction*, *inequality*, and *accident*) stem from alternative histories of the sorts of dangers faced when integrating systems within a normative social order. In other words, while these research communities have adopted tools and mathematical formalisms that purport to represent and work on discrete social phenomena, in fact the tools themselves are sociotechnical interventions, and their elaboration is justified according to historically-sedimented perceptions of risk. Rather than systems that represent and affect specific social objects (e.g. people, institutions), we advocate for the concept of AI as a process of elaborating normative commitments whose technical refinement generates unprecedented positions [55]. From these positions, novel sociotechnical questions can be revealed, resolved, or deferred.

IV. Towards Clinical Training for Graduate Pedagogy

How can researchers and practitioners learn to anticipate sociotechnical risks? Awareness of abstraction traps may corroborate an appreciation of risks, but it does not provide the tools with which to anticipate or understand them. For example, pedagogical reforms based on coursework drawn from Science and Technology Studies, Philosophy, and Law may inspire a requisite caution in technical practitioners. However, this caution remains insufficient to define the problem space of appropriate uses of AI. Instead, it will be necessary to encourage the coordination of technical and social scientists on these matters. In what follows, we interrogate this by evaluating possible reforms in graduate pedagogy.

Graduate students are a fruitful site for intervention for three primary reasons: 1) their educational role in shaping the next generation of engineers, 2) their role in pushing forward emerging areas of research and 3) their future as management and decision-makers at technology companies. Students should have the ability to recognize when a single development pipeline is trying to engage in multiple abstractions simultaneously because its metaphors are confused (e.g. the fact that certain AI Safety formalisms [18], understood in terms of a principal-agent game, can function both as a form of mechanism design and as a kind of interface between user and robot). It is further important that they have the ability to contest, merge, or even dissolve these frames if necessary. This will entail a major cultural transition in how the goals of graduate training are defined, moving away from failure-avoidance engineering in controlled environments to the responsible integration of technology in human contexts.

While there are efforts to widen the scope of a technical education and augment it with political and ethical training [56], [57], a truly sociotechnical graduate education would teach the skills of how to draw a technical bounding box as well as how to communicate those decisions to the publics that will have to reckon with the potential benefits and harms of new technology. Education cannot carve up the world into specific problem domains, but it could help coordinate concerns in a constructive manner that enables the development of context-appropriate validation metrics, as others have begun to do by synthesizing common technical pitfalls [5].

While coursework lays the foundations for research, it cannot provide a descriptive ontology that would exhaustively capture sociotechnical risks in advance of active inquiry. Anticipating and mitigating such risks requires an immersion in the relevant social context, becoming richly familiar with its phenomenology from the human standpoint. Only by doing this is it possible to register the system specification in terms of the concrete normative stakes rather than abstract approximations of optimal behavior. This entails an ontological shift away from a purely mechanistic description of the domain in favor of a clinician's perspective, comparable in scope and significance to the emergence of modern medical and legal clinics [58], [59], [60]. We believe a distinctly clinical approach to social problems—engaged and prolonged consultation, direct provision of service, relationships with clients, hands-on education overseen by professors—is the best approach.

Technical work will always rely on abstraction and framing to describe the environment in which a system is designed to function. It falls on technical researchers and practitioners to understand how to specify such a *bounding box*, decide which frames and abstractions are valid and tractable as well as commensurate with stakeholder concerns, and articulate their choices to relevant communities with varying technical backgrounds. We see this "clinician's eye," entailing effective framing and communication, as the most promising potential

outcome of reforming AI pedagogy at the graduate level, and defer further investigation of clinical approaches in the context of CS education to future work.

