Axes for Sociotechnical Inquiry in AI Research

Sarah Dean, Thomas Krendl Gilbert, Nathan Lambert and Tom Zick

Abstract—The development of artificial intelligence (AI) technologies has far exceeded the investigation of their relationship with society. Sociotechnical inquiry is needed to mitigate the harms of new technologies whose potential impacts remain poorly understood. To date, subfields of AI research develop primarily individual views on their relationship with sociotechnics, while tools for external investigation, comparison, and cross-pollination are lacking. In this paper, we propose four directions for inquiry into new and evolving areas of technological development: *value*—what progress and direction does a field promote, *optimization*—how the defined system within a problem formulation relates to broader dynamics, *consensus*—how agreement is achieved and who is included in building it, and *failure*—what methods are pursued when the problem specification is found wanting. The paper provides a lexicon for sociotechnical inquiry and illustrates it through the example of consumer drone technology.

Index Terms—Artificial intelligence, Social implications of technology, Sociotechnical systems

I. INTRODUCTION

Recent years have seen increasing public awareness of the profound implications of artificial intelligence (AI) applications 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. Meanwhile, scholars in Science and Technology Studies (STS) are working to equip system designers with toolboxes to help ground these motivations [1], promulgating the conversation on critical making [2], [3] and value-sensitive design [4] towards technical practitioners.

Due in part to a gap between the pace and investment in advancing fundamental AI technologies and reflecting on its potential harms, identifying points of engagement between critical theory and various AI domains remains challenging. Furthermore, emerging subdomains such as AI Safety, Fair Machine Learning (Fair ML), and Human-in-the-Loop (HIL) Autonomy lack a common origin, and subsequently hold different conceptualizations about the relationship between their systems and society [5]. While these differences make it challenging to develop interventions with field wide relevance, the distinct lines of inquiry pursued by each subfield present opportunities to research AI's increasingly central role in social relations. As sites of investigation, these domains are not only some of the most promising case studies for refining
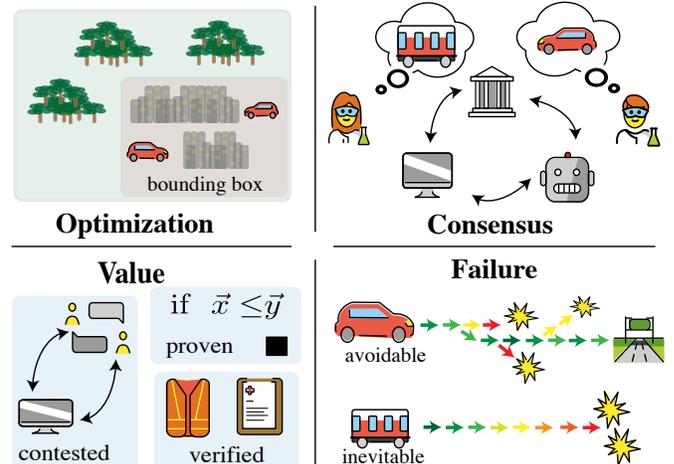
Fig. 1. We propose four vectors for investigating the sociotechnical engagement of various branches of AI research: *value*—what does a field deem of merit, *optimization*—how the designed system within a problem formulation relates to the real-world dynamics, *consensus*—who and what is in the conversation towards agreement, and *failure*—what are the assurances if the system proves inadequate. Engaging with these axes is necessary to evaluate the potential social impact of emerging AI technologies.

theories of sociotechnics over the next decades, but are also viable as beachheads for the normative appraisal of AI as a sociotechnical phenomenon.

To lay the groundwork for evaluating the relationship between AI and the sociotechnical, we portray these subfields as modes of sociotechnical inquiry in their own right. We do not forward a specific sociotechnical lens, ontology, or mode of critical reasoning in order to evaluate them. Instead, we frame these subfields' encounters with sociotechnics as *the problem of defining the interface between systems and reality*. This problem requires the designer to manage how specific technical possibilities depart from and intervene on the state of the world [6], and to define (often implicitly) what success and failure look like in terms of design practice. These design practices may reveal uncertainties, alternate technical formulations, and counterfactual scenarios – in the process generating unprecedented positions of sociotechnical engagement. From these positions, various interpretations of the design situation are simultaneously possible and amenable to inquiry. In this way the social—the matrix of relationships through which design interpretations are forged and considered—emerges *sui generis* from the technical.

AI Safety, Fair ML, and HIL Autonomy pose the directions of inquiry differently to themselves, taking the form of distinct *abstractions* whose assumptions mediate the divide between systems and social order. These abstractions entail different foci and procedures when technical problems arise, giving each community a distinctive research culture. They have

distinctive flavors of inquiry and values that guide that inquiry, particular spaces of investigation, and different limits to what one may or may not think about as "in scope" (i.e. is able to be modeled). While the fields' parameters of inquiry often have substantial overlap, the small differences can yield relationships with sociotechnics that vary widely when actualized. The sociotechnical constraints on inquiry are thus irreducibly phenomenological: a space of interactive and affective concern that is shaped by researchers and in turn shapes the engineering and data science curricula that look to these subfields for guidance. Each space is not a set of precise definitions ready-made for computation, but a continuum whose elements are contested, and in which definitions are proposed, debated, refined, adopted, or rejected.

