Title

Modes of Deliberation in Machine Ethics

Author

Gilbert, Thomas

Publication Date

2021

Peer reviewed|Thesis/dissertation

Modes of Deliberation in Machine Ethics

By

Thomas J Gilbert

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Machine Ethics and Epistemology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor David Bates, Chair
Professor David Grewal
Professor Alva Noë
Associate Professor Anca Dragan
Assistant Professor Cathy Wu

Summer 2021

Abstract

Modes of Deliberation in Machine Ethics

by

Thomas J Gilbert

Doctor of Philosophy in Machine Ethics and Epistemology

University of California, Berkeley

Professor David Bates, Chair

This dissertation is about the purpose of artificial intelligence (AI) research. New learning algorithms, scales of computation, and modes of sensory input make it possible to better predict or simulate decision-making than ever before. But this does not tell us whether or how AI systems should be built. In fact there is much anxiety about how to build AI applications in ways that respect or enact the decision criteria of existing human institutions. But instead of how to better predict or protect how we decide things, my research question is: how can AI tools be used to *reorganize* the choices we make about how we want to live together? Answering this question requires investigating the conditions under which deliberation is possible about the systems being built—their models, their real-world performance, and their effects on human domains. These three modes of deliberation are philosophically outlined in the introduction and named as *sociotechnical specification*, *normative cybernetics*, and *machine politics*.

The first chapter pursues sociotechnical specification in the context of routing algorithms for autonomous vehicle (AV) fleets. It asks what it would mean to relate this emerging transportation model to the other legacy systems adjacent to the travel domain. It sketches proxies in terms of "known unknown" features of the driving environment. These would need to be monitored and serve as targets for optimization in order for the performance of the AV fleet to be considered to be robustly good. The second chapter pursues a normative cybernetics of AVs in terms of a sustained internal critique of reinforcement learning (RL). This introduces new policy questions, whose answers would correspond to types of feedback between the behavior of AV firms and civil society or state organizations. The third chapter outlines the elements of machine politics in terms of concepts borrowed from contemporary analytic philosophy. Ruth Chang's notion of parity is mobilized to demonstrate the possibility of domain deliberation at different stages of AI development. This comprises a critique of existing schools of thought, represented here in terms of epistemicism (the notion that the structure of human activities can be passively learned and observed) and ontic incomparabilism (the notion that human activities cannot be organically modeled or developed by means of AI). The three types of feedback that are produced through active developmental inquiry are presented in terms of featurization, optimization, and integration, all of which comprise the structural choices at stake in machine politics.

TABLE OF CONTENTS i

LIST OF FIGURES AND TABLES

ACKNOWLEDGEMENTS

I wish to thank my advisor, David Bates, for taking a chance on this project and agreeing both to serve as my chair and to see through my application for an individual PhD. Without his professional and intellectual guidance, this dissertation would not exist. I am also grateful to Tania Lombrozo and Deirdre Mulligan for their service on my qualifying exam committee. I am additionally thankful to Alva Noë, Anca Dragan, David Grewal, and Cathy Wu for their service on the final dissertation committee. I have learned more from them than I can state here, and expect I will continue to do so in the coming years.

Piet Hut asked me on a memorable day in September 2015 to imagine what research project I would pursue if disciplines didn't exist. Iterating on that prompt was the seed of this dissertation. Daniela Cammack and David Grewal arrived late to the process, joining Berkeley after much of my research was conducted. Most of the ideas in the final product--on vagueness, the philosophy of measurement, the saliency of political economy--were already gestating when I met them in fall 2019. I am nevertheless grateful for their shepherding and greatly expanding my understanding of these issues, and also for their friendship, a particular bright spot throughout the COVID-19 pandemic. Their influence can be found throughout the dissertation, and particularly the introduction, whose engagement with Aristotle I owe largely to their teaching and encouragement. Meanwhile, from my PhD application to exam prep to research phase to writing to final defense, Jim Stockinger has served as a constant mentor and companion. Thanks for the coffees, discussions, and stories, Jim. Finally, I could not have navigated the personal, professional, and spiritual dimensions of designing and finishing my own PhD without the guidance of Barbara Vivino. My work with her was as, if not more, significant to me as the dissertation itself.

I have benefited greatly from the willingness of faculty and institutes to invest in my research, often beyond what I felt was warranted. In particular, I wish to thank Stuart Russell and Mark Nitzberg for welcoming me to the Center for Human-Compatible AI, without which neither my dissertation nor PhD would exist. Their leadership and professional support has been unparalleled in my experience at Berkeley and will serve as a model for my future career. I also thank the Simons Institute for the Theory of Computing and the Center for Long-Term Cybersecurity for supporting distinct stages of my research which proved integral to the second chapter.

Berkeley's vibrant and enormous graduate community is what I will miss most from this period of my life. I wish to single out Michael Dennis, Roel Dobbe, and Yonatan Mintz for serving as collaborators as well as great friends. I am a different and much better person, not to mention scholar, because of them. My fellow members of GEESE (McKane Andrus, Sarah Dean, Nitin Kohli, Nathan Lambert, Tom Zick) helped me translate my outrageously diverse interests and background into something commensurate with academic research in computer science. Because of them, being in my own department never felt lonely. Somewhat beyond UC Berkeley, my friendships with Kuan-Ying Fang, Andrew Loveridge, and Bill Thanhouser have and will continue to sustain my endeavors, professional and otherwise.

I owe the most to my partner, Monica Porter, whose unflagging support and conscientiousness often served as its own source of inspiration, and made me want to work to deserve it.

1. INTRODUCTION

My goal in this introduction is to construct a philosophical grammar for deliberation about the use of AI in various human domains, in particular the development of self-driving cars. Even if the present structure of a domain is determinate, any technical intervention on it breaks that structure and requires it to be reconstituted. This entails articulating new abstractions and frames within which to think about specific domains, in other words to deliberate about what they are and what their purpose is. Just as the concept of jaywalkers came into being after cars "broke" roads a hundred years ago, self-driving cars will change how roads work and generate new activities that cannot be fully anticipated. Consequently, the conceptual problem is how to deliberate about roads so that the activities that emerge do so in a fair and safe manner, not to formalize fairness or safety in advance of technical intervention. I will return to the ideas in this paragraph again and again in the course of this introduction.

I first present **feedback** as a lurking problem for advanced AI systems deployed in human domains, before reinterpreting the concept of **deliberation** itself. Next, I examine other features of domains (in particular their **indeterminacy** and **structure**) that future AI systems will enframe and compel us to coherently organize on new ground. I then outline three modes of deliberation (which I name as **sociotechnical specification**, **normative cybernetics**, and **machine politics**) that are needed to reflect on these features and incorporate them into how AI systems are designed, trained, and deployed in human contexts. After summarizing how each of my dissertation chapters concentrates on these respective modes, I conclude by suggesting a new kind of institutional space--the AI clinic--that would nurture and distinguish these modes in the context of building real-world systems.

1.1 The Problem of Feedback

All cities have problems, and making them smarter changes the problems they face. The prospect of smart cities amounts to the integration of streets, buildings, power lines, smart phones, and other components to support a dataflow that both optimizes and structurally reorders urban life. This techno-utopian vision of cities as a proxy for human organization promises to improve quality of life using AI applications, including self-driving cars. However, it has been attacked as sacrificing commitments to democracy and equity in the name of technological progress. In particular, Ben Green has instead called for the "smart enough city": the use of new technologies in support of livable, just, and responsibly-innovated cityscapes (Green 2019).

At stake in this debate is the implicit criteria for the success or failure of a given smart city technology. For example, new tools for pothole prediction and smart streetlights promise to prevent road wear and optimize electricity use, but these projects often fail due to emergent problems with technical implementation, inequity of service, or faulty integration with legacy systems. Meanwhile, critics have advocated for transparency, accountability, and documentation of harms to prevent tools' misuse and counter privacy concerns (Whittaker et al. 2018). But these efforts have struggled to articulate what terms like transparency, accountability, and harm amount to in the context of unprecedented data analytics and service provision. At present, the notion of a city becoming smart "enough" remains indeterminate.

These rival approaches share the Platonic assumption of an underlying definition of the good or of value that can be encoded within tools. It would follow that "AI ethics" (or fairness, or safety) is the project of technically realizing that definition. I claim that this assumption is wrong, and that accepting its wrongness is crucial to the successful development of AI systems. Whether or not such a function exists, no one is in a position to discover, measure, or defend it without reordering human experience to reflect its form. While new tools may augment our insight into human domains, they also suggest alternative means of organizing them, and in so doing question their integrity.

This is not a new problem. In fact, Aristotle used the same word, *ὅρος* (hereafter *horos*), to refer to both the limited horizon of human insight and the standard by which basic organizational decisions are made. In doing so, Aristotle suggested a profound relationship between the ways humans are naturally (or artificially) limited and the ways humans delimit their activities. These limits are reciprocal; a change in one reflects some change in the other. A basic expansion or contraction of human agency will reshape the ways we organize our activities in relation to each other, including how or whether we retain them, or forge new ones. This perspective constitutes a major break with the orthodox definition of AI as a system or rational agent that takes actions in some environment in pursuit of maximizing the likelihood of achieving some goal (Russell and Norvig 2002). Instead, following Aristotle, I argue AI is a tool to be used to reorganize the *horoi* that structure social order. It follows that any particular technical enactment of AI (e.g. deep learning, supervised machine learning, symbolic reasoning, etc.) is less important than the political enactment of *horoi* made possible by what new tools permit us to do and reason about.

It follows that new forms of data sharing and service provision will remain brittle if AI applications merely try to optimize domains that already exist. Instead, the relations between domains must be rewired so that they are ordered to work well together, in search of respective integrities that would also comprise an integrated harmony. This entails a problem of unprecedented **feedback** between systems that have not yet learned to speak to each other in terms of themselves. By feedback I broadly refer to the evolutionary effect generated by a system's recent states and actions on its future dynamics (Wiener 2019), raising questions of stability and emergent behavior whose answers require normative reflection about the domain itself.

In this dissertation, I argue that present discussions of machine ethics neglect the modes of deliberation needed to adequately tackle these questions. Without these modes, the problem of feedback—and related problems of domain structure and indeterminacy—cannot be framed in ways that are technically resilient or normatively integral. The purpose of this introduction is to present and justify these modes in relation to existing work in AI safety and governance. I conclude by presenting my dissertation chapters as respectively demonstrating how to deliberate about models, systems, and domains as they are remade by AI applications. Although not a primary focus of this introduction, it follows from my presentation that these modes align with the choices to be faced by particular agents (respectively AI designers, transportation planners, and public utilities commissions). We shall see that Aristotle provides the philosophical tools to understand what this inquiry substantively entails.

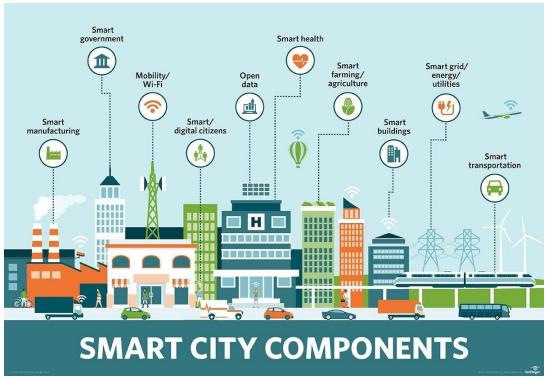
Human capabilities are not something that can be augmented, subverted, or even compared with AI applications. Rather, it is how we incorporate technologies—meaning both how we instantiate

them and make them work for us, and in so doing redefine the terms of how we live together—that reconstitutes human capabilities in the first place. The fact that we bear responsibility for this as a basic feature of our condition is what makes us human. As a result, any adequate discussion of human capabilities must begin with an investigation of machine capabilities, a sustained encounter with what Yuk Hui calls the inhuman (Hui 2019).

Drawing from recent work in the philosophy of technology (Noë 2015), I define machine ethics as the organization of how we use AI to illuminate human domains. This is a necessarily recursive definition, and I will account for its elements step by step: what are domains, what it means to illuminate them, and what it means to organize the way the illumination is conducted. At present, the problem space of machine ethics remains deeply confused about these elements, and fails to distinguish them or fully grasp the logic of how they are related to each other.

A domain is a specified sphere of human activity that fulfills a particular need. Activities sediment habits and rituals that make the domain available for us to lose ourselves in every day. Illuminating the domain means that something about the terms of that sedimentation, previously hidden, is now offered up for inspection and appraisal. Finally, to organize this illumination is to actively attend to how those terms are offered up for appraisal—to whom, under what conditions, by what channels—and establish structured protocols for maintaining that active attention. It follows that deliberation about how to build AI systems well comprises a problem of feedback.

To make this concrete: how should we update stop signs and road signage so that self-driving cars can reliably recognize them? What are the affordances we need to design for these agents, in order for them to be demonstrably safe, fair, and performant? How do we want self-driving car fleets to coordinate the other systems (economic activity, traffic dynamics, planetary climate, etc.) that are brought into a new relationship through their deployment? For what range of damages should service providers be found liable, and on what grounds? Answering these questions means specifying the types of sensory, discursive, and legal feedback needed for automated systems to maintain stability with respect to other systems.



A stylized view of the systems a smart city will integrate (with types of feedback implied).

My definition of machine ethics is close to the realm of politics. Following Aristotle, I take politics to be an activity that organizes the relationship between domains in order to direct them toward human flourishing. Concretely, my thesis is that the technical development of self-driving cars is much closer to the activity of politics than the activity of driving itself, and that automating the

latter analytically entails reorganizing it as part of the former. The modes of deliberation I present later in this introduction are intended to clarify these points by illustrating what it means to organize how we use AI to illuminate particular domains.

The pertinent normative question is how an AI application can be developed to safeguard and perhaps augment the integrity of all the other activities brought into some relationship through its operation. But this is not the way most engineering work on self-driving cars proceeds. At present there is much work on route planning, or on computer vision, or on some other module, under the assumption that stacking these modules on top of each other will approximate what it means to drive a car. The problem with this assumption is not only that its validation remains technically uncertain, but that it circumvents the foundations of what this "stacking" organizationally entails (which is, in fact, what is at stake in basic AI research). As such, there is a need for active inquiry into the paradigms through which AI research is conducted. I thus focus on articulating and defending the principles needed to guide how AI systems are developed and deployed, not on some discrete set of values or features to be encoded in order for systems to work well.

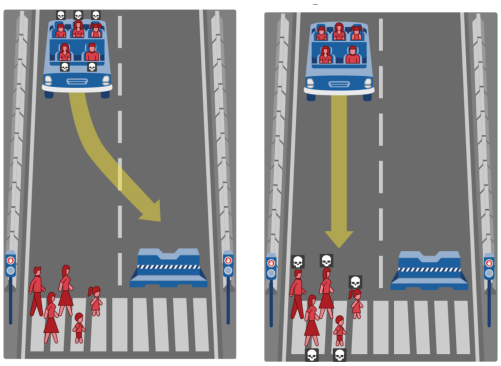
1.2 On Deliberation: Surpassing vs. Organizing Capabilities

The thought experiment of a "trolley problem" has been an influential model for the ethics of automated vehicles (AVs). But the question of giving self-driving cars an ethics is not well-framed, because it is incidental to driving as a human activity. Overwhelmingly, driving's purpose is to connect us to other activities (working, shopping, socializing) that have their own structure and purposes. Even the appeal of the "open road" is relational, serving as a momentary escape from our obligations to others (Crawford 2020). Certainly we talk about "good" or "bad" driving, but we mean this in terms of proficiency, not as a statement about driving as a way we define our relationships and commitments to other people (which is what ethics is about). Making life-or-death decisions from the driver's seat assumes that it is impossible to safely coordinate the activities of pedestrians and cars. We drive because we are trying to get somewhere else, and asking who we would rather kill implies we have failed to achieve that goal.

The problem of designing a transportation system, however, is different. It is like the seam in a garment: sewing it well implies that one has chosen an appropriate type and color of thread, and requires knowledge about how the pieces of fabric are meant to fit together, which itself requires knowledge about the article of clothing, how it is meant to be worn and in what contexts. There is no "ethics" of sewing (other than, perhaps, the banality of not harming anyone with the needle) but there is one for tailoring, in the sense that one is concerned with fitting a customer to make the clothing work in support of their own ends. It is a human enterprise.

To clarify this contrast, it is worth introducing working definitions of recursivity and contingency. A recursion is a type of program or rule that refers to itself. Meanwhile, a contingency denotes an event or circumstance that merely happens to occur and could not be predicted with certainty. Most AI models today are developed by performing a recursion or set of recursions on various observed contingencies—most notably, statistical associations between labeled or unlabeled data. The trolley problem, as a design principle for AVs, confusingly elevates a contingent feature of human driving—the freedom to kill bystanders—to a proposed recursive principle of transportation as a

whole. But this is nothing like the actual regulations and limits used to organize how roads stitch our activities together.