V. Conclusion

The work of defining "sociotechnical" problems in AI development is ongoing. Systems themselves often make symbolic reference to situations, environments, or objects that are assumed to lie behind their representations unreflectively, allowing the same mathematical structures to propagate without interrogating key metaphorical frames. This norm results in a practice incommensurate with other expert professions' standards of liability. Along with the inconsistency between subfields, this makes it hard to define what constitutes an AI expert and how responsibility should be assigned when systems fail. Looking from the outside in, the legal and philosophical communities cannot enforce standards that are neither backed up by established forms of expertise nor immediately translatable outside the context of technical-mathematical formalism, meaning case law and abstract ethics cannot fully determine or guide sociotechnical regulation.

Given this normative indeterminacy, we argue there is no ready-made delineation of which technical tools are suited to which social problems, and instead look to prospective interventions nurturing new forms of inquiry into inherited notions of risk. On this view, interventions would embrace the notion that elaborating on sociotechnical problems and procedures is essential to the task itself, and practitioners would understand the sociotechnical simply as part of what they do. We argue this is the more sure path to effective norms for distinct subfields of AI development, and thus to the aims of Public Interest Technology.

References

[1] M. S. Ackerman, "The Intellectual Challenge of CSCW: The Gap Between Social Requirements and Technical Feasibility," *Human–Computer Interaction*, vol. 15, no. 2-3, pp. 179–203, Sep. 2000. [Online]. Available: https://www.tandfonline.com/doi/full/10.1207/S15327051HCI1523_5

[2] J. Saltz, M. Skirpan, C. Fiesler, M. Gorelick, T. Yeh, R. Heckman, N. Dewar, and N. Beard, "Integrating ethics within machine learning courses," *ACM Trans. Comput. Educ.*, vol. 19, no. 4, Aug. 2019. [Online]. Available: https://doi.org/10.1145/3341164

[3] M. Skirpan, N. Beard, S. Bhaduri, C. Fiesler, and T. Yeh, "Ethics education in context: A case study of novel ethics activities for the cs classroom," in *Proceedings of the 49th ACM Technical Symposium on Computer Science Education*, ser. SIGCSE '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 940–945. [Online]. Available: https://doi.org/10.1145/3159450.3159573

[4] C. Fiesler, N. Garrett, and N. Beard, "What do we teach when we teach tech ethics? a syllabi analysis," in *Proceedings of the 51st ACM Technical Symposium on Computer Science Education*, ser. SIGCSE '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 289–295. [Online]. Available: https://doi.org/10.1145/3328778.3366825

[5] A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian, and J. Vertesi, "Fairness and abstraction in sociotechnical systems," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019, pp. 59–68.

[6] S. Russell and P. Norvig, "Artificial intelligence: a modern approach," 2002.

[7] H. L. Dreyfus, L. Hubert *et al.*, *What computers still can't do: A critique of artificial reason*. MIT press, 1992.

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

[9] R. Sun, "Connectionism and neural networks," *The Cambridge handbook of artificial intelligence*, p. 108, 2014.

[10] N. Bostrom, "Ethical issues in advanced artificial intelligence," *Science fiction and philosophy: from time travel to superintelligence*, pp. 277–284, 2003.

[11] E. Yudkowsky *et al.*, "Artificial intelligence as a positive and negative factor in global risk," *Global catastrophic risks*, vol. 1, no. 303, p. 184, 2008.

[12] S. Armstrong, N. Bostrom, and C. Shulman, "Racing to the precipice: a model of artificial intelligence development," *AI & society*, vol. 31, no. 2, pp. 201–206, 2016.

[13] R. Kurzweil, *The singularity is near: When humans transcend biology*. Penguin, 2005.

[14] N. Bostrom, *Superintelligence*. Dunod, 2017.

[15] A. Ramamoorthy and R. Yampolskiy, "Beyond mad? the race for artificial general intelligence."

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

[17] N. Soares and B. Fallenstein, "Agent foundations for aligning machine intelligence with human interests: a technical research agenda," in *The Technological Singularity*. Springer, 2017, pp. 103–125.