Textbooks, advisors, journals, and other socio-historical determinants pattern each space of concern and help practitioners internalize it. Still, as the technical possibilities expand in scope and significance, each community must come to terms with its own continuum and create new tools that purport to solve its own sociotechnical problems. This makes phenomenological comparison appropriate for revealing the distinctive elements of each space. Beyond historical genealogy, comparison helps us appreciate the tools for what they are: adjuncts to inquiry that help resolve or defer computational indeterminacies and semantic equivocations.

In this work we aim to reach two audiences. First, we map techniques of deliberation available to practitioners for the purpose of reflecting on abstractions in their own research. Second, we identify contested spaces of normative engagement for external scholars to call their own. We begin by reviewing the three subfields, pointing out key research themes. Next we present our axes of phenomenological comparison, namely *value*, *optimization*, *consensus*, and *failure* as highlighted in Fig. 1. We then conclude with a case study to illustrate the practical stakes of these differences of mindset in the context of a specific near-future AI system: consumer drone technology. Together, our hope is that these audiences will make sustained and constructive engagement on sociotechnical questions achievable for other subfields that are still nascent or have not yet been created.

II. HISTORICAL BACKGROUND

Recent examples of technical work grappling with the societal implications of AI include 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). The methods by which the subfields define and pursue their own sociotechnical questions reflect upon the history and culture of the communities. In what follows, we outline the motivating concerns of each subfield and identify key developments in their evolution. For a more complete analysis see [5].

A. AI Safety

The field of artificial intelligence (AI) has often situated itself within a wider disciplinary context. Famously, at foundational summits at Dartmouth and MIT, computer scientists,

logicians, and psychologists came together to chart a course artificial intelligence researchers could follow to arrive at human-like capabilities with machines [7]. That course, which laid the groundwork for "symbolic" or "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 [8], [9] or to biologically-inspired, connectionist strategies of learning [10]. This branch of AI has also borrowed much inspiration from the economic principles of *mechanism design*, an approach to game theory in which the designer first specifies a desired outcome and then designs agent incentives so that the outcome is achieved [11], [12], [13], [14], [15].

AI Safety began to solidify from these interdisciplinary underpinnings around growing concerns that AI systems could become capable of endangering humans and society writ large [16], [17], [18]. Motivated both by longstanding concerns about the possibility of a "technological singularity" [19] as well as recent expansive applications of machine learning in critical infrastructure domains, many AI Safety promoters fear that AI research projects are approaching a level of capability that will expand beyond their control [19], [20].

Regardless of the likelihood of such a scenario, the nascent field of AI Safety has arisen to preemptively confront these dangers. Despite recent high-profile endorsements from computer scientists and philosophers such as Stuart Russell [21] and Nick Bostrom [20], AI Safety is still a nascent research community facing debates over when and if their concerns will manifest. 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.

B. Fair Machine Learning

The field of machine learning (ML)—primarily concerned with the science of making decisions and models from data—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. For example, a classifier (or model) trained with labelled examples will be evaluated 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. 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].

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].

Importantly, FAccT is a strong locus of interdisciplinary thought within computer science. Building upon model-focused concepts like explainability, blendings of technical and legal concepts of recourse [34] and contestability [35] are employed to approach community goals of accountability. 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].

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 visions of human-robot interaction (HRI) at its core [38]. Early work was closely related to the study of human factors, an interdisciplinary endeavor drawing on engineering psychology, ergonomics, and accident analysis [39]. With advancements in robotic capabilities and increasing prevalence of 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 Institute

of Electrical and Electronics Engineers (IEEE) International Symposium on Robot & Human Interactive Communication. By incorporating principles from the social sciences and cognitive psychology, HRI practitioners formulate a control problem with predictions and models of human behavior.

Since then, 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 novel computational methods for autonomy and the prospective market penetration of new technologies.

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 [40], [41]. 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 [42], [43], 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 such as an explicit *reachability* safety criterion [44].

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 Association for Computing Machinery (ACM) or IEEE Conferences on Human-Robot Interaction indicates an emerging interest in how robotic systems interact with society more broadly. 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. PROPOSED TOOLS FOR SOCIOTECHNICAL INQUIRY

In this section we outline axes that can be used to unpack the assumptions and bounding boxes used by different subfields to interpret the interface between their technology and society. These axes are intended to knit together a phenomenological articulation of the sociotechnical dimensions of AI and provide points of engagement for qualitative reflection. They can be understood in terms of four modalities of abstraction available to present and future AI systems: *value*, different interpretations of the interface that delimit how abstractions are evaluated; *optimization*, to what extent the system design in a problem formulation matches the true domain dynamics; *consensus*, methods of reaching agreement on evaluations;

failure, how to handle and compensate for the limits of the abstraction and solution. Key terms from this analysis with respect to the subfields of AI presented in Sec. II are shown in Tab. I. As each community is quite young, these tastes are not set in stone, and researchers do have the ability to cross-pollinate abstraction strategies between camps.