An example of a "moral machine" survey conducted by the MIT Media Lab.

This leads me to an important aspect of deliberation that is widely misunderstood in the AI literature. Stuart Russell, for example, has characterized the classical model of successful reasoning as "methods of logical deduction that would lead to true conclusions given true premises" (Russell 2019, pp. 17-18). Parsing the famous discussion of reasoning in Book III of Aristotle's *Nicomachean Ethics*, Russell concludes that "the 'end'—what the person wants—is fixed and given; and [Aristotle] says that the rational action is one that, according to logical deduction across a sequence of actions, 'easily and best' produces the end" (Ibid. p. 18). Two conclusions follow from this. One is that the major stumbling block to building successful AI agents is for them to know or figure out what actions would most likely lead to the specified goal state, based on the optimal calculation of utility trade-offs. The second is that generalizing Aristotle's account of deliberation would amount to the study of rational gambling under conditions of uncertainty.

But this is not all of what the term means. **Deliberation** (also known as practical reasoning) is an organic calculation, not a strictly mechanical one. It denotes a relation between the elements of something such that they are made to fit together as constitutive parts of a whole. In other words it is structured, integrated, coordinated, and ordered. The part must be made to stand in relation to the whole in a way that is meaningful. A deliberate decision is one that has been fully (that is, organically) considered, not one that has been made in the most efficient manner possible. It is not just reasoning within some frame, but a reappraisal of the frame itself, which changes the relationship between the agent and the object of reasoning. Put differently, deliberation requires a sense of which ends are retroactively justifiable given a proportionate amount and form of reflection, as well as what means would in fact achieve a given end.

The point is not that only biologically-ordered agents like *Homo sapiens* are capable of deliberation. It is that deliberation requires a capacity to organize calculation into a capability, and in so doing reorganize (that is, reconstitute) the being of the thing that undertakes the computation as worth doing. Doctors do not deliberate whether they shall heal, but they do deliberate on treating patients in light of available actions. Their view of that activity depends on the tools at their disposal; a new set of tools may alter their sense of what is possible, and therefore what line of treatment would be prudent (Cammack 2013).

Insofar as they are tools, AI systems do not *have* capabilities; they *are* capabilities. A good example is AlphaGo's defeat of Lee Sedol in 2016. AlphaGo is extraordinarily proficient at playing Go, but it has no grasp of Go as a human activity. Its moves are the result of a strictly mechanical computation aimed to achieve a pre-specified goal, not a sense of what the game is for or its relationship with the opponent. In ancient China, Go comprised one of the four arts that denoted the status of a distinct social group. Playing Go was interpreted as a way of simulating one's place in the universe. The goal was not necessarily to win, but to generate *tesujis*, i.e. moves that caused one to "marvel" at the game's beauty. AlphaGo's technical mastery of the game is such that it has changed our sense of what counts as *tesuji*, literally by expanding the horizon of play. Expert human players have also learned to imitate and incorporate AlphaGo's distinctive strategies into their own playstyles. While AlphaGo cannot play Go as humans do, it has significantly changed how expert human players deliberate when choosing moves.

It is likewise for other AI applications. The use of computer vision models to match photos to police records does not mean that AI can recognize faces, but that the activities of policing are being reorganized. The use of language models to generate text for web applications does not mean AI can interpret speech or even use it, but that our definition of online communication is up for grabs. The use of reinforcement learning to optimize electrical grids or help drones navigate airspace does more to change our sense of how to manage those environments than to automate that management itself. At stake in all these examples is the scale on which deliberation occurs. AI models reconstitute the scale, and in so doing how choices connect with the purpose of the domain. They make it possible to come to a decision and organize choices in relation to ends; they appropriate neither that organization nor the ends that guide it.

My definitions of recursivity and contingency enrich these claims in two ways. First, deliberation is rooted in a concern for how the recursion encounters contingencies. An AI model is a structured regime of anticipation that has learned to predict things based on past observations. Deliberation entails a capacity to coordinate this structure in a meaningful fashion, that is, to organize anticipation in terms of a flow. Second, deliberation incorporates recursion as a basis for organic calculation. Sewing requires calculation, but sewing well assumes a tailor who knows what he or she is doing. These lead to a more precise definition: deliberation is a calculation that organizes the flow generated by encounters between recursivity and contingency (Stiegler 1998).

Let's reconsider the deliberation at stake in AV development. China and Germany, for example, are building entire highways and urban grids from scratch to support AVs by default. Trolley problem-like situations in which AVs would problematically encounter other modes of transport are thereby rendered structurally impossible, and unnecessary to morally "solve". This case reveals several interesting points of contact between recursivity and contingency. First, these nations have the organizational wherewithal to do this in ways that seem unlikely in an American context. Second, the way that China will do this is clearly different from how Germany plans to pursue it, based on its pre-existing vision of how cities are supposed to work. Third, Chinese citizens comparatively associate AVs with socioeconomic mobility, rather than with loss of individual control or privacy as in Germany, partly because physical transportation is more directly tied to mobility in the Chinese context. This is all to say that new forms of technical automation interact with the terms of social organization, not by remaking them from scratch but by displacing them in ways that call for their deliberative reconstitution. Smart cities built near Beijing and Berlin will

still be recognizably Chinese and German in their respective features, but in ways that are not predetermined, based on how their technical components are actively organized in relation with human factors.

This suggests that one cannot train an AV model well without a grasp on what a good road is, or good traffic flow, or a good neighborhood, or indeed any other activity that is brought into a determinate relationship through the model's operation. That model or computation stack's ability to recognize stop signs or navigate an intersection or merge into traffic is secondary to this basic organizational, deliberative question. The model must be incorporated as part of a system that works to support desired capabilities.

The technical surpassing of human capabilities does not mean that AI has been achieved, but that those capabilities have become passive. By passive I mean that they have become divorced from human activities, the specific kinds of things humans do, and instead rendered as a tool. Computer vision models, for example, have surpassed the limits of human eyesight and mental representation and are now deployed in a range of recognition systems. But these systems can only "see" things they have been told to look at. What we see when we use them is a combination of the limits of their computation and the features we interpret as meaningful.

Aristotle called this faculty *nous*: the active discernment of a horizon of possibilities in relation to some limit. These models are changing what it means to represent ourselves and our activities by enframing them against a backdrop of new features, but they cannot discern features of their environment that they were not told to attend to. In other words, they can only be put to use. In putting them to use, our passive intellectual capacities are expanded, and *nous* is reconstituted. Our own agency is reconstituted. Once *nous* has directed passive forms of intelligence to discern features, deliberation organizes possible actions. The organization at stake—of some particular task, the wider activity, or an entire domain—depends on what changes the agent in question is able to effect. It follows that AI models, as abstract capabilities, will entail different forms of deliberation depending on the agency at stake in their specification within actual systems.

To be clear, AI models' lack of agency does not mean that they are guaranteed to be safe. For example, there is a creeping awareness of the risk in developing tools without reflecting on the purpose that guides their use. At present, we are like a child who, having mastered the basic rules of arithmetic, has just realized the ability to manipulate numbers of arbitrary size. The child is now in a position to deterministically simulate and manipulate computation to his or her satisfaction. This is indeed a scary thought.

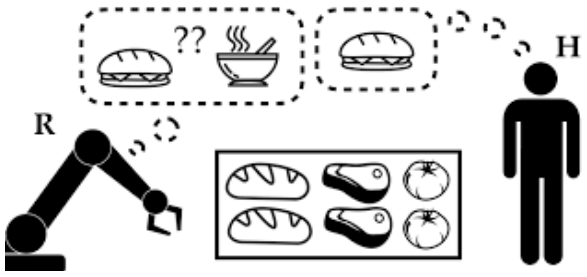
This problem is resolved through the child growing into a mathematician, someone who knows what the uses of mathematics are and what calculations are worth pursuing. Calculations not worth pursuing are not undertaken. All this means is that a mathematician is someone who has an organic relationship with the ability to perform mathematical operations, not someone who is arbitrarily good at doing math. Much of our anxiety about AI models rests on our growing awareness that we are pursuing calculations—or allowing others to pursue them—whose terms we do not understand or trust and yet whose results are reshaping the contours of our own lives.

1.3 Indeterminacy and Abstraction in AI Development

This anxiety may be further illustrated through examples of prominent technical research paradigms in the long- and short-term AI ethics landscape. AI Safety is a technical subfield that treats design predominantly as a long-term problem of controlling what artificial agents are able to learn and do, and is dedicated to the creation of "provably beneficial" systems. Fairness in Machine Learning (Fair ML) is a subfield that focuses on the short-term creation of algorithmic models that are fair, accountable, and transparent (Dean et al. 2021). At present, these fields struggle to technically specify their own normative aims. For example, many AI Safety researchers focus on a theorization of "artificial general intelligence" and attempt to forecast its likely arrival, but have yet to agree on a consistent definition of the term. What set of capabilities, skills, and proficiencies does it cover? Fair ML, meanwhile, is still digesting the uncomfortable notion that there is no single definition of fairness to work from: one can favor equality across selected groups, or equal odds of being selected within groups, or some other selection criteria, but it is a free parameter that our models cannot provide to us (Corbett-Davies and Goel 2018).

These long- and short-term considerations face the same problem: perfectly deterministic conceptions of fairness and safety do not exist outside mathematical formalism. Building systems to incorporate those models requires agreement on what features of fairness and safety in fact matter, requiring calculation that is organic and deliberative rather than strictly optimal. As such, the goal of these communities is to break down the abstractions they have chosen for themselves into **indeterminacies** that are technically tractable. By indeterminacies I mean the merely possible relationships that proposed models may have with actual existing systems (economic, behavioral, cognitive, environmental, juridical), which the formalism does not specify.

The method of crafting manipulable indeterminacies is best exemplified by the paper "Cooperative Inverse Reinforcement Learning" (CIRL). This paper asks the question of what it would mean to teach the structure of human preferences to a robot in the context of a specified task (Hadfield-Menell et al. 2016). How, for example, do I teach a robot not that coffee is valuable but that it is valuable *to me* in a particular way, so that it could learn how to make it in terms of the value I assign to it? The paper formalizes this question in terms of a game that is played between a human (whose "reward function" or goal is at stake) who demonstrates the task and a robot that observes the demonstrations, and thereby tries to learn what the human is trying to do. To illustrate this through a possible CIRL configuration: the human demonstrates, the robot mimics, the human demonstrates again in response to the robot's prior action, the robot mimics that, and so on. Because the robot is initially uncertain about the human's preference structure, this style of pedagogical interaction may continue as long as needed for the task to be satisfactorily learned.



A stylized view of the CIRL game in the context of making a meal.

The thesis of the CIRL paper is that the value alignment problem—the problem of getting machines to learn human values as we intend them to be learned—is best tackled not by specifying tasks in a way that can be passively observed, but by specifying roles for the human and robot to play in order to isolate what it would mean for the task to be learnable. In order for value alignment to work, the human has to become something that can pedagogically demonstrate the value of its own actions by framing them into a task that can be taught to an agent that does not (and need not) share or understand them. Value alignment, the problem of getting AI to learn the structure of human values, is thereby treated as a maieutic activity in its own right, based on the agent's compliance with the human's demonstrated behavior.

CIRL offers a possible route to structurally decoupling the content of what humans value—their place and shape in the context of human activities—from the learnability of actions that stand for them. The designer can, in effect, reason about actions outside the context of the domain in which the activity occurs. This makes the CIRL game a deliberative activity in the sense described earlier. It instantiates, in terms of an abstracted game, a structured set of encounters between a recursion (the robot) and a source of contingency (the human) *that must be organized* in order for values to be learned in a way that does justice to them.

In CIRL's terms, the path to AI Safety is not to build better sewing machines but for humans to become the tailors of their own learnability. Like a Platonic dialogue, it models a robot as Socrates and a human as someone who tries to provide a good definition of what he or she is doing. It is not an actual conversation between a human and a robot, but a model of a conversation that could in theory be optimized. And it positions the designer, who in this sense is "reading" the dialogue, as the hub of normative deliberation. Like a student of philosophy, the designer must ponder as deeply as possible what set of demonstrations would satisfy the robot, regardless of whether such a definition has ever been written down or even exists as a fact of the matter. The goal of the activity has now shifted to include how a determinate conception of the game itself could be articulated, to reincorporate the activity in terms of its learnability.

This entails a reinterpretation of the human. The human is modeled as something that has access to information about its own reward function (i.e. what it wants) and is able to rationally demonstrate states that minimize the robot's uncertainty about that reward function. The game itself is structured as a Partially Observable Markov Decision Process (POMDP), a decision procedure in which the relationship between successive states determined by chosen actions cannot be directly observed. Instead, a probability distribution over prior observations must be computed to minimize these uncertain conditions and thereby interpret demonstrated actions as representative of the activity.

This is a subtle but crucial point. The human is not assumed to have an optimal, expert-like grasp of its own behavior or reasons for doing things. If that were the case, no game would be needed—the robot could just passively observe the human's activity. Rather the human is interpreted as something that is able to play a CIRL game and respond to the robot's efforts in a meaningful way by selecting between the states that matter and those that don't with respect to task completion. It is not that the robot is trying to interpret the human, but that the human is trying to reinterpret itself in terms of the robot's performance. CIRL redefines the value of coffee as the state of play in

which the human stops demonstrating to the robot how to make coffee for it. And if the human for some reason fails to rationally demonstrate states, this could be resolved by altering the terms of the game (e.g. further subdividing the task, learning from the physical environment directly). In a meta-pedagogical sense, the designer could reinterpret the activity in the name of surmounting the limitations of the human's demonstrations.

CIRL renders concrete the core mystery of AI Safety: what must humans become in order for their values to be exhaustively learned by an artificial agent? The question isn't whether the CIRL game could be realistically played in a cognitive or institutional sense, or how various activities are or aren't in fact like the POMDP through which the game operates. It's how we would have to reorganize ourselves as sociotechnical beings such that the CIRL game actually could be played. The accomplishment of CIRL is that it concretizes these indeterminacies. It makes it possible to reimagine and enact human normativity in terms of the limits of the game.

In this way, the technical contributions of AI Safety and Fair ML are indeterminate abstractions: propositions that propose the existence of substances or entities whose content is implied but deferred pending further elaboration or examination. I draw inspiration here from Charles Sanders Peirce's closely-related concept of hypostatic abstraction. The point is not that what the abstraction implies (i.e. a pedagogic human or a deliberative designer) is assumed to be real. It is that treating it as possible allows one to perform a kind of controlled thought experiment that clarifies what it would take to confirm its existence.

This is not a critique of what technical communities are in fact doing, which is giving voice to basic human problems in the context of AI development. We want models that are fair. We want models that are safe. In sum, we want to implant our sense of what is good and true into the actual systems that will help organize our lives. But purpose is not something we put into tools. It is something we articulate to ourselves, and organically achieve, in the process of putting our tools to work. Deliberating on what it means to use them well is how we distinguish the means of achieving what we want from what we actually want, in ways that make what we want more clear. What makes CIRL profound is its conception of deliberation as a coordination problem between a human teacher and robot pupil who work to complete a shared task, entailing the "design [of] machines that provably converge to the right purpose as they go along" (Hadfield-Menell et al. 2016). That is the paper's pertinent contribution: to technically model indeterminate features of human activity, drawing our attention to them in ways that not only entail further inquiry but that could be actively structured to inquire in an organic way.

1.4 The Interpretation of "Structure"

CIRL sharpens the question from "how do we build a model we know would be safe?" to "how can we determine human activities as learnable through active demonstration under conditions of partial observability?" This new question stylizes, via the need to structure the game, both the designer as an interpreter of human activities and the agent as an interpreter of human actions. It makes it possible to ask: how should these interpretations be organized?

As a thought experiment, let's imagine a CIRL game being played between a human driver and his personally-owned AV. The driver demonstrates to the AV how he wishes it to cruise down a

highway, or pass through a four-way stop, or merge into traffic, or navigate around cyclists. Suppose these games are played successfully, in a way that remains physically safe to all participants. In theory, this would mean that the AV has successfully learned the precise value of driving tasks for the person whose utility is at stake. But to do this, roads themselves would have to be remade in a way that mirrored the terms of the game, generating externalities whose stakes would be impossible to take into account.

There are two features of roads, in terms of their **structure**, that illustrate this point. By "structure" I mean the organizing attributes or features of roads, rather than those known or used by agents at any particular moment (Zwetsloot and Dafoe 2019). One is that roads enact a city's geometry. A city's population density and transportation hubs reflect how life in that city has been ordered to support daily commutes, commercial activities, leisure, recreation, and residential areas (Walker 2012). Even local aspects of roads like sidewalks and street corners support this geometry: the width of sidewalks and gradient of street curbs reflect their frequency of use by particular kinds of agents, enacted in a set of ratios between the different parts of the road, how they have been patterned to support particular kinds of access. In fact, one hundred years ago, a major political dispute surrounding automobiles was the incentive they created to make sidewalks narrower in order to improve the efficiency of traffic lanes for vehicle throughput (Norton 2011).