[18] D. Hadfield-Menell, S. J. Russell, P. Abbeel, and A. Dragan, "Cooperative inverse reinforcement learning," in *Advances in neural information processing systems*, 2016, pp. 3909–3917.

[19] M. Brundage, S. Avin, J. Wang, H. Belfield, G. Krueger, G. Hadfield, H. Khlaaf, J. Yang, H. Toner, R. Fong *et al.*, "Toward trustworthy ai development: mechanisms for supporting verifiable claims," *arXiv preprint arXiv:2004.07213*, 2020.

[20] I. Gabriel, "Artificial intelligence, values and alignment," *arXiv preprint arXiv:2001.09768*, 2020.

[21] K. Grace, J. Salvatier, A. Dafoe, B. Zhang, and O. Evans, "When will ai exceed human performance? evidence from ai experts," *Journal of Artificial Intelligence Research*, vol. 62, pp. 729–754, 2018.

[22] A. L. Samuel, "Some studies in machine learning using the game of checkers," *IBM Journal of research and development*, vol. 3, no. 3, pp. 210–229, 1959.

[23] N. JNILSSON, "Learning machines: Foundations of trainable pattern classifying systems," 1965.

[24] F. Rosenblatt, *The perceptron, a perceiving and recognizing automaton Project Para*. Cornell Aeronautical Laboratory, 1957.

[25] M. Olazaran, "A sociological study of the official history of the perceptrons controversy," *Social Studies of Science*, vol. 26, no. 3, pp. 611–659, 1996.

[26] D. Pedreshi, S. Ruggieri, and F. Turini, "Discrimination-aware data mining," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2008, pp. 560–568.

[27] B. Friedman and H. Nissenbaum, "Bias in computer systems," *ACM Transactions on Information Systems (TOIS)*, vol. 14, no. 3, pp. 330–347, 1996.

[28] S. Barocas and A. D. Selbst, "Big data's disparate impact," *Calif. L. Rev.*, vol. 104, p. 671, 2016.

[29] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proceedings of the 3rd innovations in theoretical computer science conference*. ACM, 2012, pp. 214–226.

[30] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and Machine Learning*. fairmlbook.org, 2018, http://www.fairmlbook.org.

[31] F. Calmon, D. Wei, B. Vinzamuri, K. N. Ramamurthy, and K. R. Varshney, "Optimized pre-processing for discrimination prevention," in *Advances in Neural Information Processing Systems*, 2017, pp. 3992–4001.

[32] M. B. Zafar, I. Valera, M. Gomez-Rodriguez, and K. P. Gummadi, "Fairness constraints: A flexible approach for fair classification." *J. Mach. Learn. Res.*, vol. 20, no. 75, pp. 1–42, 2019.

[33] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Advances in neural information processing systems*, 2016, pp. 3315–3323.

[34] B. Ustun, A. Spangher, and Y. Liu, "Actionable recourse in linear classification," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019, pp. 10–19.

[35] D. K. Mulligan, D. Kluttz, and N. Kohli, "Shaping our tools: Contestability as a means to promote responsible algorithmic decision making in the professions," *Available at SSRN 3311894*, 2019.

[36] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial Intelligence*, vol. 267, pp. 1–38, 2019.

[37] U. Bhatt, M. Andrus, A. Weller, and A. Xiang, "Machine learning explainability for external stakeholders," *arXiv preprint arXiv:2007.05408*, 2020.

[38] B. Kulynych, R. Overdorf, C. Troncoso, and S. Gürses, "Pots: protective optimization technologies," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020, pp. 177–188.

[39] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. Daumé III, and K. Crawford, "Datasheets for datasets," *arXiv preprint arXiv:1803.09010*, 2018.

[40] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru, "Model cards for model reporting," in *Proceedings of the conference on fairness, accountability, and transparency*, 2019, pp. 220–229.

[41] M. A. Goodrich and A. C. Schultz, *Human-robot interaction: a survey*. Now Publishers Inc, 2008.