A. Value: What are the nature and metrics of value?

The names AI "Safety", "Fair" ML, and HIL "Autonomy" ought not be taken too seriously. There is broad agreement across communities that socially responsible and effective AI systems must be simultaneously safe, fair, and independent from constant human oversight [45], [46], [47], [48]. Instead, these fields' differences are rooted in incommensurate interpretations of these labels: in terms of how they pertain to system design and which technical abstractions they largely take for granted. As such, these terms are better conceived as metaphors that more or less problematically outline what the subfields themselves see as their distinctive research contribution to the AI landscape, and how they claim to identify the key features of a well-defined system.

Each subfield has a particular style to its inquiry. This manifests in two ways: a fundamentally shared sense of the value in particular modes of abstraction, and an accepted range of play for elaborating, modifying, refining, or discarding particular features of the abstraction. Together, these elements constitute the distinctive "tastes" present in each community.

AI Safety evaluates a given abstraction strategy in terms of its promise for *epistemic alignment*, conveyed in answer to the following query adapted from [21]: *Have all sources of human concern been modeled in terms of uncertainty, with probabilities able to be learned for every logically possible human preference?* The elegance of this approach is that it sidesteps a potentially fraught appraisal of what human values are at stake in a given domain, in favor of computationally modeling them by observing the human behaviors present in that domain. A well-designed system will thereby learn how to determine the right thing to do, and a designers' job is merely to clear this threshold rather than purport to answer substantive questions of meaning or significance.

Yet there is some disagreement on whether and how designers can maintain this criterion. One example debate is the tension between alignment and corrigibility, i.e. the ability to intervene on and correct the behaviors learned by the system once it has been designed [49], [50]. Corrigibility would not be a pressing concern for a perfectly-designed system, as guarantees would be in place on exhaustive value learning and the perfect simulation of human preferences. At some limit, it is not even clear such a system would need to be overseen by humans at all. There is also tension between empiricist and theoretical approaches to epistemic alignment: is it better to model probabilities through unprecedented scales of computation and data collection [51], [52], or by refining the application of expected utility theory to particular domains [53], [54]?

Fair ML instead focuses on the *distributive guarantees* made possible by the abstraction [55]: *Has a discursive formulation of distributive choices been framed or made tractable in relation to a contemporary computational procedure, whose implications are well-defined?* This field does not claim to translate human behaviors into abstracted forms of knowledge, but to translate the normative criteria already relied on by human institutions into an algorithmic setting. Because many of these criteria are implicit or lack specification, this translation constitutes a major research challenge. It is bolstered by efforts to make the resulting abstractions transparent and accountable in addition to technically fair.

Values within Fair ML are partially articulated through contention regarding what transparency and accountability actually mean. One example is research on explainability, which tries to transform the system from a "black box" (i.e. a system whose outputs do not have a understandable cause) into one whose input-output mappings are paired with a descriptor that can be interpreted [56], [57]. Critiques of exclusively formal work claim that the "mathiness" culture of the Fair ML community causes an insensitivity to social context [58], [59]. This is perceived as problematic because the integrity of Fair ML research depends on leveraging computational tools to meaningfully represent human concerns, based on inquiry into the boundary between computational and discursive approaches to social problems. This tension is integral to the distinctive Fair ML research agenda [60], [61].

Finally, HIL Autonomy concentrates its attention on *precise control* in the context of system operation [62]: *Has the system been set up to operate well in the presence of human activities, such that the boundary between human and system can be known and effectively controlled?* There is less of a focus on translating the lived experience of some human domain into an algorithmic model. Instead, there is a sustained effort to balance abstract learning procedures with domain-specific assumptions, in order to generate guaranteed means of control over system behavior. The designer must appraise the domain in terms of how it actually works, rather than via desired behaviors or distributive priorities.

In lieu of normative debates, there are outstanding questions related to robust learning and comparing various methods for human control. A prominent example is legibility as a criterion for determining the boundary between system and environment: autonomous vehicles may perform suboptimal behaviors (like acting submissively at four-way stops) if it aids coordination with human drivers [63], [64]. This is distinct from both transparency and corrigibility, as the goal is not to aid diagnostics or intervene on value learning but to incorporate legibility directly into the system behavior and leverage this to improve model performance.

B. Optimization: What is the relationship between the system and the broader environment dynamics?

The formulation of structured optimization problems is the primary avenue for the representation of value in technical fields. Optimization refers to how system performance is rendered effective or sufficient according to specified metrics. This reflects how domain dynamics figure into the model specification—what a system is able to do, how it is framed,

Inquiry	AI Safety	Fair ML	HIL Autonomy
Value	Epistemic Alignment	Distributive Guarantees	Precise Control
Optimization	Mechanism Design	Classification	Reachability
Consensus	Forecasting	Debate	Certification
Failure	Provable Benefit	Contestability	Stability

TABLE I

AXES OF COMPARISON FOR SOCIOTECHNICAL INQUIRY. THESE TERMS SERVE TO HIGHLIGHT ESSENTIAL PROPERTIES OF EACH SUBFIELD ALONG THE SET OF AXES WE PROPOSE. WE ELABORATE AND EXPLORE RELATED PROPERTIES IN SECTION 3.

and who can control or oversee its functioning. The problem of optimization thus constitutes a separate axis of comparison between fields, to examine the decision of which environmental dynamics are in or out of scope for technical work.