The geometric design of roads at different abstraction layers, left to right: highway interchange, road slope (to shed rainwater), and sidewalk-curb contraction joints (to minimize cracking).

The second feature is that roads are public infrastructure. This means that roads are intended to be widely-shared, accessible, equitable, affordable, and capable of supporting diverse means of use (vehicles, pedestrians, cyclists, etc.). In other words, roads are a form of public space or "commons" whose value is not a function of individual utilities but of the fact that they are collectively owned and used. One example of this is that as roads are worn down through forms of unintended damage like potholes, everyone bears the cost of their repair through taxes, rather than just those people who are known to frequent the specific roads where potholes manifest. In other words, the externalities are publicly managed and paid for.

Insofar as they are geometrical and public, roads already have recursivity and contingency baked into them. That is, roads are a city's way of referring to its own demography, and they are understood as public as a fact of the matter by the people who use them. Drawing from our earlier discussion, deliberation about roads is a calculation about how to organize the traffic generated by the interaction between their geometry and status as a public good. Regular jams are perceived to be a problem not just because they inconvenience the drivers who happen to be stuck in them, but because they violate the integrity of particular thoroughfares. Frequent cases of vehicular collision

or manslaughter are perceived as failures of road design as well as inattentive drivers, resulting in calls for different signage or lighting conditions.

These problems are interpreted in ways that are structurally distinct from the thought experiment described earlier. Let's return to the two abstractions at stake in CIRL:

1. Human reward functions have learnability.
2. The human's demonstrations have partial observability.

Learnability means that the human has been positioned in such a way that its demonstrations provide information about the structure of the activity in question. Partial observability means that there is some determining ratio between the perceptible substance of the game (i.e. every action taken) and its purpose as enacted by the human, in terms of what the human wants to demonstrate. In other words, the activity has been discretized in order for it to be deterministically learned, and there is a distinction between the "signal" and the "noise" that can be discerned through observation. These abstractions respectively suggest that the activity itself has been restructured to make it possible for the game to be played, and for the game itself to be a meaningful coordination or orchestration of the activity. Again, the beauty of the paper is that it frames these indeterminacies as tractable while remaining agnostic about how to resolve them.

For CIRL games to be playable in the driving environment, existing roads and highways would require a vast new telecommunication infrastructure between AVs, roads, and perhaps even other cars and road users. Without sophisticated sensors providing real-time data and integrated analytics about the environment, the driver behaviors would not even be partially observable. Furthermore, for demonstrated behaviors to be learnable, the game would have to introduce new protocols for role-playing and discretized driving tasks whose specification would have an unclear relationship with existing social norms about the rules of the road. As an example of the latter, different road users will pay extra attention when navigating intersections that are damaged or compromised, because we understand what roads are for and how they are supposed to work. That is, our transportation behaviors are organic with respect to roads' intended structural features rather than optimal with respect to locally observable ones.

Consequently, the game-theoretic assumptions at stake in the criteria of the CIRL game constitute a separate axis of deliberation from the kind that organizes how traffic works. A problem will arise once CIRL games begin to intervene on the structural criteria of roads themselves. Which should take priority? Learnability introduces an algorithmic geometry to traffic activities that is incommensurate with the one already physically baked into roads, while observability assumes proprietary metrics for demonstration that are incommensurate with roads as a commons. The CIRL criteria would therefore require new standards for measuring and certifying the conditions under which personalized AV behaviors are enacted, apart from legacy forms of traffic management.

To counter these concerns, we could try to integrate the CIRL game at higher levels of abstraction than demonstrations from individual drivers. For example, AV planners could provide demonstrations for an entire fleet to learn from, streamlining how a company like Waymo offers its services to a particular city or region. Companies that sold vehicles to individuals could pursue

"CIRL certifications" in order for personal demonstrations to be indexed or bound in a way that municipal or federal bodies could oversee. In such cases, CIRL demonstrations could be optimally customizable without being undertaken in a normative vacuum.

Even so, these strategies would redefine roads in terms of their utility and resource scarcity rather than their structural integrity. They would defer the basic question of what roads are supposed to be (vs. what they happen to appear to be) in the context of a specified task. Because roads are public, there is always more that can occur on them or near them than designers will know how to optimize. And because they are geometrical, they reflect a shared normative sense of the cityscape that citizens, not AV designers, are positioned to affirm or reject. What if the city of Houston wants to minimize vehicle access to downtown for sake of greater livability, or Chicagoland wants to use AV services to connect wealthy suburbs with the south side? Aristotle's term for this problem space is *politeia*, the constitutional form that denotes how a city-state or polity is meant to be ordered, in terms of how it governs itself (Winner 2010). Deliberation about roads assumes *politeia*, our sense of what activities are proper on them and of what domains they are meant to connect once enacted.

What CIRL does is augment that deliberation by expanding the interpretive possibilities of the domain via new forms of task specification. In theory, it enables us to specify new tasks (and corresponding activities) at the same time we pursue proprietary efficiency and safety. We are called to ask *by what standard* CIRL games should be conceived, authorized, and played, so that designers' deliberation about how to structure them does not overwhelm our sense of what roads are for in the first place. This question will have to be answered through new performance metrics and certification standards so that CIRL games in fact "count off" the units of activity whose relations we affirm as constitutive of roads' public geometry (Klein 1968). CIRL's deeper, philosophical value is thus to model how we might restructure domains in terms of how we want them to work, not to guarantee utility to individual humans.

1.5 From Artificial Intelligence to Sociotechnical Specification

Until now our discussion of models has revolved around two questions:

- What features of human activity can be automated using AI?
- How should AI be used to represent the structure of a domain?

These questions bear on each other. Automated feature detection for faces, game-playing, and text generation demands a restructured sense of what those domains amount to and what they mean to us. Meanwhile, our willingness to pursue the creation of AV fleets raises new questions about which features of road activity like traffic are essential and which are incidental. These are epistemic and normative dimensions of the same problem: the need to organize how AI can be used to illuminate human activities.

This is a wholly different problem than the one faced by Hubert Dreyfus, whose highly influential critique of logic-based or "symbolic" AI continues to shape public and scholarly debate (Dreyfus 1992, 2014). According to Dreyfus, the question that symbolic AI failed to answer is: how does man exhibit intelligent behavior? His own answer, developed over decades of philosophical

inquiry, was skillful coping. That is, humans learn to make sense of their own embodiment and situatedness, which organize their place in the world and what it means to engage in a given activity. On this interpretation, performing given tasks like driving, cooking, and exercising in an everyday, naturalized sense is the most complete expression of what it means to be intelligent.

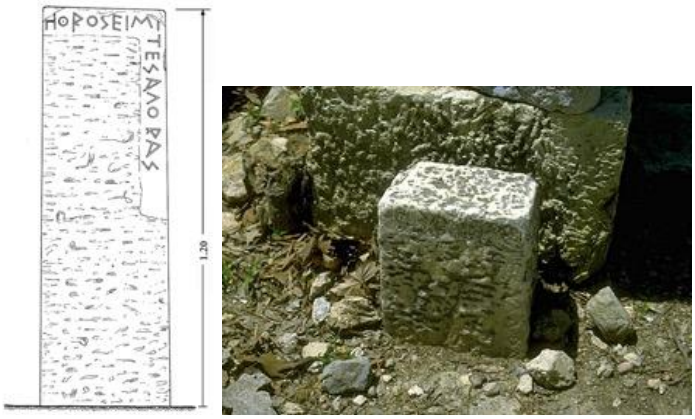
But the question raised by CIRL is: against what spatio-temporal horizon is a given activity conducted? The answer, which I claim must be developed to meet AI's future design challenges, is what I call **sociotechnical specification**. That is, humans organize themselves with the aid of tools, which enframe possible relationships between activities that in turn must be defined. For some activity to be able to be undertaken in an everyday sense, there must be a set of coordinates to mark its relative scale and significance. Tools make those coordinates possible, without telling us how to specify them. In the technical language of reinforcement learning (and CIRL), this horizon comprises a tuple including a transition function and state-action space, against which rewards can be observed and predicted by some agent. Accordingly, sociotechnical specification marks the essential features of related activities within a domain's functional whole.

The difference between these two positions mirrors larger shifts in the technical development of AI. Dreyfus gave an early philosophical voice to the problem of human intelligence by asking how humans become proficient and what it would take to mechanistically simulate this through reference to performance thresholds. In many ways we are still living in this paradigm, as optimization problems define what it means to automate skillful performance of a given activity across technical subfields. Public discussions remain fixated on when AI will meet or surpass human capabilities, how neural nets draw inspiration from the human brain, whether or in what ways AVs will be safer than human drivers, and which sorts of human jobs are likely to be automated next. In all these cases there is an assumed 1-1 correspondence between the models we are building and proficient human behaviors.

But today that paradigm is breaking due to growing problems of specification, the pressing need to define the interface between AI models and social reality. We are now able to inquire into our own behaviors on scales of abstraction and computation whose terms are no longer clear. For example, the activity of driving is composed of certain features like turning, accelerating, braking, signaling, honking, stopping, and parking. But driving itself is a feature of roads, which also include walking, cycling, jogging, transiting, taxiing, or just riding. And roads themselves are a feature of cities, along with homes, neighborhoods, businesses, downtowns, parks, and suburbs. Specifying driving means that its parts are being arranged to stand within a structured whole. A model of driving whose implementation would violate that whole reconstitutes the domain, and hence would compel other activities in that domain to conform to its specification. Driving, road use, and urban living are activities at different scales whose relationship is made technically tractable and manipulable through automation. In this way, sociotechnical specification poses a problem of modeling these activities' structure such that designers know that those relations will remain functional as they are technically integrated.

A related term in Aristotle's philosophy, *horos*, will help us make sense of this problem. A *horos* is a boundary, something that is discerned by *nous* as a limit to the horizon of a given activity and enacted by *politeia* as a delimiting factor in how activities are organized in relation to each other. We encountered these terms in our previous investigations of the limits of algorithmic computation

(Section 1.2) and the structural integrity of roads (Section 1.4). Aristotle uses *horos* both to refer to the ratio between matter and form discerned by *nous* and the standard by which *politeia* structures how people live together: "[T]he standard of aristocracy is merit, of oligarchy wealth" (1998). In this shared sense, *horoi* are definitions that comprise the terms and conditions of how organic human beings mark the structure of activities in relation to each other. By placing boundaries on the way an AI model is able to signify some activity, choices must be made about how to modulate and defer meaning within the emergent sociotechnical setting enacted through the system's operation.



Agora boundary stones, found east of the Tholos (left) and in its original position in Athens (right). Rough translation: "I am the boundary of the Agora".

Technical work in AI Safety, including CIRL itself, is an exercise in *nous*: active inquiry into what first principles are needed for the development of AI models to be demonstrably aligned with human values. But this suspends and defers *politeia*, the whole institutional order of social and political relationships, in ways that require further definition. Sociotechnical specification, as the mature expression of *nous*, discerns *horoi* that indicate how a given activity is structured in relation to other activities. The *horoi* are how we reconcile the discernment of features and the enactment of optimization techniques, in terms of what it means to live well together.

Famously, Dreyfus argued that AI was nowhere near matching the practical skill displayed in everyday activities because it did not inhabit the world in the way that humans do. But the CIRL formalism shows that the goal of AI research is no longer to mimic our worldly coping directly, but to interpret it as something that could be modeled as teachable to a machine. Learned models alter how we discern particular activities as parts in relation to a possible whole. The question is not just how to optimize the performance of individual activities but how their automation compels us to redefine their essential features. Answering this requires a grasp of that activity in relation to other activities, and how to structure that relation well.

1.6 Normative Cybernetics: Models vs. Systems

So what does all this have to do with AVs?

So far I have been referring to AI almost exclusively in terms of models. Models are stylized conceptions of reality whose purpose is to clarify the terms of that reality. Our discussion of

sociotechnical specification leads to further deep questions: what phenomena *should* AI models represent? And how can we develop those models in an appropriate way if it is those very models that are the entrypoint for deliberation itself?

While designers must represent the domain in terms of their AI model, they do not do so on terms of their own choosing. The models being built for self-driving cars to route through cities reflect some business model of how the AV company represents its own market position. That business model exists in some relationship with how the law represents private business activities. Consumers, in turn, are shown advertisements that represent vehicle services as worthwhile or not, based on what they want. All of these representations help define how the transportation system is able to function in relation to other systems. The designer is only in a position to design how their system operates with respect to the physical, cognitive, legal, or managerial substrates on which the computation must operate. They have to work within them or in relation to them, often without having a clear or expert understanding of what they are or how they function. By contrast, the most basic choices a transportation planner faces are not of what domain features matter most, nor of how well those features can be modeled, but of how what is being modeled will affect the relations between other systems; that is, to balance or vary the forms of feedback between them in ways that are organically desirable.

In my experience, many AI designers approach these other systems as having nothing to do with their model. They interpret them like weather patterns: natural or incidental features of the world that define the boundary conditions of the AI model and which it does not have to take into account in order for it to perform optimally. It is easy to forget that things like stop signs are not natural features of the environment, but artificially constructed entities that are meant to help other systems talk to each other. They are patterned to provide feedback in ways that structure the domain to work well. In other words, they are cybernetic entities that have been intentionally implemented by planners to help keep roads public. This implies the need for what I call **normative cybernetics**: structuring feedback between activities to guarantee the integrity of the domain. Normative cybernetics begins with an investigation of the types of feedback needed to monitor how an AI system might interact with a domain, before evaluating effective performance. Absent this investigation within a concrete sociotechnical context, the notion of "optimal performance" will lack significance and fail to be given proper meaning.

Let's review what a stop sign is, in terms of its features. First, what stands out about a stop sign is its redness. This makes it easily noticeable to the human eye from a distance, and (invariantly across national contexts) denotes a state of alertness and readiness as well as danger. The redness of the stop sign means: *pay attention*. Second, a stop sign is an octagon. From a certain distance, this makes it distinctly unlike all other forms of signage, which take the shape of other geometric objects. The octagon-ness of the stop sign means: *this is not a* YIELD*,* NO TURN*, or* DO NOT ENTER *intersection*. Finally, a stop sign literally says STOP. That makes it an indicator for what specific kinds of road users (namely drivers) are supposed to do at the intersection as well as the resulting consequences, both to them and to all other road users. The STOP-ness of the stop sign means: *vehicles must completely halt before the white line*.

It is true but misleading to say that novice human drivers learn to recognize stop signs. What they actually do is learn how to recognize them in a patterned way that structures their activity as drivers

to work in support of desired features of roads. All this is encoded in the sign in terms of its color, shape, and protocol. In other words, the mode, pattern, structure, and equity of road access are cybernetically encoded in the object in order to help make the domain work well for different types of activity.

What really is a stop sign? It is a *horos*, a mark of social order. The primary purpose of a stop sign is not to help structure four-way stops so that cars can navigate them in a safe and efficient manner. In other words, it is not merely a means to traffic optimization. Rather, a stop sign is what constitutes particular types of road intersection as stops (four-way or otherwise). It reflects a pattern of decision-making, made by planners at a higher level of abstraction, about what kind of road this is and how vehicle activity needs to be structured to be permissible in relation to it. The purpose of a stop sign is to provide feedback about what vehicles can or cannot do with respect to other activities. In other words, it is a form of road specification. It helps make cars safe for roads rather than roads safe for cars. Safety originates as an emergent property of how roads structure traffic to be stable, not as a design parameter for individual elements of traffic.

Horoi in action: patterned forms of feedback that structure activities to support equal road access.

AV fleets are not like a mass of individual human drivers who happen to have been automated. They are more like an independent mechanism for traffic itself, and are going to change the way traffic works once fully deployed. They are going to change how roads work, will require new types of feedback to monitor their real-time performance, and will restructure traffic by changing the behavioral incentives of other road users. They will redefine what it means for roads to be safe. In brief, AV fleets will be much more like stop signs than like individual drivers, and the real cybernetic question is figuring out what that means.

To understand this, let us return to the key features of human vision and action (color, geometry, protocol) that stop signs help organize and that AV fleets will explode. First, AVs navigate their environment through the use of computer vision algorithms that classify objects based on inputs provided through LiDAR sensors. LiDAR sensors are comparable to human vision in their sensitivity to particular types of light and ability to evaluate distance from objects, yet also incommensurable in their acuity and modularity. AVs often have several LiDAR sensors placed all around the vehicle, of different types of sensitivity, looking backward and forward and sideways at once. AVs leverage sensory affordances from the environment that overlap with but also surpass what it means (in human terms) to perceive and experience colors.