[42] L. Bainbridge, "Ironies of automation," in *Analysis, design and evaluation of man–machine systems*. Elsevier, 1983, pp. 129–135.

[43] R. Baheti and H. Gill, "Cyber-physical systems," *The impact of control technology*, vol. 12, no. 1, pp. 161–166, 2011.

[44] A. Banerjee, K. K. Venkatasubramanian, T. Mukherjee, and S. K. S. Gupta, "Ensuring safety, security, and sustainability of mission-critical cyber–physical systems," *Proceedings of the IEEE*, vol. 100, no. 1, pp. 283–299, 2011.

[45] D. Sadigh, A. D. Dragan, S. Sastry, and S. A. Seshia, "Active preference-based learning of reward functions." in *Robotics: Science and Systems*, 2017.

[46] C. Wu, A. M. Bayen, and A. Mehta, "Stabilizing traffic with autonomous vehicles," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 1–7.

[47] A. Bajcsy, S. L. Herbert, D. Fridovich-Keil, J. F. Fisac, S. Deglurkar, A. D. Dragan, and C. J. Tomlin, "A scalable framework for real-time multi-robot, multi-human collision avoidance," in *International Conference on Robotics and Automation*. IEEE, 2019, pp. 936–943.

[48] C. Bartneck, K. Yogeeswaran, Q. M. Ser, G. Woodward, R. Sparrow, S. Wang, and F. Eyssel, "Robots and racism," in *Proceedings of the 2018 ACM/IEEE international conference on human-robot interaction*, 2018, pp. 196–204.

[49] R. Sparrow, "Robotics has a race problem," *Science, Technology, & Human Values*, vol. 45, no. 3, pp. 538–560, 2020.

[50] D. Herath, J. McFarlane, E. A. Jochum, J. B. Grant, and P. Tresset, "Arts+ health: New approaches to arts and robots in health care," in *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 2020, pp. 1–7.

[51] G. B. Verne, "Adapting to a robot: Adapting gardening and the garden to fit a robot lawn mower," in *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 2020, pp. 34–42.

[52] H. T. Javed, M. O. Beg, H. Mujtaba, H. Majeed, and M. Asim, "Fairness in real-time energy pricing for smart grid using unsupervised learning," *The Computer Journal*, vol. 62, no. 3, pp. 414–429, 2019.

[53] C. Utz, M. Degeling, S. Fahl, F. Schaub, and T. Holz, "(Un)informed Consent: Studying GDPR Consent Notices in the Field," in *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*. London United Kingdom: ACM, Nov. 2019, pp. 973–990. [Online]. Available: https://dl.acm.org/doi/10.1145/3319535.3354212

[54] M. Nouwens, I. Liccardi, M. Veale, D. Karger, and L. Kagal, "Dark Patterns Post-GDPR: Scraping Consent Interface Designs and Demonstrating their Influence," p. 12.

[55] R. Dobbe, T. K. Gilbert, and Y. Mintz, "Hard choices in artificial intelligence: Addressing normative uncertainty through sociotechnical commitments," *arXiv preprint arXiv:1911.09005*, 2019.

[56] C. Barabas, C. Doyle, J. Rubinovitz, and K. Dinakar, "Studying up: reorienting the study of algorithmic fairness around issues of power," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020, pp. 167–176.

[57] J. Moore, "Towards a more representative politics in the ethics of computer science," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020, pp. 414–424.

[58] T. N. Bonner, *Becoming a physician: medical education in Britain, France, Germany, and the United States, 1750-1945*. JHU Press, 2000.

[59] M. C. Romano, "The history of legal clinics in the us, europe and around the world," *Diritto & Questioni Pubbliche*, vol. 16, p. 27, 2016.

[60] R. S. Haydock, "Clinical legal education: the history and development of a law clinic," *Wm. Mitchell L. Rev.*, vol. 9, p. 101, 1983.