In AI Safety, the dynamics of the system are not explicitly modeled, but are generally deferred to the optimization of the system. The paradigm draws heavily on the game theoretic underpinnings and is very much analogous to mechanism design. Traditionally, mechanism design maintains a relative indifference to the resulting behavioral dynamics so long as the desired end result is achieved – what matters is the parameters the designer has placed on that end result, as well as a willful inattention to how those parameters are met. For example, autonomous vehicles might be made safe by first modeling the entire roadway and then creating behavioral incentives for all road users that would align their activities with the model specification.

Approaches taken in HIL Autonomy developed from origins in control theory, and thus in contrast tend to begin with the underlying dynamics that define the domain (whether physical, informational, or social). As such, the dynamics serve to ground the model abstraction and are used with a formal verification procedure to guarantee that the system exhibits desired behaviors. The dynamics themselves are treated as sacrosanct, either because they are natural and thus beyond designer control (like Maxwell's laws describing electromagnetism) or protected by law or custom and thus outside the system specification [65]. To return to the above example, HIL Autonomy theorists would instead rely on deterministic physical constraints (e.g. physical laws of motion and following distances between vehicles) and human factors evaluations of the road environment to achieve safety through an analysis of *reachability*: controlling to prevent the worst possible model state (i.e. crashes) by calculating all possible evolutions of an environment and likelihood of danger [66].

Fair ML, however, relies predominantly on supervised learning within which domain features cannot be changed and which is not canonically represented in terms of feedback loops or evolution over time [60], [67]. The technical focus is not on defining ends or protecting these dynamics but surgically intervening on them through "one-shot" *classification* decisions in order to meet specified thresholds of fairness, accountability, or transparency. ML is typically treated as a tool for precisely achieving institutional priorities that are already in place, like credit scoring [68], school admissions [69], and predictive policing [70]. Correspondingly, Fair ML practitioners may be more concerned with accounting and assigning responsibility for crashes – which may entail rigorous data collection on where and how autonomous vehicles are involved in crash events – than minimizing their likelihood.

We can compare these approaches by appraising their assumptions about the interface between system and world. In AI Safety, the model is meant to intervene on and structurally transform the dynamics of the domain in order to guarantee whatever ends the designer has specified. In Fair ML, the model serves as a mere snapshot of the domain dynamics and is meant to administer the proper classification according to canonical metrics. Meanwhile for HIL Autonomy, the model acts more as a hook into the "natural" dynamics of the domain, and is designed to respond to feedback in ways the designer can control. In this way, each subfield is more or less intensely concerned with specification or optimization based on whether the goal of abstraction is to transform, manage, or redesign the dynamics of the domain.

C. Consensus: What are methods for reaching agreement on risk and who is included in the process?

Each subfield's orientation towards problems of interest is only partially determined by their distinctive tastes, values, and favored approaches to optimization and specification. The particulars of how subfields reach consensus on these topics is illustrative of further differences in how practitioners behave and develop technology. By contrasting the approaches to reaching consensus, it is possible to highlight the varying levels of deference to authority, legacy of regulatory structure, and different timescales of interest.

A common feature of the AI Safety community is *forecasting* future AI capabilities against various time horizons based on expert opinions [71]. For example, it is common to forecast when autonomous vehicles will achieve full autonomy and when consumers will adopt said technology [72]. This process favors the opinions of highly regarded researchers in AI and Computer Science. The problem of reaching consensus on the risks of AI development thus relies heavily on figures of authority who are largely members of technical fields without substantial social science training or affiliation. Interdisciplinary voices include vigorous discussions surrounding AI policy [15] and ethics [73] as they relate to forecasts of future technological capabilities, often based in long-term scenario planning. This focus on far-off implications makes apparent the extreme timescales of interest to researchers in this subfield. However, there is a corresponding focus on minimizing the uncertainty of predictions in order to balance these projections. The idea of exponential technological

growth underlies the assumed importance of diligent, early investigation: progress will seem slower in the short-term and simultaneously be ever-accelerating in the long run.

Comparatively, the subfield of Fair ML is marked by its practice of interdisciplinary *debate*. Positioned from existing frameworks of nondiscrimination law and theories of social justice, the community is comprised of academics in Law, Policy, and Science and Technology Studies in addition to Computer Science. There is further an appeal to understand and improve the lived experiences of marginalized populations. In contrast to AI Safety's respect for authority coming from positions of power, the Fair ML community recognizes how individuals with limited power are often early indicators of harmful technology due to disproportionate exposure and compounding effects of overlapping technologies. The subfield thus has a clear focus on existing inequalities, while articulating concerns about how near-term deployment of tools will change social relations, such as automated scoring and surveillance [74]. Fair ML champions the idea that the risks of AI take place on human timescales and have human impacts, so those affected should be part of the conversation.