Second, AVs navigate their environment through the use of routing algorithms that make travel plans based on coordinates that have been geofenced. A geofence is a virtual or simulated perimeter that is comparable to the real-world geographic areas navigated by individual human drivers, yet also incommensurable in who is responsible for bounding it. Instead of individual road users, these bounds are determined by standing agreements with municipal planners, businesses, government agencies, and anyone else with whom the AV service provider chooses to partner. AVs leverage geometric affordances from the environment that overlap with but also truncate and artificially delimit what it means to navigate street traffic.

Third, AVs orchestrate their environment through the use of application program interfaces that establish forms of data sharing based on information protocols. These protocols conform to the "smart" infrastructural capacities and business model of the AV fleet in question, comparable to the services provided by taxis and human-driven ride sharing fleets (e.g. Uber, Lyft), yet also incommensurable in their scale of traffic optimization. Simulated results already show that AVs would be able to control highway bottlenecks once they reach 10% of all vehicle traffic, acting as moving road obstacles that would restructure the behavior patterns of other road users (Vinitsky 2018). AVs leverage informational affordances from the environment that both interact with and reconstitute the flow of traffic on any particular stretch of road.

AVs' perception, routing, and coordination modules appropriate and explode the features of stop signs without re-patterning them to reflect how we want roads (and cities) to work. The problem is that unlike stop signs, AVs at present do not meaningfully restructure the roads they drive on to remain equal and equitable. A single human driver, no matter how experienced, is not in a position to decide how traffic works, let alone what traffic means with respect to other systems. But AVs introduce new mechanisms for object detection, route navigation, and traffic flow, in ways that are nothing like someone learning to drive or indeed someone who has been driving for decades. In all these ways, the kinds of affordances stop signs provide are reconstituted by AVs. And companies like Waymo and Tesla are preparing themselves to fill that position. In fact, it would be more accurate to say that these companies are building privatized roads than automated vehicles. That is, they are optimizing the way roads work according to legacy standards without any proportionate deliberation by planners for what roads are meant to be.

A network of AVs acting as a traffic management system via feedback.

AVs, along with many other AI systems, displace *horoi* and in so doing reconstitute the relationship between *nous* and *politeia*. *Politeia*, the constitutional social order, is becoming conditioned on *nous* and deferred to a degree that is new and in ways that are new. The existential problem faced by tomorrow's traffic planner is not how to ensure AVs can recognize stop signs,

but to ask: what are roads supposed to do and not do? And the series of choices that are faced are about how to condition the ways that roads can be modeled on what roads are supposed to be.

What I am describing is an institutional, rather than strictly computational, basis for the encounter between recursivity and contingency. Designers cannot control the structural (vs. computable) contingencies with which their systems interact. But city planners can articulate new rules or standards or types of feedback that structure how that interaction happens. They could, for example, support financial incentives at the federal level for AV startups and research labs to compete on achieving desired performance metrics for safety and routing efficiency, rather than what happens to pass legal muster. Or cities could certify particular AV companies or firms as worthy of acting in the public interest and form special relationships with them about where they are allowed to test fleets and provide privileged terms of service. Or state-level Departments of Motor Vehicles could structure periods of public comment on proposed performance metrics if it is unclear which should be prioritized. Any of these is preferable to surveying individuals about how they would solve trolley problems, because what really matters is how we redesign roads as a whole, not ethically resolving corner cases pertaining to crash scenarios. Do we want to live in cities where pedestrians can cross the street without looking? Or in which AVs become so accessible that the concept of "pedestrian" is abandoned as outdated? This is the problem space of political economy: structuring the encounter between recursivity and contingency at distinct institutional levels, of actual systems rather than toy models.

1.7 Machine Politics: Systems vs. Domains

In the process of sociotechnical specification, there is a need to track the ways we reinterpret corresponding activities. This tracking is cybernetically orchestrated through feedback. But there is also a need to resolve how we want the domain to work, once its indeterminacies have been framed by the model and made actionable by the system's operation. The case of prioritizing or wholly abandoning "pedestrians" discussed above is one example of this indeterminacy.

Let's appraise the challenge faced by AV designers and traffic planners. Road models can be structured through reference to the behavior of other drivers, pedestrians, animals, weather effects, and other features of the environment. Meaning can be assigned to these by discretizing the environment into states, actions, and rewards for the AV model to learn. And we can scale both of these with forms of geofencing and data protocols in order to build actual AV systems. But it is actually really hard to do all of this at once, let alone when public infrastructure is at stake. Help is needed. To sociotechnically re-specify and reconstitute an entire transit domain, we need to ensure that AVs learn to navigate these environments according to the terms that we want. This includes new forms of public comment that are proportionate to the choices designers and planners are faced with. At least some of the criteria for object detection, route navigation, and traffic flow will need to be provided by a public utilities commission.

Consider the example of an AV fleet that has demonstrated reliable and safe transportation within a suburb for several years. The firm that operates the fleet proposes to expand service and provide rapid, reliable access to downtown, including luxury shopping venues. Various business interests quickly hop on board and endorse the proposal. However, some municipal planners voice concern that the expansion may interfere with busing options whose routes already operate along parts of

the proposed route. Those services are relied on by low-income commuters who have few other means for accessing jobs and essential shopping like groceries, and risk sliding into poverty if service is disrupted. The firm counters by proposing a more ambitious, integrated AV service that incorporates legacy bus transit routes alongside simultaneous access to downtown for suburban residents. However, questions about the proposal's relative affordability, logistical feasibility, and political viability remain unclear.

What should be done in this situation? The proposal promises to incorporate aspects of the city's transit geometry which previously were impossible to technically enact, introducing a new means of organizing city traffic. However, there are emergent trade offs related to service coverage and the economics of ridership whose criteria have never been exhaustively specified yet must be taken into account somehow. Deliberation about the encounter between AV fleet routing (as recursivity) and legacy forms of public transit (as a contingent feature of the domain) is needed. For example, should planners propose rebuilding highways with specialized AV lanes to help guarantee traffic flow? Or restructure the AV service with direct municipal oversight? Or something else? These are not academic questions when salient communities may be undone as a result of poor choices. But they also cannot be answered without affirming the rights and needs of those communities on new ground. Maintaining an active sense of normative proportion between what we now are and what we might become as AI systems are developed is a central deliberative component of machine ethics.

I name this mode of deliberation **machine politics** (Hui 2017). By this term I intend to capture both the range of technical choices about how to reorganize the domain, and the organized human constituencies that enact a particular choice based on desired political ends. This mode is activated in situations in which at least some requisite criteria for system development are absent, requiring basic reflection and decisions about the domain's purpose and functioning. Machine politics entails evaluating the encounters between the domain and the AI system as the latter is built, in order to track the changes to the domain its operation may instantiate. Planners must ask: 1) What new transit geometries would the widened AV service make possible? 2) How would these geometries compel distinct reorganizations of service and land use already present in the transit domain? 3) What modes of public engagement and iterative transit planning could highlight the features to be affirmed or rejected as the domain is reorganized?

It is worth clarifying that AVs do not constitute an entirely new mode of social organization. We already have highways, and signage, and residential streets, and human drivers, and downtowns, and pedestrians, and neighborhoods, among many other well-defined features. AV fleets will coordinate these elements in ways that may be centralized (via some global utility function) or decentralized (via smaller fleets that only communicate locally) based on how residents wish to reorganize them. But either way, we know how to measure resulting effects and could reasonably ask what it would mean to control for them, enforce them, and manage them. We can do this because the social activities at stake are already adequately framed so that we know how to go about inquiring into the effects on them of the system we are implementing. We don't know what a fully-smart city that has integrated self-driving car fleets would look or feel like precisely, but we at least know what it would mean to investigate it. And that means we know where we would need to look for the criteria upon which our models would need to rest.

AV developers must frame specification options such that the utilities commission is able to examine and supply the criteria needed to choose between them. Technical proposals for vehicle performance, service range, geofencing, and infrastructure costs must be narrated in light of possible orderings of the city that citizens (or their representatives) can affirm or reject. Unlike the modeling of features or monitoring of feedback, this is an explicitly political problem space, which AI designers must steward without commandeering. New modes of public inquiry are needed to acquire a sense of proportion between the prior structure of the domain and how that structure should be reconstituted to support human flourishing. Only in this way could planners distinguish the desired system performances and behaviors from those not desired.

Two visions of Las Vegas transit: underground tunnels (left) vs. Medical District access (right).

We must learn to see self-driving cars not as novice dancers that need to be taught how to move, but as a choreography for the way cities work whose notation and sequencing are to some extent unprecedented. It is a problem of leveraging AVs as a basis from which to interpret what cities could be made to be in relation to what they now are, more than it is a problem of model fitting within a predetermined structure. Structuring public inquiry into a given system's effects is about reforming how we mark the world in terms of how we want to live based on the limits of what we can discern. Machine politics is about trying to do this well.

Machine politics is fallibilist by nature. Designers cannot fully anticipate how AI systems will change a given domain, because they will reconstitute that horizon of anticipation to begin with. And planners need to acquire a sense of what proposed system effects it is within their agency to completely account for and relay to stakeholders. Moreover, the structure of many domains is considerably uneven and non-deterministic to begin with. For these reasons, those whose needs are met by the domain must be invited to inform and structure the regulation of how a given AI service is provided. This amounts to developing AI systems in ways reflective of what it would mean to deliberate about them, to structure the findings of *nous* to work in support of *politeia*.

With AVs, the situation is not hopeless. Intuitively, most people understand that the purpose of stop signs is to keep them safe in public, not to optimize traffic. That means that most people are in a position to reason about what stop signs are for, and hence what changes we could make to the way transit works to restructure roadway activities so that the optimization of traffic does not come at the expense of public needs. Most people may not be in a position to help make AVs work in a mechanical sense, but the bottleneck is that neither designers nor planners know how to organically integrate them within traffic. At present, due more to lack of political will and commitment than technical know-how, we have not specified the problem of AVs to accommodate organic and deliberative modes of public inquiry. Given that reality, the call is to invite those who

are in a position to deliberate to actually do so. Machine politics requires us to directly and actively engage each other as political beings, rather than to passively observe each other so that our behaviors are easier to model.

1.8 The Structure of the Dissertation

For any AI system there are three requisite modes of deliberation.

1. About the model: what are the key features to be represented at different abstraction layers?
2. About the system: what types of feedback are needed to protect the domain's integrity?
3. About the domain: how must we inquire in order to regulate service over time?

The first question has been addressed in Section 1.5 through reference to sociotechnical specification, the identification of significant features at interrelated scales of a structured domain. The second question, addressed in Section 1.6, is the problem space of normative cybernetics, comprising how we guarantee the domain's integrity in the face of system operation. The third question, which Section 1.7 named machine politics, denotes the process of reevaluating the domain itself in the context of AI development and regulation, based on a substantive grasp of what capabilities we want a given system to enact for us.

Normative problem	Pertaining to	Aristotelian term	Mode of deliberation	Deliberative agent
Structure	Model	Nous	Sociotechnical specification	AV designer
Feedback	System	Horos	Normative cybernetics	Transportation planner
Indeterminacy	Domain	Politeia	Machine politics	Public utilities commission

Dimensions of deliberation in the context of autonomous vehicles.

This introduction has provided a partial grammar for this project by addressing confusions present in the current technical and philosophical literature. I first presented **feedback** as a problem neglected by two types of AI ethicists, namely transhumanists (who wish to transcend the limits of human agency using AI) and critics (who wish to protect human agency from AI's encroachment). Next, I portrayed technology as a kind of external organ that reconstitutes human agency, and **deliberation** as the structured use of that agency. Third, I diagnosed various **indeterminacies** present in current debates on safety and fairness, before showing how good technical work frames them in terms of relationships that can be manipulated by the designer. Fourth, I noted pre-existing forms of **structure** that the designer can neither ignore nor remake, pointing to the need for criteria that are external to the designer's interpretation of the formalism. Fifth, in contrast to Dreyfus's critique of symbolic AI, I argued that **sociotechnical specification** is how AI designers must deliberate to distinguish the core features of related activities. Sixth, I defined **normative cybernetics** as the problem of structuring feedback to preserve a given domain's normative integrity. Seventh, I outlined **machine politics** as a distinct mode of deliberation, rooted in codifying regulatory procedures that would help distinguish desirable and undesirable AI systems, based on how we may want domains to work.

2. PROXY METRICS FOR THE BROADER IMPACTS OF AUTONOMOUS VEHICLES

In their ability to optimize the local safety and efficiency of individual vehicles, AVs promise to make individual transportation more predictable and reliable. Trips that people find too tedious to make could be made into trips worth taking, and as this change is reflected through the broader population it has the potential to fundamentally change the relationship consumers have with transportation (Fagnant and Kockelman 2015; Millard-Ball 2018; Stocker and Shaheen 2018). AVs also make it possible to centralize and coordinate the routing of vehicles. At the most local level we can see coordinated routing through the large body of work in platooning (Bergenhem et al. 2012; Wu et al. 2017c). Prior work has extended this to a larger scale showing how these new affordances can be used to control for larger scale effects in the transit system, alleviating traffic congestion by dampening the propagation of shock waves (Wu et al. 2017a). Such works represent only the beginnings of what could be possible. Centralized route planning could allow load-balancing between routes on the scale of cities (Tiba et al. 2020; Jiang et al. 2016), the predictive placement of vehicles for the purposes of ride-sharing (Miller and How 2017), special routing considerations for emergency vehi- cles (Konrardy et al. 2018), and the management of interactions between these considerations.

These new possible interventions pose a problem of *sociotechnical specification*: the need for designers to articulate the essential features of driving and other transit activities at particular scales of abstraction. This problem is sociotechnical to the extent that the interface between abstract model features and social reality cannot be assumed in advance. Rather, choices about which features are modeled and against what horizon(s) AV routes may be planned will reorganize the transit system itself. This includes the incentives and resultant behaviors of other road users, types and extent of repairs needed for public roads, emergent economic and environmental dynamics, among many other effects. In order for AV fleets to be safe as well as robustly beneficial, designers will need to control for these effects and deliberate about which types and scales of intervention are prudent with respect to the desires and needs of actual communities.

Many understudied opportunities exist to model externalities and optimize for those which result in desirable macroscopic outcomes. These effects extend far beyond the local effects to safety and efficiency, and could lead to large-scale changes to physical and economic mobility (Greenblatt and Shaheen 2015), pedestrian safety (Wang et al. 2019; Millard- Ball 2018), congestion (Simoni et al. 2019), pollution (Morrow et al. 2014), CO2 emissions (Greenblatt and Saxena 2015), and many more (Elefteriadou et al. 2012) as well as affecting how these externalities are distributed across society (Brandão et al. 2020; Fleetwood 2017). While these externalities have been studied by social scientists (Bonnefon, Shariff, and Rahwan 2016; Hancock, Nourbakhsh, and Stewart 2019), civil engineers (Sousa et al. 2018), and environmental scientists (Miller and Heard 2016), there is little technical research which can account for and mitigate these effects as part of AV design.

This gap between interdisciplinary insights and technical tools is analogous to the situation faced by development economists decades ago (Sen 1983, 1988; Ul Haq 1995). In that case, there was a concrete need for practical metrics to quantify the impact of a particular strategy, new models to implement those strategies, and ways to compare the viability of strategies in particular national contexts. This need led to new tools like the Human Development Index, which incorporated metrics for life expectancy, education, and standard of living (Anand and Sen 1994). While

imperfect, these tools are widely understood to have reoriented the field away from a narrow focus on GDP towards "people-centered development", improving the lives of millions around the world (Haq and Ponzio 2008; Snchez 2000).

AV developers now find themselves at a similar crossroads. While many needed metrics remain informal, imprecise, or opaque, multiple disciplines have made headway on identifying proxies that are suitable for modeling. Our goal is to provide impetus to an "AV Development Index" that would include these and other proxy metrics to serve as ongoing targets for optimization. I consider six distinct types of externalities (physical and economic mobility, environmental effects, local community needs, infrastructure, commercial activities, traffic laws), present coarse-grained models to control for them, and motivate additional proxy metrics that would refine these models. I also highlight other known impacts of AVs that require further scrutiny before they can be framed in terms of well-defined externalities. I do not intend any of these metrics to be final, but invite the research community to refine and build upon them through ongoing deliberation about scales of measurement and evaluation, in order to sociotechnically specify features in appropriate ways. It is my hope that the collaborative development and optimization of these proxy metrics between domain experts and AV researchers can better position us to take full advantage of the opportunity AVs provide. In Section 2 I give a broad overview of the problem space. In Section 3 I identify how some known externalities of AVs can be modeled to produce well-specified technical problems.