The subfield of HIL Autonomy rests on a legacy of public-private coordination and *certification*. For example, the aerospace industry is characterized by collaborations between federal regulatory agencies (e.g. Federal Aviation Administration) and large companies (e.g. Boeing) [75]. Engineering expertise and safety culture play an important role in developing regulation, but once created, the resulting specifications do not usually require further debate. There is a clear focus on broadly agreed upon notions of physical safety. Even when metrics are contested, the process is better characterised as a re-evaluation than a deliberative or normative reflection. Modern research on HIL Autonomy, such as algorithmic human robot interaction (HRI) looks to fields like psychology for tools to plan around human behavior, but isn't yet characterized by a strong culture of including voices of non-technical disciplines. Furthermore, the subfield of HIL Autonomy is largely concerned with the instantaneous behaviors of physical systems, where rapid computation is necessary for preventing instabilities. Rather than considering the accumulation of risk over years as technologies are developed, the main concern is on this minuscule timescale of harm caused by system instability.

D. Failure: Is failure inevitable, how is it contained, and how is it measured?

The way in which a domain handles failure is perhaps the most relevant to identifying regulatory interventions. Yet, the manner in which failure is abstracted varies across the three subfields we explore in this work. The types of failure modes considered, who or what is meant to identify them, and the way through which they are addressed are all distinct. The modes of failure considered are essentially the counterparts of the value abstraction around which each field is oriented, while the manner in which failure may be identified and addressed is influenced by the optimization and consensus axes respectively. In this way, the conception of failure within each of the fields we discuss is very much a function of the abstractions discussed earlier in this section. In this discussion it is useful to further decompose the failure abstraction into two components: *robustness*, or how likely a system is to fail, and *resiliency*, whether the system can fail safely. How AI Safety, Fair ML, and HIL Autonomy construe and emphasize these dimensions of failure provide valuable insights into their conception of the sociotechnical.

AI Safety is primarily concerned with the robustness of a system and the notion of safety by design. The aim is to develop *provably beneficial* autonomous agents. If an agent is aligned to human values, it will act in the best interests of humanity, even when situations arise that had not been considered at design [21]. In this way, resilience is delegated to the optimization the agent performs post-deployment. Success, with regards to failure, is a system that adapts and avoids potential pitfalls. Accordingly, the conception of failure in AI Safety is predominantly focused on objective misalignment, or the mismatch between the objectives the system was designed to achieve and those it was optimized to achieve [53]. This conception of failure is captured in the classic Paperclip problem, in which an agent directed to efficiently produce paperclips ultimately destroys humanity on its singular path towards its objective. The stakes on which this dominant failure mode operate, motivate the off-switch intervention [76]. In other words, if the system fails, the only recourse is to shut it down. In this sense, the notion of resilience is not explicitly optimized for in AI Safety.

Fair ML on the other hand is oriented particularly around resilience. As a subfield, it operates on inherently opaque principles (e.g. fairness, equity, etc.), meaning that to some extent, system failure is inevitable. Though the values of the system are generally formalized and explicitly modeled, system failure is not. Rather, a safe interface between the the system and society is an iterative one. Resilience in Fair ML requires that the system be *contestable*. Contestability is a discursive process that engenders system qualities like explainability and transparency [35]. If a system has the capacity to engender harm through this process or cannot accommodate proper iteration, it may be deemed an unfit application of ML [77]. Notably, this notion of refusal has been considered in the context of facial recognition technology [78], a potentially high-stakes application in which safe iteration towards a more robust system is challenging.

Finally, HIL Autonomy is oriented about both robustness and resiliency. In other words, a successful system must be designed to operate safely, but must also be capable of failing safely. Safety is often defined in terms of *stability*, where all states of a system converge in a guaranteed manner to a desired value. Certification processes require that the likelihood of instability and failure remains small, necessitating a complete accounting of the environment and explicit modelling of potential failure modes given a specified environment [79]. These stringent requirements often necessitate the practice of over-specification, in which the system is hierarchically decomposed so that failure modes are understood as the result of simultaneous failures in subsystems [80]. The probability that each subsystem fails must be suitably small, motivating

an emphasis on robustness. The probability that subsystems fail simultaneously must also be small, promoting redundancy and resilience in system design. In this way, the specification of safety in the context of HIL Autonomy guides system designers to hold explicit conceptions of both components of safety.

IV. Case Study on Sociotechnical Inquiry: Consumer Drone Technology

Intelligent consumer drones are beginning to interface with society in many ways: personal entertainment, commercial logistics, local security, and more [81]. In this section we study sociotechnical inquiry into consumer drone technology from the perspective of different AI subfields introduced in Sec. II. The case study highlights the axes for comparison shown in Tab. I. Individual consumer drones are gaining substantial capabilities by leveraging multiple advances in artificial intelligence, with computer vision, for example, enabling autonomous tracking and exploration in commercially available products [82], [83]. Future iterations of these devices can take many forms of autonomy—from a security risk as *slaughterbots* [84], to virtual assistants that have more direct integration with the social [85], and to groups of agents working together to accomplish at-scale distributed tasks [86].

HIL Autonomy most directly integrates with the sociotechnical of consumer drones—the technology is studied from a lens of minimizing physical harm. Drawing on its roots in control theory, HIL Autonomy considers consumer drones as an open problem of robust and agile control in the presence of human uncertainty. Broadly, the subfield optimizes the system dynamics by wrapping a bounding box around the robot itself, and the environment, which includes humans. At this point, a reachability problem can be formulated to provide numerical guarantees against human collision. To determine which guarantees are necessary, designers might look to an official certification process for various behaviors. The process of certification must be revisited, such as with recent updates to the Federal Aviation Administration's code permitting drone delivery at night and moving towards required pre-flight registration. HIL Autonomy arrives at consensus before and during deployment, primarily to consider the potential failure modes of the designed controllers (e.g. crashes due to algorithmic or mechanical error).

As drones develop and gain more hierarchical and abstract versions of autonomy, AIs planning in the space of goals and intents may create concerns of mis-interpretting human values. AI Safety is concerned with long-term intelligence growth and the alignment between drones, drone swarms, and humans. AI Safety practitioners thus focus optimizations on the long-term behavior of the system at large rather than dangers of specific control commands. To arrive at agreement on the prevention of harm, system designers focus on forecasting outcomes and risks strictly before deployment. Specifically, AI Safety might look to permanently prevent cybersecurity risks around autonomous agents targeting humans with direct harm, similar to the situation portrayed in *Slaughterbots* with independent decisions [84]. For an AI Safety practitioner,

failure in this regard involves pervasive risk to society via numerous dangerous drones without well modeled intents.

Fair ML focuses on the interactions between drones and existing norms, both at the individual and societal levels. The potential for harm arises due to models trained from frequently biased datasets with sometimes biased algorithms. For example, computer vision may be used to identify individuals for payments, monitoring, or other personally identifiable tasks. Fair ML focuses on pinpointing specific forms of inequity—instead of optimizing a model parameter, such as validation accuracy, practitioners might search for solutions in the forms of better datasets and debiasing algorithms. There may be a loop between consensus and failure, where particular inequities motivate debates over solutions. In contrast to other subfields, Fair ML might see some technological failures as indicative of an impossibility, and will refuse to accept certain tools (such as for surveillance). To arrive at consensus after an adverse event, stakeholders are ideally included in the conversation, a practice of great value to development of the sociotechnical.

Both AI Safety and Fair ML take the broader lens and consider how the behavior of a drone, or a swarm of drones, can re-define sets of interactions within a given society. The potential for surveillance and control with at-scale adoption on intelligent drones poses critique from the subfields in different ways. It provides a second entry point into consideration of the sociotechnical: how such systems can augment power structures or perpetuate demographic profiling. The Fair ML community may focus on mitigating harms of individual interactions, but when harm is not controllable, they opt for a flat refusal of technological adoption. AI Safety would view the at-scale swarm problem relatively similarly to the epistemic alignment of individual drones. These differences highlight how current problems in HIL Autonomy are less likely to change dramatically in scope, as they are focused on improving the capabilities of individual agents. On the other hand, as increasingly intelligent behaviors become possible, the views of AI Safety and Fair ML practitioners are more likely to develop at the same pace of technological development. As with any technology, the future path of consumer drone development and adoption is not certain, so regularly re-evaluating pathways of future use is crucial.

V. Conclusion

Given the pace and breadth of the field, constructive tools for sociotechnical inquiry into AI research and technological development are needed. In this paper, we have presented a framework for sociotechnical inquiry along axes of *value*, *optimization*, *consensus*, and *failure* as an initial toolbox to gauge a research domain's relationship with society. The subfields of AI Safety, Fair ML, and HIL Autonomy do not represent a total or exhaustive landscape of social interfaces of AI development, but those fields that have matured recently towards their own lens of sociotechnics. Future study in other wide directions of AI research such as cybersecurity, ML in healthcare, and more are strongly warranted as directions for future work. We hope that this paper helps nurture a community of engaged scholars examining AI's possible interfaces

with society, and encourages further lexicons that facilitate translations between technological and social development.

Acknowledgments

References

[1] 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.

[2] M. Ratto, "Critical making: Conceptual and material studies in technology and social life," *The Information Society*, vol. 27, no. 4, pp. 252–260, 2011.

[3] A. Hamraie, "Mapping access: Digital humanities, disability justice, and sociospatial practice," *American Quarterly*, vol. 70, no. 3, pp. 455–482, 2018.

[4] P. Sengers, K. Boehner, S. David, and J. J. Kaye, "Reflective design," in *Proceedings of the 4th Decennial Conference on Critical Computing: Between Sense and Sensibility*, 2005, p. 49–58.

[5] M. Andrus, S. Dean, T. Gilbert, N. Lambert, and T. Zick, "AI development for the public interest: From abstraction traps to sociotechnical risks," in *2021 IEEE International Symposium on Technology and Society (ISTAS)*, 2021.

[6] H. Rosa, *The Uncontrollability of the World*. John Wiley & Sons, 2020.

[7] S. Russell and P. Norvig, *Artificial intelligence: a modern approach*. Prentice Hall, 2002.

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

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

[10] R. Sun, "Connectionism and neural networks," in *The Cambridge handbook of artificial intelligence*. Cambridge University Press, 2014, pp. 108–128.

[11] D. Hadfield-Menell, S. J. Russell, P. Abbeel, and A. Dragan, "Cooperative inverse reinforcement learning," in *Advances in Neural Information Processing Systems*, vol. 29, 2016, pp. 3909–3917.