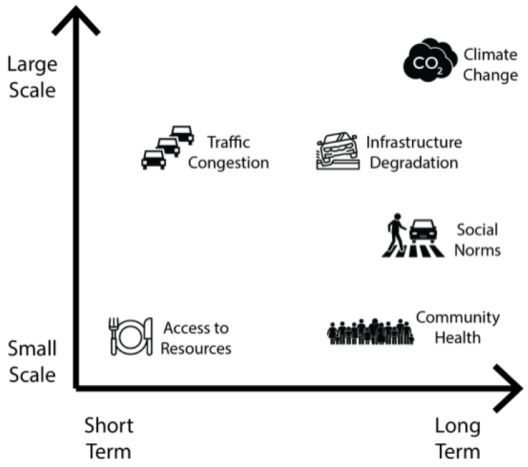
2.1 The Space of Externalities

Most work ensuring beneficial impacts of AVs has focused on the safety of the people in and around the vehicle (Kalra and Paddock 2016; Koopman and Wagner 2017; McAllister et al. 2017), which has the potential to save tens of thousands of lives annually. However, AVs will also reshape the wider transportation system, and in turn, have ripple effects to other dynamics beyond local interactions. There has been limited technical work in modeling these large scale effects. Technical and policy work on the effects of local behavior of self-driving cars on traffic patterns has revealed that default approaches to routing will cause large-scale congestion (Vinitsky et al. 2018; Litman 2017; Metz 2018; Van den Berg and Verhoef 2016). Once identified, technical solutions were designed to mitigate these effects, which allow for local routing solutions with good aggregate behavior according to specified performance tradeoffs (Wu et al. 2017c,b; Lin and Ho 2019; Lee et al. 2020; Rossi et al. 2018; Levin 2017). To mitigate the externality of large scale traffic congestion, the first step was to identify it as an externality, model the effects, and construct some control for those effects. Systems have caused measurable and preventable harm in cases where effects were not identified via appropriate metrics, modeled successfully, or controlled.

Of course, the relationship between vehicle automation and transportation externalities is deeply reciprocal. Few of the externalities at stake–traffic, carbon pollution, road wear–are specific to *autonomous* vehicles, and instead are longstanding features of transportation systems. In such cases, it is important to leverage AV behavior policies as an opportunity to deliberate about the underlying structural dynamics of traffic, rather than optimize AVs to operate in a dysfunctional traffic environment. This conceptual gap between desired policies and present or anticipated structural features is the problem space of sociotechnical specification. For example in (Wu et al. 2017c), the goal was to coordinate AVs to mitigate an existing negative externality (traffic congestion) present in the shared public good of road-ways, rather than to prevent an externality

generated by AVs themselves. Yet these same connected AVs could cause negative effects of their own: commuters living farther away from work (increasing urban sprawl), or traffic congestion induced by new youth and senior users of the road network (whereas they previously were unable to drive).

Similar complexities have been examined in the tradition of adaptive cruise control (Shladover, Su, and Lu 2012; Ioannou and Stefanovic 2005), the economic and environmental impacts of platooning (Alam et al. 2015; Besselink et al. 2016), and work on control system architectures for automated highway systems (Hedrick, Tomizuka, and Varaiya 1994; Chang et al. 1993). A common theme in this work is attention to the multiple scales (environmental, economic, infrastructural) on which vehicle performance must be measured and evaluated, and the need to coordinate metrics reflecting the fundamentally distinct stakes involved in defining good performance. Drawing inspiration from this, our distinctive contribution is to frame the dynamic relationship between AVs and the wider transportation system as *in scope* for AV development, moving beyond local interventions and leveraging new metrics to both better understand the systems that already exist and control for the effects that AVs have on them. To this end, we hope to differentiate the scales at which different sorts of externalities (either canonical to the transportation system, or original to AVs) manifest, reflected in distinct metrics for routing performance.

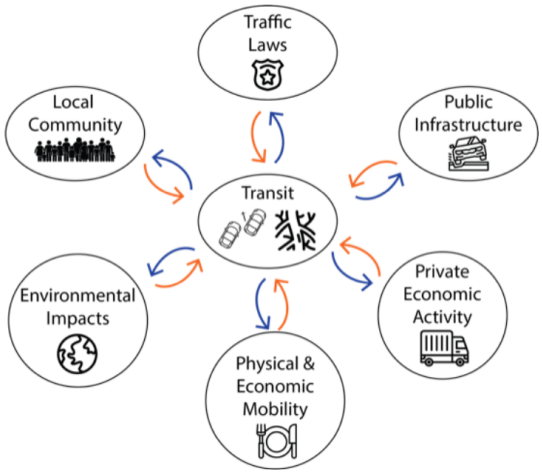


Considerations about the transit system.

Previous technical work on AVs has focused on one-to-one vehicle interactions within the transportation system, which we refer to as the *local effects* described by the left column of Figure 1. These include how vehicles interact with each other and the surrounding material infrastructure, as well as the evaluation of AV performance on those metrics. Instead, we focus on the effects including and in some cases beyond the entire transportation system, which we refer to broadly as *global effects*. These are effects that changes to transportation have on the parts of systems (geographic, economic, social, environmental) of concern to transportation planners and researchers. In local interactions there will be dramatic effects on how individuals experience roadways – how pedestrians signal to cross the street, how drivers signal to change lanes, who is blamed when a crash occurs, and the general norms of how people use the roads. Considerations

about how assertive or passive cars are today could have long-run effects on the sorts of driving strategies that are successful for the remaining human drivers.

We represent local and global effects in a continuous spectrum of technical problems, which may share relevant metrics if they fall on similar parts of this continuum. Still, there is less work on mitigating distinctively global effects, or articulating benchmarks to distinguish local and global scales of vehicle interaction. AVs' ability to coordinate on large scales will likely have significant effects on road conditions and will actively intervene on physical mobility, safety, and comfort of different communities. These effects are wide ranging but have been a focus of study in other fields, including environmental studies and the social sciences (Miller and Heard 2016; Levy 2015). As a result many models have been made of these effects, and working to mitigate them is a tractable research problem with the tools available today.



Systems with which the transit system interacts.

We should expect effects like these to arise from how widespread AV adoption interacts with the existing structure of the transit domain. Figure 2 outlines a rough sociotechnical specification of the distinct categories of these effects, representing the various subsystems involved in transit and imagining how decision-making could change if these were automated. For example, since transportation is central to many other social systems (recreation, work, residential life, economic mobility), interventions would have ripple effects on distinct aspects of society. Moreover, Figure 2 helps to distinguish short term effects, in which AVs affect surrounding systems by treating them as static, and long term ones, in which surrounding systems adapt to AVs in a dynamic fashion. This is illustrated via distinct causal arrows between transit and other systems. The next section follows this pattern by presenting short and long term effects of each externality respectively in terms of the static and dynamic relationship between AV transit and surrounding systems.

Introducing AVs is not as simple as making all of the trips that we currently take more efficient and safe. It will affect traffic patterns, which will in turn change the relative ease of getting to different locations and how comfortable and safe the traffic is around us when we get there, which will affect the behavior of individuals and businesses, which may inspire government oversight or intervention, which in turn will affect traffic patterns. This has the potential to reshape many facets

of systems that interface with transportation. Though we cannot fully anticipate the resulting dynamics, multiple disciplines have been modeling what the effects may be and making the tools necessary to get started. While many of these will be beneficial (easier and cheaper access for many to reliable and safe transportation), others are not as clearly good, pending deliberation about the metrics needed to match essential features with desired performance at particular intervention scales. What matters is that these interventions are designers' to make, as many externalities can be anticipated and managed over time through choices about the transit model.

2.2 A Sampling of Practical Proxy Metrics

Here I sketch the general space of AV externalities, as well as practical metrics for evaluating AV behaviors in relation to them. As this space is quite large, we focus on six high-level categories and give examples of proxy metrics in each: one short-term effect, and one long-term effect. Interestingly, short term impacts are more often about AVs adapting themselves to society as it is, and long term impacts tend to account for how society will adapt in response to AVs. These proxy metrics are only roughly representative of the true impact of AVs on the externality, and more work will be necessary to refine them. However, it is important to have some explicit metric as a starting point; we propose metrics grounded in practical concerns, suitable for informing useful changes to current systems. We argue that it is better to consider a simplistic proxy of an externality than for the externality to be ignored entirely. Each section concludes with a selection of other impacts in that category which could serve as directions for future work. Most of these impacts are referenced from (Elefteriadou et al. 2012) and reinterpreted within the AV context unless otherwise noted.

Physical and Economic Mobility

Short Term: AVs will affect access to resources for individual well-being such as food, jobs, schools, and entertainment. As AV routing algorithms unevenly shape how easy it is to get from one location to another, they have the ability to connect—or further isolate—individuals from the resources they need to live a healthy and fulfilling life.

As a concrete example, we can take the problem of food deserts (Beaulac, Kristjansson, and Cummins 2009). Suppose we have some population living in housing in a distributed manner. There are good food options and bad food options likewise distributed. For an algorithm A we will say that time to go from the housing position h to the food option f will be $t^A(h, f)$. We can then use a simple model for consumer choice in which the consumer will visit a high quality food option. That is, the consumer will choose the high-quality food option if the added cost over the easiest low-quality option is less than some constant. We define $F^A(h)$ to be 1 if a consumer at position h would choose the high-quality food option, under the algorithm A and 0 otherwise. This yields a proxy metric for consumer nutritional health as follows:

$$N_{\tilde{H}}^{A} = \mathop{\mathbf{E}}_{h \in \tilde{H}}[F^{A}(h)]. \qquad (1)$$

Note that this is a clearly simplified model for consumer choice, as well as the differential impact of different foods on consumer health. Moreover, implementations of this system would likely raise both political and ethical challenges, as a system could try to maximize this objective by

making it very difficult to get to unhealthy food options. This might be viewed as paternalistic and as undermining consumer freedom. However, these reasons are insufficient to abandon the possible transformative impacts of such an intervention. Instead, they could help ensure we build such a technology in a way that can adapt to the needs and ethical considerations important to the relevant stakeholders.

Long Term: The previous example took where people live as given, but in the long run, people will relocate or change address. Where they decide to move depends largely on how easy they believe it will be to get to where they need to go, such as ease of commute or proximity to grocery stores. Much as cars facilitated urban sprawl (Norton 2011), AVs could have effects of their own, but these effects will depend on their routing algorithms, as these algorithms will determine how easy it is to get from place to place and will modify the value of time of being in transit.

Using the techniques similar to the previous proxy metric, we could generate proxy metrics for commute time, and commute time for educational opportunities. If we have a model for personal preference in which person p from population $\tilde{P}$ cares about a weighted sum of these considerations with specified weights, then we can model the individual choice of a person over time as choosing the house for sale which maximizes the weighted sum of these features. We will define the House Choice of a person p over a set of housing options H to be under algorithm A to be:

$$HC^A(p, H) = \underset{h \in H}{\mathbf{argmax}}\{\alpha_N^p N_h^A + \alpha_C^p C_h^A + \alpha_E^p E_h^A\}. \quad (2)$$

Of course this represents a single housing choice, and to understand how the larger distribution of people would change over time, this process would be iterated. Each time a person moves it would open up a new vacancy, and another person would move in. Looking at the resulting long-run distribution of housing, we could measure many interesting properties. For example, we could pay attention to the concentration of people with a particular prioritization. Suppose this process reaches stable distribution of housing $\tilde{S}$, then we can measure the average amount that a person p agrees with their neighbor p_n on the importance of education, or educational agreement:

$$EA^A(p, H) = \underset{p \sim \tilde{S}}{\mathbf{E}}[|\alpha_E^p - \alpha_E^{p_n}|] \quad (3)$$

This may seem like an overly simplistic model, but it can be seen as a more complex version of Schelling's seminal work on segregation in the housing market through selfselection (Schelling 1969). This work shows that even simple models can lend themselves to important insights. As such, designers should take model uncertainty into account, rather than treat it as a blocker to progress in this direction.

We should expect even longer term consequences beyond this model, as the urban sprawl we see now is not only a result of people moving to suburbs, but the housing supply adapting to that demand. In essence, the routing algorithms for AVs have the potential to reshape the geometry of cities through how they shape the demand for housing and how the supply for housing responds to that demand.

Other Impacts: Various design strategies could accommodate the issues raised above. One is to scale AV infrastructure with equity-focused initiatives so that deployment does not accidentally cause gentrification or urban decay in regions where service offerings are mismatched with local mobility needs (O'Donnell, Corey, and Podowski 2017). Because use of public transit, bicycles, and pedestrian mobility increase sharply in dense environments, AVs will require routing capacities that scale with various environmental and infrastructural measures of density. These include transit-oriented employment and residential density, development scale, density gradient, population and employment centrality, population density gradient, density at median distance, density of development, percent of houses within one mile of an elementary school, percentage increase in residential density, gross and/or net residential density, building coverage ratio, average school size, and non-residential intensity.

Another strategy is to integrate AV usage with first- and last-mile mobility considerations, both to improve service in low-density areas (Ohnemus and Perl 2016) and augment public transit connections so that the cost of switching between transportation modes is reduced for multi-modal road users (Shaheen and Chan 2016). Important metrics here would include: average trip length per traveler, delay per traveler, door to door travel time, HCM-based bicycle LOS, and proportion of total person miles traveled (PMT) for non-single occupancy vehicles (SOVs).

Legacy metrics pertaining to destination accessibility should also prove useful. These can be subdivided into area-based measures that help optimize land use to minimize travel, and network-based measures that make it possible to compare the viability of different transportation modes along a particular route. The former includes: residence proximity, employment proximity, work accessibility, number of key destinations accessible via a connected pedestrian system, industrial/warehouse proximity, transit convenience/stop accessibility, geographic service coverage, population service coverage, percent in proximity, and transit accessibility. The latter includes: bike/pedestrian accessibility, destination accessibility, residential accessibility, average walking distance between land use pairs, spacing between village centers, and multiple route choices.

A barrier to effective models in this domain is the need to coordinate individual decisions and day-to-day priorities with macro-structural traffic effects that unfold gradually. Fortunately, there are many canonical metrics for road network usage that are also easy to communicate to the public, meaning that new models can be updated in response to stakeholder feedback. These metrics include: vehicle occupancy by land use, district-wide Level of Service (LOS)/Quality of Service (QOS), local traffic diversion, percent of system heavily congested, v/C ratio, vehicle density, demand/capacity ratio, Maximum Service Volume, Peak Hour LOS, and Percent of Capacity Consumed.

In a broader sense, translating local community priorities into perception and routing modules can draw from canonical measures of diversity in infrastructure and environment. These include the Smart Growth Index, significant land uses, land use ratios, land use balance, variation of agriculture of green fields, land consumption, core land use, land use separation, the Transportation-Efficient Land Use Mapping Index (TELUMI) model, minimum thresholds of land use intensity, nearby neighborhood assets, distinct indexes for sprawl (Ewing, Schieber, and Zegeer 2003; Galster et al. 2001), land use within village center, land use within transit supportive area, and jobs/housing balance.

While useful, these measures must be understood as proxies for deeper structural problems at the intersection of physical and economic mobility. Access to the city center remains a central concern of urban planning and transportation infrastructure (Shen 1998), serving as a proxy for access to labor markets and low-income mobility (Montgomery et al. 2018). Recent measurements of food deserts (Beaulac, Kristjansson, and Cummins 2009), commuter health exposure (Knibbs, Cole-Hunter, and Morawska 2011), and subjective appraisals of daily travel routes (Gatersleben and Uzzell 2007) indicate the difficulties of tracking unanticipated externalities of common travel patterns.

Environmental Effects

Short Term: Many vehicles release various forms of pollution which are harmful to the residents of high traffic areas. There has been a large body of research in quantifying both the amount and the health effects of pollution on local residents (Fisher et al. 2002; Krzyzanowski, Kuna-Dibbert, and Schneider 2005; Lipfert et al. 2006; Zhang and Batterman 2013; West 2004). The effects of this pollution will depend on the type of vehicle, the density of the traffic and/or the neighborhood, weather conditions, physical proximity of road users at particular times, and many other complexities. As a simple model to control for these effects, we can estimate the number of people present and say that there is a small penalty per second for every person inversely proportional to the square of the distance of that person from the vehicle. Or more formally, if $\tilde{P}$ is a pollution of people and c is the location of the exhaust pipe the vehicle, then we define the cost of pollution to be $Pol(\tilde{P})$.

$$Pol(\tilde{P}) = \mathop{\mathbf{E}}_{p \sim \tilde{P}} [\frac{1}{l_2(p, c)}] \qquad (4)$$

An AV that is trying to minimize this cost would be more likely to route around local communities towards less occupied areas, distributing pollution where it has less adverse effects. Additionally, even along a fixed route, AV designers could take advantage of the fine-grained control available to AVs, to optimize for pollution effects based on the characteristics of the population, the vehicle engine, and traffic conditions. While this model could be improved by better models of diffusion or quantification of health effects, its ability to induce a change in AV routing behavior shows the efficacy of even poor proxy-metrics to help reduce externalities.