[12] J. Leike, M. Martic, V. Krakovna, P. A. Ortega, T. Everitt, A. Lefrancq, L. Orseau, and S. Legg, "AI safety gridworlds," *arXiv preprint arXiv:1711.09883*, 2017.

[13] S. Natarajan, G. Kunapuli, K. Judah, P. Tadepalli, K. Kersting, and J. Shavlik, "Multi-agent inverse reinforcement learning," in *2010 Ninth International Conference on Machine Learning and Applications*, 2010, pp. 395–400.

[14] V. Kuleshov and O. Schrijvers, "Inverse game theory: Learning utilities in succinct games," in *International Conference on Web and Internet Economics*, 2015, pp. 413–427.

[15] 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.

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

[17] E. Yudkowsky, "Artificial intelligence as a positive and negative factor in global risk," in *Global catastrophic risks*. Oxford University Press, 2008, p. 184.

[18] 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.

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

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

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

[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. J. Nilsson, *Learning machines: Foundations of trainable pattern classifying systems*. McGraw-Hill, 1965.

[24] F. Rosenblatt, "The perceptron—a perceiving and recognizing automaton," Cornell Aeronautical Laboratory, Tech. Rep., 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 Trans. Inf. Syst.*, vol. 14, no. 3, p. 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*, 2012, pp. 214–226.

[30] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and Machine Learning*, 2018. [Online]. Available: 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." *Journal Machine Learning Reseach*, 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] M. A. Goodrich and A. C. Schultz, "Human-robot interaction: A survey," *Found. Trends Hum.-Comput. Interact.*, vol. 1, no. 3, p. 203–275, 2007.

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

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

[41] 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.

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

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

[44] 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*, 2019, pp. 936–943.

[45] S. Russell, D. Dewey, and M. Tegmark, "Research priorities for robust and beneficial artificial intelligence," *AI Magazine*, vol. 36, no. 4, pp. 105–114, 2015.

[46] M. Whittaker, K. Crawford, R. Dobbe, G. Fried, E. Kaziunas, V. Mathur, S. M. West, R. Richardson, J. Schultz, and O. Schwartz, *AI now report 2018*. AI Now Institute at New York University New York, 2018.

[47] H. Shi, R. V. Prasad, E. Onur, and I. G. M. M. Niemegeers, "Fairness in wireless networks: Issues, measures and challenges," *IEEE Communications Surveys Tutorials*, vol. 16, no. 1, pp. 5–24, 2014.

[48] R. Alur, *Principles of cyber-physical systems*. MIT press, 2015.

[49] N. Soares, B. Fallenstein, S. Armstrong, and E. Yudkowsky, "Corrigibility," in *AAAI Workshop: AI and Ethics*, 2015.

[50] R. Carey, "Incorrigibility in the CIRL framework," in *Conference on AI, Ethics, and Society*, 2018, pp. 30–35.

[51] G. Irving and A. Askell, "AI safety needs social scientists," *Distill*, vol. 4, no. 2, p. 14, 2019.

[52] R. V. Yampolskiy, "Personal universes: A solution to the multi-agent value alignment problem," *arXiv preprint arXiv:1901.01851*, 2019.

[53] N. Bostrom, "The superintelligent will: Motivation and instrumental rationality in advanced artificial agents," *Minds and Machines*, vol. 22, no. 2, pp. 71–85, 2012.

[54] K. E. Drexler, "Reframing superintelligence," 2019. [Online]. Available: https://www.fhi.ox.ac.uk/reframing/

[55] L. T. Liu, S. Dean, E. Rolf, M. Simchowitz, and M. Hardt, "Delayed impact of fair machine learning," in *International Conference on Machine Learning*, 2018, pp. 3150–3158.

[56] W. Samek, T. Wiegand, and K.-R. Müller, "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models," *arXiv preprint arXiv:1708.08296*, 2017.

[57] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[58] O. Keyes, J. Hutson, and M. Durbin, "A mulching proposal: Analysing and improving an algorithmic system for turning the elderly into high-nutrient slurry," in *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–11.

[59] Z. C. Lipton and J. Steinhardt, "Troubling trends in machine learning scholarship," *arXiv preprint arXiv:1807.03341*, 2018.

[60] S. Corbett-Davies and S. Goel, "The measure and mismeasure of fairness: A critical review of fair machine learning," *arXiv preprint arXiv:1808.00023*, 2018.

[61] C. L. Bennett and O. Keyes, "What is the point of fairness? Disability, AI and the complexity of justice," *arXiv preprint arXiv:1908.01024*, 2019.

[62] J. F. Fisac, A. K. Akametalu, M. N. Zeilinger, S. Kaynama, J. Gillula, and C. J. Tomlin, "A general safety framework for learning-based control in uncertain robotic systems," *Transactions on Automatic Control*, vol. 64, no. 7, pp. 2737–2752, 2018.

[63] A. D. Dragan, K. C. Lee, and S. S. Srinivasa, "Legibility and predictability of robot motion," in *Conference on Human-Robot Interaction (HRI)*. IEEE, 2013, pp. 301–308.