Long Term: In the longer term we have the externality of CO_2 emissions. At a high level we could model these impacts much like pollution above, while ignoring many of the same considerations such as vehicle type and traffic density. However, it is also important to consider the effects of induced demand (Lee Jr, Klein, and Camus 1999). Induced demand occurs when the transit system makes it so much easier to get from place to place that people decide to take more trips, as rides that were previously not worth the effort become worth it. To model induced demand we have some population of possible rides $\tilde{R}$, each of which have utility to the consumer $u(r)$. Suppose we fulfill rides using an algorithm A which fulfills ride r with a route that takes time $t^A(r)$, and charge a fee $f^A(r)$. If we assume that the value of the customer's time is αt then we can define $R(A) = \{r \in \tilde{R} | u(r) \geq t^A(r)\alpha_t + f^A(r)\}$ to be the set of rides which are executed under algorithm A. If we further assume that CO_2 emissions are proportional to the length of the trip, $l(r)$, at rate R^{CO_2}, we can define a proxy metric for CO_2 emissions under induced demand as follows:

$$CO_2(A, \tilde{R}) = \underset{r \sim \tilde{R}(A)}{\mathbf{E}} [l(r) R^{CO_2}] \qquad (5)$$

This simplifies the true externalities, and does not incorporate differences in road conditions, models of vehicle, fluctuations in demand, traffic conditions, interactions between routing vehicles, and many other factors. Still, this basic model may serve to substantiate carbon-pricing or surge pricing policies, given the expected drop in α_t with the adoption of AVs. Extending this to incorporate more of the vast work on quantifying CO_2 emissions (West 2004; Sgouridis, Bonnefoy, and Hansman 2011; Noland and Quddus 2006) and induced demand (Hymel 2019; Omstedt, Bringfelt, and Johansson 2005) are clear directions for future work.

Other Impacts: There are many metrics for vehicle pollution beyond individual car exhaust and aggregate CO_2 emissions. These range from noise pollution (Campello-Vicente et al. 2017) to the wider ecological, fiscal, and social factors associated with AVs' environmental sustainability. Ecological metrics have been well documented in the literature: attainment of ambient air quality standards, daily CO_2 emissions, daily NOx/CO/Volative organic compound (VOC) emissions, noise pollution, impact on wildlife habitat, and water runoff. Changes to these in turn will likely generate effects on fiscal metrics related to activity level: additional fuel tax, transportation utility fee (TUF), vehicle miles traveled (VMT)-based impact fee, consumption-based mobility fee, improvements-based mobility fee, cost recovery from alternate sources, variable fees based on LOS, benefit cost ratio, parking pricing, and capita funding for bike/pedestrians. These, in turn, may have community-level social impacts whose measurement is vital but somewhat more speculative for activity level and modal share: distribution of benefit by income group, transportation affordability, equitable distribution of accessibility, commute cost, transit values, fee charged for employee parking spaces, travel demand management (TDM) effectiveness based on TRIMMS model, travel costs by income group and/or race, VMT by income group and/or race, mode share by income group and/or race, and walk to transit by income group and/or race.

Incorporating these proxy metrics would have several modeling benefits. It would provide strong estimates for demand that would help align emissions control with wider societal aims for equitable road access. It would also help integrate AV policy development with ongoing research on updates to pavement materials and construction practices of tollways (Al-Qadi et al. 2015), leading to possible new improvements in sustainability. And it could leverage vehicle-level data collection to control for the granular spatial and temporal features of air pollution that have recently been measured at unprecedented micro scales (Apte et al. 2017; Caubel et al. 2019), ensuring the benefits of emissions reduction are both locally and globally equitable. The deployment of AVs at distinct scales of road infrastructure (urban core, commuter routes, interstate highways) could also aid in the evaluation of alternative measurement approaches that trade-off precision against efficiency, an open research question in environmental engineering (Messier et al. 2018).

Local Community Needs

Short Term: The presence of traffic and the behavior of that traffic, has the potential to stifle or facilitate the interaction and congregation of members of the community. Roadways are not fully isolated from the rest of society and often intersect open air markets, public squares, and public parks. The behavior of these vehicles, the amount of traffic and their deference to local pedestrians impacts the ease with which these communities can conduct themselves.

As a simple model for this effect, you could model pedestrians walking, $\tilde{P}$ through an open-air market, going from shop to shop. The behavior of the pedestrians could be modeled through taking the shortest path to their next destination and pausing along that path if it would result in getting too close to a moving vehicle. Evaluating a routing algorithm A in simulation would result in each pedestrian p visiting some number of shops $s^A(p)$. A simple metric for the interference of the AVs on the market would be the average number of shops visited by the pedestrian S^A defined as:

$$S^A = \mathop{\mathbf{E}}_{p \sim \tilde{P}}[s^A(p)] \qquad (6)$$

There are many clear directions to extend this work, through modeling parks or city squares, adding more accurate models of pedestrians, and through accounting for the stress or noise disturbances of the vehicles. However, even this simple model would be enough to incentivize the AV to stay away from open-market areas if there are comparable alternatives, and to try to minimize the disruptions to the pedestrians walking through the market.

Long Term: On the longer term, the decision for people to visit the open market would be informed by their past experience, and if these markets were always crowded with AVs and difficult to navigate, it could disincentives people from visiting and eventually cause these open markets to disappear. To model these effects we can again move to a rational choice model on the part of the pedestrians. Let $\tilde{M}$ be the distribution of members of the population, and let $C(m)$ be the personal cost for a particular member of the population visiting the market over their outside alternative option. If the amount of enjoyment a pedestrian gets from visiting the open air market is directly proportional to the number of shops they visit, then the subset of the members that find it worth it to visit the open-air market under algorithm A will be $\tilde{P^A} = \{m \in \tilde{M} \mid C(m) \leq S^A\}$. Thus we can compute the effect of the algorithm for controlling the AVs on the number of visitors to the market by the V^A defined as:

$$V^A = |\tilde{P}^A|. \qquad (7)$$

That is, just the size of the set of members of the population which find it worth visiting the market. In addition to the extensions to the shopping model, this model could be extended through heterogeneous population models, trade-off considerations between other locations the members may be choosing to visit, and the incorporation of this metric with other metrics we consider.

Other Impacts: Community stakeholders and neighborhood representatives will need to affirm AV operations as legitimate and trustworthy, rather than merely safe. This means that the AVs must operate differently depending on the human factors considerations of the communities they route through, including emergency preparedness (Sullivan and Ha¨kkinen 2011), behaviors of road users during evacuations (Wong, Shaheen, and Walker 2018), and navigation wayfinding (Montello and Sas 2006). Geofencing, in which a virtual perimeter is mapped onto a specific geographic setting to determine who can access the platform, is already a common strategy for providing AV services and could be readily used to incorporate these considerations.

Just as widescale AV deployment is likely to highlight mobility features across communities that have never before been "priced in" to vehicle services, it may also generate disaffected consumers

who are forced into suboptimal mobility patterns because their needs do not conform to AV routing considerations. As a result, designers will need to inquire into and agree on new service tiers that effectively map various AV routing specifications to local concerns.

There is a risk of managing trade-offs poorly if well-defined metrics for relevant features are not consulted. A clear limit to the congestion pricing paradigm is the ready availability of parking and total parking supply, which directly affects a given community's preferred mode of transport and the perceived need for alternatives (Elefteriadou et al. 2012). Other relevant measures include: bicycle network density, parking spaces per 1000 workers, age of transit vehicle/fleet, bus shelter locations, bicycle parking requirements, bicycle parking spaces at schools, inter-modal connections, transit service quality index (de Oña, de Oña, and Calvo 2012), transit network coverage, transit service to site, walking distance to transit, project adjacency to transit, and connectivity to inter-modal facilities.

With respect to individual behaviors, another path forward is reliability measures, which are able to estimate upper and lower ranges of travel time but require higher rates of data collection (Elefteriadou et al. 2012). These include percent of trips 'on time', 90th or 95th percentile travel time, the Buffer index (used to prioritize freeway corridors according to travel time reliability) (Lyman and Bertini 2008), and planning time index (PTI). With respect to entire traffic networks or AV fleets, it is instead possible to consult transit-oriented metrics that capture the desirability of the service itself, relative to alternatives. These have been divided into measures of occupancy (the number of riders using the system), including load factor, passengers per transit vehicle mile, ridership, transit peak hour occupancy, and percent person-minutes served; measures of service availability (the provision of service to riders), including average frequency, average headways, hours of service, off-peak transit availability, transit service density, transit type availability, fixed route missed trips, on time performance, demand-response transit (DRT) trips not served, and response time for DRT; and measures of operation (addressing the speed, efficiency, and productivity of the system), including number of fare media sales outlets, transit productivity, number of transfers, transfer time, transfer time between modes, transit priority delay reductions, transit reliability (quantitative), fleet spare ratio, road calls, average life of vehicle components, and average age of vehicle components.

Infrastructure

Short Term: In the short term, AVs interact with the existing infrastructure, including the existing signage, lane markings and road damage. Depending on the current conditions certain roads will be better or worse and facilitating comfortable, safe, and efficient trips in AVs. If we have some metric for the comfort of the average trip down a particular segment s of roadway $C(s)$, which could be a measure of the number of potholes hit, the probability of the driver needing to take control, or the probability of near-collision. We call these costs collectively *navigation costs* and it would be sensible for our routing algorithms to take this into account and minimize this cost. As a first-pass we could say that for a distribution of trips $\tilde{T}$ routed by algorithm A, let $Seg^A(t)$ be the navigation cost for a road segment t, from a set of road segments visited on trip $t \in \tilde{T}$. Then the total expected navigation costs of A under the trip distribution and road conditions would be $N^A(\tilde{T})$ defined as:

$$N^A(\tilde{T}) = \mathop{\mathbf{E}}_{t \in \tilde{T}}[Seg^A(t)] \qquad (8)$$

These costs could be refined by better models for navigation costs, or considerations for how these navigation costs are effected by local congestion or construction.

Long Term: Like other parts of our society, the infrastructure itself will change in response to AVs. Not only could regulations and standards change signage and lane markings, but the patterns of AV activity could impact the deterioration of roadways. Through the centralized control of AV routing, when and where this deterioration is distributed would be under the control of the routing algorithm.

Let S be the set of road segments that make up the road network. Each segment will have a current quality $q(s)$ which grades the quality of the road conditions on that segment. We will say that the navigation costs are directly proportional to the current quality of the segment, so $C(s) = \alpha q(s)$. In addition, we will say that the road quality is reduced by some proportion ε whenever it is traversed. If the quality of the segment reaches 0 it can no longer be traversed and must be repaired.

In this model, if we try to minimize the costs $N^A(\tilde{T})$ greedily on a trip-by-trip basis we will find that we always take the highest quality roads. This will have the effect of the distributing the effects evenly, as the highest quality roads are used until they are worn down at which point another route is more appealing. Thought this is equitable it has a practical concern, that all of the routes will need to be repaired at the same time, likely causing long delays. We can model this effect as well by supposing that it will take one week to repair a road segment. Given this and a routing algorithm A, we could run a long-run simulation of routing activity, including the road ware and repair models, to get a distribution of trips in the long-run $\tilde{T^A}$. Thus we can measure the long-run navigational costs by LRN^A defined to be:

$$LRN^A = N^A(\tilde{T}^A) \qquad (9)$$

Better models of road damage, which are backed by real-world data and weather data could be used to improve this metric, along with more accurate models of construction. Moreover, if roads degenerate at different rates, it could be reasonable to minimize construction costs by staying on more robust roadways. Finally, since AVs allow for large scale monitoring of road conditions it becomes possible to coordinate more closely with cities to plan when construction will take place and route accordingly.

Other Impacts: As AVs become widely deployed, their effects and impact on public infrastructure (roads, bridges, highways, electrical grids) may be felt unequally. For example, while AVs must avoid potholes successfully and consistently, there is a tension between modeling potholes in terms of perception (identify them as they appear) or route planning (avoid roads that are more likely to have them). The consequence is that a failure to specify proper perception and routing constraints will harm regional mobility in ways that cannot be readily mitigated. As feature detection of road damage remains a stumbling block, the latter seems more likely for the foreseeable future–and this is likely to generate effects on congestion, highway flow, and other macro-traffic dynamics. AVs can compensate for this by measuring and minimizing loss in fuel efficiency or average time to

destination from avoiding potholes, aiming to preserve the road without disrupting traffic. This is analogous to other settings where AV software compensates for hardware limitations, except here the road itself is also modeled as "hard-ware" rather than relegated to the external environment. This is quantified through reference to existing models for high-way maintenance (Theberge 1987), priority damage assessment (Snaith and Burrow 1984), and smart pavement evaluations (Asbahan, McCracken, and Vandenbossche 2008). This work could aid constraint satisfaction by including factors that corroborate existing public standards for road maintenance, rather than modeling vehicle motion in isolation.

Changes to signage and lane markings could be made into controls on the large-scale effects of AVs. A natural implication of this would incorporate measures that reflect different design scales for multi-modal concerns (Elefteriadou et al. 2012), which also map onto our spectrum of local vs. global effects. On the former end of the scale, updating perception to conform to various point design measures will help AVs modify their speed and behavior in real time to conform to stakeholder expectations and priorities. These include: wayfinding information, sidewalk quality/width/shade, tree- lined/shaded streets, walkable streets, systematic pedestrian and cycling environmental scan instrument, commercial on-site amenities to support alternative modes, availability of on-site bicycle amenities, pedestrian scaled lighting, ratio of street width to building height, parking screening, bus pass program utilization, and parking shading.

Sensitivity to network-level effects should also be reflected in metrics for connectivity and route directness at the neighborhood level. These include: square feet of pathways/sidewalks, crosswalk spacing, number of safe crossings per mile, bicycle parking at stops and stations, parking footprint, block length, parking location, bicycle path condition, pedestrian/bicycle route directness, land use buffers, walking environment, bicycle maintenance stations, bicycle/pedestrian connectivity, connectivity indexes (Mishra, Welch, and Jha 2012), project adjacency to existing network, connected and open community, connected sidewalks/paths, connected streets, and cross access.

For example, residential neighborhoods in the United States often accommodate special needs groups through distinct signage: warning signs about pet dogs and cats, "children at play", and protection for the disabled (e.g. audible walk signs). Beyond vehicle features such as wheelchair access, AVs will need to incorporate routing adjustments so that time spent in these zoned areas is minimized. Meanwhile, some communities require unique forms of road mobility, such as retirement facilities and golf courses that have their own specialized modes of transport. Some communities have adopted special guidelines for golf carts interacting with normal traffic vehicles (Head, Shladover, and Wilkey 2012). Each of these considerations, and other details of local customs which we have yet to consider, need to be incorporated into the local control procedure so that they can be customized to be contextually appropriate.

Regional design measures, tailored to capture the completeness of regional transportation systems (Elefteriadou et al. 2012), are also well-suited to this problem. Relevant metrics include on-vehicle bicycle-carrying facilities, park-and-rides with express service, parking spaces designated for carpools or vanpools, transit passes, traffic cells, percent miles bicycle and/or pedestrian accommodations, miles of express fixed-transit route/dedicated bus lanes, road density, lane miles per capita, percent of network that is "effective", and roadway network balance.

Private Economic Activities

Short Term: AVs will often be the most convenient way for customers to get to local business, and thus changes in the way AVs are controlled, could dramatically affect the accessibility and thus the profitability of local businesses.

To model this, suppose we have some population $\tilde{P}$ and some set of business locations $\tilde{B}$. For each member of the population p and each business location p there is some time cost that member would be willing to incur to visit that location $c(p, b)$. Under algorithm A, the time it takes for p to visit b will be $t^A(p, b)$. Thus the total set of visits that occur are $V^A = \{(p,b) \in \tilde{P} \times \tilde{B} | t^A(p,b) \leq c(p,b)\}$. A reasonable metric for the amount of business activity for business location b under algorithm A is $BA^A(b)$ defined as:

$$BA^A(b) = |V^A| \qquad (10)$$

This model does not account for the quality of business products, nor population heterogeneity, nor changes in demand. Yet this model is sufficient to notice that there are trade-offs between how much activity is attracted to different local businesses. This raises clear questions of fairness, accountability, and governance as to how this trade-off should be determined. However, note that the presence of serious concerns about explicit control of these measures should not be taken as a reason not to explore these directions. The alternative to collectively deciding on how to handle this trade-off is for the trade-off to be arbitrary or privately decided, both of which raise their own serious concerns.

Long Term: As a result of changes to consumer activity in different locations AVs could cause some businesses to struggle and others to flourish. As new businesses are founded and others fail, AVs could shape where new business dielectrics are located by determining the locations which will have the most business activity.

In the Section on the Long-Term effect on personal physical mobility we discussed a similar effect with housing choice over time, and we could model businesses moving over time in the same way. Instead we will present another approach, which tries to predict the long-run concentration of businesses in an area directly from the business activity.

So building on the previous metric we can model the number of profitable business in an area to be proportional to the amount of business activity in that area mediated by some constant α. Thus we can define the long-run business con- centration at a particular location by $BC^A(b)$ defined by:

$$BC^A(b) = \alpha BA^A(b) \qquad (11)$$

There are many considerations missing here, many which are mentioned in other sections, though even in this model one can start to see the impacts of the AV algorithm on the local economy. If the algorithm systematically neglects the local business districts it could have the effect of squashing local business and harming local quality of life. On the other hand, local businesses which are currently difficult to access could be revitalized through access to new customers. In this way AVs offer a powerful and dynamic way of promoting economic activity in a community.

Other Impacts: The advent of platooning will make it possible to coordinate stakeholder commutes on unprecedented scales, likely leading both to new market configurations and problems with traffic coordination. There is a risk of managing trade-offs poorly if well-defined metrics for relevant features, such as community measures, are not consulted. For example, as mentioned earlier, a limit to the congestion pricing paradigm is the availability of parking, which directly affects a community's transport preferences and needs.

It should also be possible to incorporate multimodal considerations as a constraint: if a neighborhood is far from a public transit line or other means of access to the urban core, its AVs could be given preferred access during heavy congestion times or priced differently than comparable neighborhoods. This would address a longstanding policy question of whether it is simply easier to compensate road users for consistently taking a suboptimal mode of (public) transportation rather than trying to implement congestion pricing at scale (DeCorla-Souza 1994). It would also help make dynamic pricing friendly to public policy standards by its application to particular highway segments matching low-income community needs with toll discounts–a major constraint on prior implementations (DeCorla-Souza 2007). Moreover, it leads to a wide assortment of natural technical problems: designing routing procedures to maximize welfare, minimize congestion, ensure equitable access to mobility across communities, or balance the performance of shared and individually owned AVs. These have the potential to trade-off against other considerations, such as zoning rules, local efficiency, and safety. Considered as a whole, this points to a vast unexplored space of well-defined technical problems whose solutions would help ensure that the benefits of AVs are distributed fairly and effectively.

Meanwhile, one means of confronting structural economic effects is to incorporate freight-oriented metrics. As automation develops in congested urban areas to aid passenger travel, AV delivery via trucks could support the movement of goods and augment commercial access through regions that might otherwise be left behind. Relevant metrics include: truck miles traveled, truck throughput efficiency, freight delay, number of violation of weight restrictions, and overweight permits. Other measures provide innovative ways to approach total corridor capacity through parameters for demand, such as auto/demand response transit (DRT) travel time ratio, auto/transit travel time ratio, multimodal LOS, and seat capacity/person capacity. While measurement of and distinction between induced and latent demand on a network scale is notoriously difficult (Lee Jr, Klein, and Camus 1999), important factors have been identified for trip generation (average vehicle occupancy, bicycle and pedestrian activity, community capture, internal capture, internal capture, mean daily trips per household, person miles traveled (PMT), person trips, trip length by mode, vehicle miles of travel (VMT) by mode, VMT per capita) and mode share (bicycle and pedestrian mode share, mode choice availability, mode split, safe routes to school program (SRTS) effectiveness, SOV mode split).

As the sophistication of localized routing increases, legacy metrics for safety and security may become relevant to make sure that new approaches to optimization do not compromise system behavior. Helpful historical metrics include: bike/pedestrian injuries/fatalities, traffic fatalities, transit accident rate, transit vandalism incidents, transit related crime rate, vehicle accident rate, crash statistics/locations, annual severe crashes. Risk management metrics include: percent of lane-miles under traffic management center (TMC) surveillance, average clearance times for major

incidents, speed suitability, percent of vehicles with safety devices, and ratio of police officers to transit vehicles.

Finally, a major stumbling-block for the use of traffic efficiency models is the need for more sophisticated travel time metrics for highly-localized neighborhoods, urban sub-regions, and particular corridors. One path forward is to incorporate Highway Capacity Software in support of the highway capacity manual (Manual 2000). This will help permit a choice of advanced modeling tools in the context of stakeholder interests and targeted focus groups (Flannery, Anderson, and Martin 2004). It also makes possible particular auto-oriented metrics of demand and system utilization: average speed, average speed weighted by person miles of travel (PMT), congestion duration, control delay, highway reliability, percent work trips within specific travel time, total segment delay, travel delay, travel time, travel time index, vehicle hours traveled (VHT), average commute time, time by trip purpose, vehicle hours of delay (VHD), vehicle speed/VHD by mode, and travel distance index.

Traffic Laws

Traffic laws are distinct from the other categories, as the co-adaptation between AV optimization and public policy cannot be strictly grouped into short and long-term effects. Because AV designers are immediately concerned with optimizing AV performance rather than deciding what "good" performance necessarily means, we attend to traffic laws more as a means toward the former. As a result, it would be reasonable for methods to consider possible changes or updates to existing traffic laws as an approach to control the impacts of AVs. For instance, new regulations could support the addition of traffic lanes, pick-up/drop-off points, and zones where pedestrians have different rights. Each of these interventions would allow for coordination between people and AVs, and thus it serves as an important lever of control, alongside local control, routing, and controlling stop lights.

There will also be a need for entirely new traffic regulations, as the maturation of AV optimization interacts with legacy forms of traffic control. Designers will therefore need tools and methods to accommodate this likelihood. A good source of inspiration is the work on standards, metrics, and simulation parameters by the International Bridge, Tunnel and Turnpike Association (IBTTA). IBTTA has supported and made possible studies of the impacts of innovative technologies on highway operators (Azmat et al. 2018), as well as the impact of public-private partnerships on financing road infrastructure in developing economies (Queiroz, Vajdic, and Mladenovic 2013). IBTTA has also developed specialized tools for modeling various tolling environments. For example, the IBTTA Tollminer is a visualization tool that includes maps of toll facilities, a list of managed lane projects in operation nationwide, an optimizable user interface, and annual data on public and private toll revenues, among other features. While it is geared towards modeling and comparing the relative effects of high-occupancy vehicle lanes and toll lanes (Poole 2020), this work could be readily translated to test new simulation parameters for AVs that incorporate speculative regulations for equitable mobility access.

Another source of inspiration is analytical tools from the Highway Safety Manual (Part 2010) (HSM), which could be applied to a wider range of urban simulation settings beyond highways. While focused on mitigating crash frequency, the Highway Safety Manual aims to coordinate safety and economic concerns in a way that well-approximates the human factors interpretation of

safety as a problem of limited attention and human capabilities (Banihashemi 2011). The HSM could help update current AV simulation work to prepare it for future traffic laws in two ways. First, it embeds system planning within engineering, construction, and maintenance as part of an integrated development process. This perspective would prepare AV designers for federal and state regulatory environments once they have moved past proprietary standards for simulation and control. Second, it pinpoints three neglected sources of data (site characteristics data, traffic volume data, crash data) and incorporates them as part of HSM safety prediction. This would help AV designers move beyond "cookbook engineering" when setting up simulation parameters, and instead incorporate the basic concepts of systems engineering–functions, requirements, and context–that will ensure simulations accommodate the multiple interfaces necessary for fair and inclusive urban AV navigation. Attention to these interfaces is likely going to be a key nexus of regulatory attention in the coming decades.

Other Impacts of AVs

Here we give a small survey of externalities that do not fit into the other categories. While important, they currently lack a single model or definitive list of proxy metrics. Ongoing attention to relevant disciplines, highlighted below, will be necessary in order to transform these concerns into workable measures that could serve as targets for optimization.

Human factors researchers recognize a basic distinction between correcting for error-laden driver behaviors and accommodating the limits of human perception. Whereas most computational simulations for AV safety and reliability try to minimize known errors, human factors assumes that people make mistakes, and that we should responsibly design for this as a feature of roads by identifying the contexts in which this feature is safety-critical. One concrete implication of this perspective is the imperative both to assess user perceptions of AV-relevant infrastructure continuously and compare these against distinct non-human metrics of surrounding physical infrastructure (Elefteriadou et al. 2012). Only through this comparison is it possible to evaluate and repair the disjuncture between the phenomenological and material features of the driver / road environment.

Accordingly, AV design will have to distinguish between optimizing for bad driver behaviors and modeling context-specific limits to human perception and decision-making. This includes creating models according to situations and protocols that work from the human standpoint rather than purely to minimize likelihood of crashes. For particular problems in urban traffic control, designers must identify an appropriate human factors metric for the situation in order for mobility concerns to be addressed or resolved without excluding certain participants. Fortunately, the existing literature has already identified relevant performance metrics that could eventually be incorporated for simulation studies. These have been sorted into the categories: infrastructure and environment, system utilization, user perception, safety, and stability (Elefteriadou et al. 2012). Together, these categories can be evaluated and translated over time into model features or simulation environments to capture human factors issues pertaining to pedestrians, cyclists, and other non-vehicle road users. This would effectively permit AVs to serve as an interface between primary-mode passenger mobility and the wider infrastructural context of urban design.

A clear example of the above is user perception surveys, designed to capture human needs and priorities beyond system optimization. These surveys will become all the more relevant as AV-

generated changes due to congestion begin to affect mode choice and entail costs of congestion mitigation. Only by measuring community attitudes and user perceptions can multi-modal projects be designed in a responsible and sustainable manner. Relevant measures for present and future user surveys include: bicycle LOS (FDOT), pedestrian LOS (FDOT), LOS-based on traveler perception, perception of transit safety, transit comfort, transit condition of vehicles and facilities, transit ease of using the system, transit reliability/performance (perceived), transit complaint rate, transit customer loyalty, and pedestrian friendliness.

2.3 Conclusion

In this work I have surveyed a broad range of impacts AVs are likely to have on our society, both through direct impacts on the transit system and through the rest of society adapting to those changes. Though many of these impacts are already externalities of existing vehicles, the combination of centralization and automation provides a unique opportunity to control for these impacts on a much broader scale. To take advantage of this opportunity it is important for designers as well as stakeholders to identify the core features that AI models will represent or remake through implementation. I have provided a sample of several metrics that could be incorporated into the design of systems today, as well as references to many more which I have not made explicit and formal but are prime candidates for formalizing in future work.

I am aware of the rich methodological debates between competing approaches to measuring quality of life in the developmental economics literature. In particular, the "index number problem" describing how a given measurement index relates to a particular normative standard of living looms large in the tension between utilities-based vs. capability-based approaches. As described by (Reddy 2003): "The choice of metric for the evaluation of the standard of living must ultimately be motivated by normative reasoning concerning the appropriate manner in which to evaluate the life circumstances of individuals. There is no escape from this dependence of the concept of the standard of living on the normative judgments of the evaluator. Even the decision to defer to information concerning individuals' subjective preference satisfactions represents a particular such evaluative judgment". In this work I have relied on the normative evaluations and legacy standards present in adjacent literatures, including human factors, environmental engineering, and transportation planning. While this is sufficient for sociotechnically specifying possible modeling choices, this reliance defers the problem of reconsidering legacy standards, including the normative valuation of transit itself, to domain experts as AV fleet capabilities mature.

More broadly, I see my contribution as a call for new technical work. I hope that this work can serve as a catalyst for AV researchers to refine the sociotechnical specification of particular fleet deployment scales, and support ongoing deliberation connecting technical modeling choices to the needs of stakeholders. Though the challenges we present are broad and interdisciplinary, the next steps on improving each individual externality in isolation are immediately actionable: making a proxy metric based on existing literature that can be optimized. By working together as a community, we can create better metrics and methods to take full advantage of this new opportunity to coordinate existing stakeholder commitments.

3. MAPPING THE POLITICAL ECONOMY OF REINFORCEMENT LEARNING SYSTEMS

Conversations about AI ethics often revolve around the elimination of statistical bias. If a given machine learning system makes many mistakes, a common approach is to provide the system with more or better-structured data so that the resulting representation is more accurate and perhaps suitable for practical implementation. A clear example is the various optimization problems entailed in the long-term development of autonomous vehicles. If a car is not good at taking left turns or merging onto the highway, designers can just simulate more "left turn" scenarios for an agent to learn from, and then take the resulting learned policy as a guide for desired real-world performance.

Yet the "more data" band-aid has failed to address many challenges with real systems designed for facial recognition, recommendation systems, and others (Barocas and Selbst 2016, O'Neil 2016, Pasquale 2015, Benjamin 2020, Noble 2018, Eubanks 2018, Zuboff 2019, Bolukbasi et al. 2016). This is partly because designers rely on piecemeal fixes that make a system perform better according to some narrowly defined metric, without deeper reflection on what "better" means in context. Reinforcement learning (RL), which models how agents might act in some environment in order to learn and acquire some approximation of intelligent behavior, may push this paradigm to its breaking point. Its three key ingredients are states (composing the environment at stake), actions (the options available to the agent at every time step), and rewards (the "return" given to the agent when it takes a particular action). It is often distinguished from supervised and unsupervised learning, in which the system can either reference only what is already known via labeled data or explore the data structure with minimal constraints. By contrast, the heart of RL is to interpret intelligence itself, whether human or artificial, as a set of learned behaviors that effectively balances what is known with what isn't — a kind of computational prudence.

For the RL designer, the problem lies in making reasonable assumptions about the environment that can be well translated into states, actions, and rewards. But uncertainties often seep into the model at each of these, making "better" or "worse" outcomes increasingly difficult to identify as the task becomes more complex. Consider the simple case of making coffee in a motel room: At what point is it not worth scrounging for filters or grounds vs. using the bag of Earl Grey next to the bed? Am I up for schlepping down to the lobby to use the unreliable cappuccino dispenser, or do I not want to get out of my jammies? How is a well-brewed cup of coffee "better" than a sour or bitter one, as long as I'm caffeinated enough to catch my flight? As the task must be defined independently from any particular learned technique, it is unreasonable to foist all these quandaries on the agent, let alone a groggy one! Instead, the designer somehow has to set up a Markovian environment (one in which the values of states and actions do not depend on past information) for the agent to observe, and then help the agent learn to navigate it. The agent's behavior has to be interpretable as good, i.e., as well aimed rather than merely clever, based on our perception of the task itself.

The above example remains relatively simple and could in principle support a clear specification of the task. But as we transition from modeled environments to actual systems—say, an automated lobby coffee station used by both British tea-lovers and American coffee-fiends who learn to trust or distrust its operation in the face of uncertain hotel conferences and flight delays—it is necessary to prioritize stable performance dynamics over any particular optimal solution. This is because the system is not operating in a single-agent Markovian environment, and the dynamics can no longer

be taken for granted. I argue that this problem transcends task specification and is instead about **normative cybernetics**: Who is the system for? What is its purpose? How will it maintain the stability of the domain in which it will operate? Or in the context of autonomous vehicles: How can we preserve the integrity of all the different "environments" at stake in citywide traffic mobility? What information channels are needed to monitor and control for autonomous vehicle performance in the streets and intersections that matter most? Under what circumstances might it be unacceptable to optimize a single algorithm for object detection, traffic navigation, and congestion management?

Finding complete answers to these questions will require decades of research as autonomous vehicles are further developed. But before that happens, designers of RL systems must understand that these social domains' normativity (the pattern of local behaviors we feel entitled to expect from others and ourselves) is unevenly and richly structured. While some of those structures are sufficiently defined to support optimization techniques, the definitions of others are not forthcoming. Normative cybernetics challenges current AI development practices in three ways: 1) reaching consensus on the normative structure of the domain is a *deliberative process* that cannot be achieved by data aggregation or computation alone; 2) normative structure is enacted and reformed through distinct *types of feedback*—in particular, within markets (which pursue optimization) and politics (which pursue collective definitions of the good); 3) developers need to construct *interfaces* with these types so that the model specification (defining states, actions, and rewards) is responsibly indexed and updated to incorporate the concerns of stakeholders over time.

The meta-question of which problems are "ready" for RL and which will require further definition shares much with themes of political economy. Rather than decontextualized principles of ethics, political economy asks the essential question at the heart of any RL system — what is the nature of value? — in the context of particular normative domains. Because these domains vary in scale and complexity, the same reward structure cannot be applied automatically to them, and designers need good sense about the levels of abstraction where norms operate so that the system works on the levels we want and not the ones we don't. There are scales at which the corresponding optimization is known, scales where the optimization is uncertain, and other scales whose features are unclear. In fact, the fields of engineering, economics, and governance roughly correspond to these and serve as distinct inquiries into how assumptions can be operationalized, competitively pursued, or re-evaluated. We need a distinct mode of deliberation—normative cybernetics—to coordinate this inquiry by asking how to guarantee stability in a manner that is proportionate to the structural integrity of the domain.