[64] C. Lichtenthäler and A. Kirsch, "Legibility of robot behavior: a literature review," 2016. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01306977/document

[65] R. Dobbe, O. Sondermeijer, D. Fridovich-Keil, D. Arnold, D. Callaway, and C. Tomlin, "Toward distributed energy services: Decentralizing optimal power flow with machine learning," *IEEE Transactions on Smart Grid*, vol. 11, no. 2, pp. 1296–1306, 2019.

[66] S. Bansal, M. Chen, S. Herbert, and C. J. Tomlin, "Hamilton-Jacobi reachability: A brief overview and recent advances," in *Conference on Decision and Control*, 2017, pp. 2242–2253.

[67] D. Madras, E. Creager, T. Pitassi, and R. Zemel, "Learning adversarially fair and transferable representations," in *International Conference on Machine Learning*, 2018, pp. 3384–3393.

[68] T. Kamishima, S. Akaho, and J. Sakuma, "Fairness-aware learning through regularization approach," in *2011 IEEE 11th International Conference on Data Mining Workshops*, 2011, pp. 643–650.

[69] A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach, "A reductions approach to fair classification," in *International Conference on Machine Learning*, 2018, pp. 60–69.

[70] A. Chouldechova and A. Roth, "The frontiers of fairness in machine learning," *arXiv preprint arXiv:1810.08810*, 2018.

[71] 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.

[72] P. Bansal and K. M. Kockelman, "Forecasting Americans' long-term adoption of connected and autonomous vehicle technologies," *Transportation Research Part A: Policy and Practice*, vol. 95, pp. 49–63, 2017.

[73] I. Gabriel, "Artificial intelligence, values and alignment," *Minds and Machines*, vol. 30, no. 3, p. 411–437, 2020.

[74] C. O'Neil, *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown, 2016.

[75] L. Ilcewicz and B. Murphy, "Safety & certification initiatives for composite airframe structure," in *46th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference*, 2005, p. 1877.

[76] D. Hadfield-Menell, A. Dragan, P. Abbeel, and S. Russell, "The off-switch game," *arXiv preprint arXiv:1611.08219*, 2016.

[77] R. Benjamin, "Informed refusal: Toward a justice-based bioethics," *Science, Technology, & Human Values*, vol. 41, no. 6, pp. 967–990, 2016.

[78] ACM US Technology Policy Committee, "Statement on principles and prerequisites for the development, evaluation and use of unbiased facial recognition technologies," Association for Computing Machinery, Tech. Rep., 2020. [Online]. Available: https://www.acm.org/binaries/content/assets/public-policy/ustpc-facial-recognition-tech-statement.pdf

[79] G. Katz, C. Barrett, D. L. Dill, K. Julian, and M. J. Kochenderfer, "Reluplex: An efficient smt solver for verifying deep neural networks," in *International Conference on Computer Aided Verification*, 2017, pp. 97–117.

[80] B. Brosgol and C. Comar, "DO-178C: A new standard for software safety certification," Ada Core Technologies, Tech. Rep., 2010.

[81] D. Floreano and R. J. Wood, "Science, technology and the future of small autonomous drones," *Nature*, vol. 521, no. 7553, pp. 460–466, 2015.

[82] E. Ackerman, "Skydio demonstrates incredible obstacle-dodging full autonomy with new R1 consumer drone," *IEEE Spectrum*, 2018.

[83] "DJI - official website." [Online]. Available: https://www.dji.com/

[84] S. Russell, A. Aguirre, A. Conn, and M. Tegmark, "Why you should fear "slaughterbots"—a response," *IEEE Spectrum*, 2018.

[85] A. Bachrach, A. Bry, and M. Donahoe, "Magic wand interface and other user interaction paradigms for a flying digital assistant," Jun. 13 2017, US Patent 9,678,506.

[86] E. Vattapparamban, I. Güvenç, A. I. Yurekli, K. Akkaya, and S. Uluağaç, "Drones for smart cities: Issues in cybersecurity, privacy, and public safety," in *2016 international wireless communications and mobile computing conference (IWCMC)*, 2016, pp. 216–221.

Sarah Dean Sarah is a PhD candidate in the Department of Electrical Engineering and Computer Science at UC Berkeley. She received her BSE in Electrical Engineering and Math from the University of Pennsylvania. Her research focuses on developing principled data-driven methods for control and decision-making, inspired by applications in robotics, recommendation systems, and developmental economics.

Thomas Krendl Gilbert Thomas Krendl Gilbert is an interdisciplinary PhD candidate in Machine Ethics and Epistemology at the University of California, Berkeley. He is an affiliate of the Center for Human-Compatible AI, and recently served as the inaugural Law and Society Fellow at the Simons Institute for the Theory of Computing.

Nathan Lambert Nathan Lambert is a PhD Candidate at the University of California, Berkeley. He is a member Department of Electrical Engineering and Computer Sciences, advised by Professor Kristofer Pister in the Berkeley Autonomous Microsystems Lab. His work explores many topics on model learning and decision making with data-driven and analytical methods. He received his BS in Electrical and Computer Engineering from Cornell University in 2017.

Tom Zick Tom Zick received her PhD from the University of California Berkeley in 2020. She is currently a postdoctoral fellow at the Berkman Klein Center for Internet and Society working at the intersection of law and AI ethics. Outside of academia, Tom has worked in municipal government, crafting data governance principles for the city of Boston.