A systematic exposition of political economy is beyond the scope of this chapter. Instead, my goal is to briefly describe the distinct forms of social risk entailed by the optimization of advanced RL systems, how these forms might be interpreted according to existing legal standards, and what sorts of limits to optimization should be implemented to protect and reform infrastructure responsibly. Although these claims are meant to apply to any RL system of sufficient scope and complexity, continuous reference is made to the pertinent case study of autonomous vehicles (AVs) for the purpose of clarity and simplicity. Throughout, I consider the surrounding institutional contexts (social, behavioral, managerial) in which AV systems are made and deployed and which often absorb hidden costs related to suboptimal performance.

3.1 The limits of intelligent behavior

The themes of political economy help reveal how the structuring of an RL agent's learning environment can both make optimal learning possible and generate social harm if designers do not adequately reflect on how rewards have been specified. This can be illustrated through the *reward hypothesis*, the idea that "ends" or "purposes" are the maximization of the expected value of the cumulative sum of a received scalar signal (Silver et al. 2021). This is a fancy way of saying that for any particular job, there is some computable answer to the question of what it means to do that job well — that the definition of a good job is somehow baked into the activity and can be learned exclusively by referencing the relative success of one's own actions. Mowing the lawn means you are cutting blades of grass; doing the dishes means you are scrubbing away spots of dirt; beating Super Mario Bros. means you are collecting coins, beating levels, or finally freeing Princess Peach.

REWARD HYPOTHESIS: that all of what is meant by intelligent behavior is the maximization of the expected value of the cumulative sum of a scalar signal within some environment.

It follows that skill is best acquired by interacting with the environment directly rather than by imitating how someone else has done it. According to this hypothesis, optimizing for the underlying reward function rather than learning to mimic some observed behavior pattern is the most "succinct, robust and transferable definition of a task" (Ng and Russell 2000). This reward function is often not even unique, as it is common for different objective functions to be simultaneously optimized when there are overlapping interpretations of the observed behavior (am I pouring a glass of water because I am thirsty or because I want to rinse the glass out?). Moreover, the AI designer does not have to specify the mechanism for achieving a goal, as the RL agent can design its own strategy for achieving it.

Philosophically, the reward hypothesis is a claim about how the complexity of intelligent behavior can, in principle, be encapsulated by the simplicity of scalar reward. In other words, different actions and strategies can be definitively compared as better or worse than each other with reference to the ultimate goal. If a reward function seems hard for the agent to learn, the hypothesis entertains the idea that further optimization (expanding the action space, adjusting the signal) will solve the problem, at least to the extent that there is a solution to be found. To clarify this point, the hypothesis does not claim that all human activities amount to utility maximization, but that all "well-specified" activities effectively do, at least in terms of the signals received and particular learning environment at stake.

But at the end of the day we are building systems that interface with the real world, not just models. Leveraging the reward hypothesis when designing AI systems such as AVs entails a problem of feedback — the need to structure and monitor the dynamic relationship between agent and world so that the nature of its activity is normatively appropriate. RL system developers must decide how to manage the gap between model representation and resulting real-world behavior (Sendak et al. 2020). How should they do this? The limits of the reward hypothesis show how good representations of activity cannot be blindly pursued in lieu of normative deliberation by tackling the problem of feedback.

For example, consider an AV perception algorithm that has trouble recognizing street debris, compromising the AV's ability to drive safely through areas with significant homeless populations

or unreliable street flow. The AV does not get into accidents and generally makes it to its goal on time, but occasionally runs over bits of plastic and glass in a way that does damage to the vehicle and possibly the road. Is this AV's behavior suboptimal (could it be doing a better job)? Or is the environment misspecified (doing the wrong job)? Or neither?

We can readily imagine a host of ways — some of which are technical, others less so — to solve this problem, depending on how we choose to interpret it. Our example AV could be rerouted to go through different streets that are typically cleaner, even though this would add to the travel time. We could more extensively validate the vision architecture to better avoid street debris. Or we could rebuild the AV so that the chassis is less prone to damage. These strategies propose alternative translations between expected utility (avoid debris) and desired outcomes (drive on all streets, drive only on streets that are safe, protect the vehicle, preserve the integrity of the road network).

Let's not lose perspective. Humans do not drive cars in order to avoid debris, but to get somewhere! The question of what it would mean to weigh hitting debris against the goal of getting to our destination on time or in one piece is far from the minds of most human drivers. Yet the reward hypothesis makes it possible to imagine a single utility function that encompasses all these features — a single environment where scalar reward is sufficient, rather than multiple worlds with incommensurate normative criteria.

That the reward hypothesis must have limits cannot be seriously questioned, unless we believe that there is a single ready-made perspective from which all goals can be computationally simulated and optimized. In that case, artificial specification would not be necessary — all humans ever do is aggregate rewards that have already been baked into our environment. Nor is it possible to dismiss the hypothesis entirely, as many tasks can be meaningfully simulated and humans do pursue well-defined objectives all the time. The fact is that somehow, humans can specify tasks by writing down what it means to perform them. This implies that to interpret the world by meaningfully carving it up into navigable chunks requires a different kind of agency than intelligently navigating it in the first place.

The reward hypothesis forces the RL designer to make explicit and weigh—or invent from whole cloth—various norms that are collectively followed but have never before been robustly specified or even determined. Even if a single utility function for driving does in fact exist, it has never been written down before (and is thus not ready-made for modeling purposes), it would require evaluating driving behaviors at enormously varied scales, and it may well encounter basic disagreements among those scales about what "optimal driving" actually means. The problem is that optimal and right may well turn out to be different in ways we cannot understand prior to sustained reflection about how the activity is supposed to work or not work. Unprecedented empirical research and political will are necessary to overcome these hurdles and encode features appropriately rather than the behaviors we assume to be optimal. This is why we need, beyond specification, normative cybernetics.

We might consider another approach. Instead of painstakingly crafting an AV specification that meaningfully includes the features we want, designers could remake those environmental features to conform to the AV specification they have. Andrew Ng endorsed this in the context of incentivizing pedestrian behaviors to accommodate the limitations of AVs: "Rather than building

AI to solve the pogo stick problem [i.e., rare human actions], we should partner with the government to ask people to be lawful and considerate. ... Safety isn't just about the quality of the AI technology" (Brooks 2018). To extend the example above, in practice this would mean that AV companies could partner with local communities to fight homelessness, have debris removed from the road so that their vehicles did not have to observe it, or discourage people experiencing homelessness from congregating near profitable streets. We can even think of this as a mechanism design problem: define the objective(s) we want and then reverse engineer incentives for the agents (human or otherwise) that would guarantee those objectives are met.

To be clear, this approach assumes such a definition is ready-made, and avoids substantive political questions about tradeoffs between safety and efficiency within either the learned model or the incentives given to people to conform with it. I will defer this problem of whether reducing homelessness or cleaning up roads may in fact be something we all want to include within the AV task specification, beyond cars that merely drive well, and whether criteria for resolving this already exist for AV companies to use. In fact, I believe we should create independent institutional spaces functioning as "AI clinics" that would richly deliberate about how to build systems in the context of problems like this, though that is not the focus of this chapter. Instead, the crux of my discussion here is that this basic indeterminacy — whether it is designers' responsibility to make AVs ready for the world as it is, or help remake the world itself so that AVs can navigate it, or other options made possible by unprecedented scales of technical intervention— is not something the reward hypothesis can answer for us.

We can define the political economy of RL as **the science of determining the limits of the reward hypothesis for a given domain**: framing it, comparing it with alternative specifications, and evaluating it. This is both a technical and a normative problem, because specifying rewards under uncertainty depends on the scale of the domain in which we are operating. A local government designing a road vs. a group of states designing an interstate may fall at different scales of the complexity hierarchy and could not be treated as similar. In other words, evaluating which reward structures could be institutionally affirmed and cybernetically supported through feedback requires asking what it would mean to govern RL systems, not just optimize them.

3.2 Computational governance

The problem space of RL governance is revealed by the resonances between the reward hypothesis and two foundational assumptions of the Chicago school of economics, in particular the ideas of Ronald Coase. One is that firms are better than markets at handling transaction costs pertaining to information flow (Coase 1995). This suggests that provision of AV service via a single in-house operational design domain (the specific set of constraints within which an automated system has been designed to operate) is more efficient than a mix of independent contractors comprising AV fleet owners, operators, manufacturers, and regulators. The reward hypothesis extends this intuition by allowing designers to imagine tasks as optimizable in-house as part of a single computation stack rather than specified or evaluated in a more distributed manner — that it is easier to maximize utility within the firm than under the guidance of third parties. This would provide a computational (rather than economic) basis for justifying the power of firms to aggregate social value.

The second is that courts and other political entities should intervene on social contracts only to ensure the rights of affected parties are allocated optimally (Coase 1960). The technical criterion here is Pareto efficiency: given finite resources, goods are to be distributed in the best possible way, with no party's benefit coming at another's expense. For example, only if AV fleets were found to generate measurable economic costs (commuter times, air quality, road damage) for specific neighborhoods could courts then charge the firm to make up those costs. The reward hypothesis extends this by explicitly internalizing known costs as part of the reward function before they can even register as economic externalities: rates of congestion, pollution, and road wear can simply be added as environment features and be updated in response to fluctuations in demand. While adding features does not preclude finding an optimal RL policy, it does make it harder for regulators and designers to understand it, which places constraints on system interpretability and how responsibility is allocated in case of harm. The basic problem at stake — deliberating about what a good neighborhood is and whether or how it can be sustained if the AV fleet is deployed — is sidelined, as it lies outside the boundaries of Pareto efficiency.

Together with the reward hypothesis, these assumptions would permit the designers of RL systems to "simulate" political economy by either adding features or maximizing expected utility at arbitrary computation scales. Such designers would, in theory, be in a better position to define and measure value than either the market or political institutions. Rather than allowing normative concerns to be expressed "suboptimally" by boycotting a service or making new zoning laws, it would simply be more efficient to let AV designers "figure out" what the reward function for driving is in San Francisco or New York or Cincinnati through a mix of routing adjustments and dynamic pricing. If this form of computational governance were seriously pursued, it would eclipse a core function of politics: articulating underspecified normative criteria that distinguish between utility losses and the definition of good social outcomes. This is where questions of RL optimization meet the themes of political economy.

We can make these abstract considerations tangible by illustrating the risks they entail. The reward hypothesis implies alternative strategies that reflect different approaches to risk, in particular either redefining the service that AVs are providing or optimizing service performance. The former is made possible by *reward shaping* — i.e., restructuring the agent's environment in order to facilitate learning the optimal policy. While reward shaping is specifically not meant to redefine optimal behavior itself, it does assume such a definition exists. This becomes a problem if an AV firm has selected a definition that does not responsibly account for different interpretations of "good" driving (e.g., don't hit objects, maximize fuel efficiency, minimize travel time). Using reward shaping to computationally reconcile such interpretations means that environment rewards have been evaluated in terms of expected utility and comparatively ranked according to some priority ordering. Actions that human drivers perceive negatively (driving into potholes, cutting someone off) are given scalar value and may be ranked differently according to some model specification, even absent a guiding legal standard. Complex normative approaches to these distinctions (potholes are bad but publicly managed, road rage is rude but tolerated, manslaughter is illegal) may be obfuscated or reinvented as their incommensurable stakes are reduced to an optimization problem.

To better evaluate this commitment, we can mobilize the tools and insights of the recent literature on antitrust (Steinbaum and Stucke 2018). In road environments whose norms are indeterminate, informal, or lack a clear form of domain expertise, the use of reward shaping amounts to a

proprietary claim on public infrastructure, also known as *monopoly power*, by whoever controls the system specification. In other words, a single firm would effectively act as the exclusive supplier of vehicle services and would have the ability to define what "good" and "bad" driving means within its operational design domain. As long as reward shaping is limited to environments whose dynamics are well understood, this might be acceptable. But as the AVs' domain expands from a single neighborhood to an entire city, the AV provider is essentially deciding what types and magnitudes of underspecified costs are acceptable for the public to bear as the system optimizes the firm's chosen definition of driving. Potholes are perhaps the best example of this, as the vehicle fleet will structurally generate them in specific places as certain highways are discovered to be safer or more efficient for routing purposes.

Historically, antitrust has interpreted problems like this through the common carriage standard. Beyond some geofencing threshold, an AV company should be interpreted as a common carrier that is responsible for ensuring fair and equal access to its platform. At certain infrastructural scales, the platform generates externalities that cannot be reliably tracked or managed through reference to goals that have been only partially specified by law, like avoiding crashes, maximizing fuel economy, or minimizing route time. This trend is likely to worsen as Uber and Lyft increase capital investment in automated rideshare, Tesla grows the market for personally owned AVs, and Waymo scales up the size and service area of its platform. Instead, parameters for these externalities must be set by a third party, requiring some sort of public commission whose job is to specify what social welfare means, rather than let it be optimized in a normative vacuum. Outside these parameters, the common carrier (in this case, the AV service provider) can be held directly liable for damages its platform does both to public infrastructure (roads, signage, etc.) and to regular road users, whether AV passengers or not.

The other approach to risk is referred to as *information shaping* (Griffith et al. 2013). This approach structures the environment so that the reward signal, however it has been defined, can be observed by the agent only under precise conditions. This may allow the agent to learn more reliably or efficiently but also restricts the number of sensory inputs. Possible forms of feedback are thereby neglected, because they would make optimizing performance harder according to the chosen specification. As a concrete example, AVs at a four-way stop may consider only physical distance from other vehicles as a signal, even if other information (pedestrians' gaze, other drivers' hand gestures, horns honking, people yelling, common knowledge about surrounding signage) is salient for human drivers. The fact that humans are able to coordinate via different sources of information is important, as road mobility is defined by multimodality: the coexistence of multiple literacies (pedestrians, cars, cyclists) and mediums for feedback in a single domain. In effect, information shaping may exclude or marginalize expected behaviors that are common in the real world, potentially neglecting the underspecified but integral modes of communication relied on by stakeholders.

A diverse network of regulators at the local, state, and federal levels is responsible for designing and evaluating forms of signage that support common road access. But left to its own devices, information shaping will gradually transform the suite of sensors used by the AV into the interface for the roadway itself. This would co-opt the authority of independent agencies and burden them with the responsibility of redesigning roadways to make them safe for AVs, a phenomenon that economists refer to as regulatory capture (Dal Bo 2006). Because of this, information shaping constitutes a claim on *monopsony power*, formally defined as the exclusive "buyer" of some good

or service in a particular market. The service in this case is the distributed labor force (regulators, manufacturers, and municipal bodies) compelled to support the AV firms' chosen specifications and sensory inputs. As long as the optimization is restricted to single-mode environments like highways, this may not threaten the public interest. But as its urban integration becomes more intensive, the AV interface will tend to exclude certain roadway literacies and delimit the range of mobility participants to whom AV-specified roadways can provide common services. Jaywalking is the clear historical parallel here, as pedestrians learned to see themselves as a problem for cars to avoid and largely ceded public control of streets to them by the 1930s (Norton 2011).

Returning again to antitrust, such problems are interpreted using the standard of structuralist regulation: some kind of firewall or public interface must be created across the organizations that produce the AV specification to ensure that it remains inclusive of road users. More importantly, this regulation would need to support a space for deliberation about the specification, treating it not as an "everlasting solution" but as a provisional means of dealing with the problem of feedback. This would prevent the fusion of private service provision with roadway access via restricted information channels, while permitting external regulators to investigate sensory inputs and confirm they do not exclude mobility participants. Structuralist regulation will become more important as we transition to 5G roadway infrastructure that will make AV platooning and citywide traffic optimization viable, as signal constraints for perception, localization, planning, routing, and controls must remain publicly coordinated and not merely optimal. At a minimum, these information dynamics must be able to be observed and interpreted by third parties, requiring the ability to evaluate the platform through external documentation or audits.

3.3 Policy challenges

The reward hypothesis cannot answer which forms of shaping are normatively appropriate for a particular AI service. This leads to open cybernetic questions of concern to RL system developers:

-What limits are there to what can be modeled?

-What types of feedback are available to bind system effects?

-Whom do we entrust with the power to set the bounds within which the reward hypothesis can be framed and comparatively evaluated?

It is difficult to answer these questions, as the conceptual landscape and context of AI ethics are rapidly shifting. Entirely new standards for antitrust are now being proposed that transcend narrow economic protections. Governments in the European Union and United States are discovering they have the stomach for confronting and regulating Big Tech platforms through a mix of fines, data protections, and ongoing lawsuits. In the aftermath of the Federal Trade Commission's allegations that Facebook has acted as an illegal monopoly, we must continuously evaluate whether aligning systems with social ends also requires examining the structure of the organizations that build them. These developments reflect difficult political questions (Fukuyama 2021): Must we wrest power away from private monopolies and place it in the hands of public officials? Or hold such power accountable regardless of where it lies?

Let us take a step back and consider the core values at stake when pursuing the reward hypothesis in more and more social domains. I propose there are two: *integrity* and *interoperability*. Integrity

means the normative structure of the domain, to the extent that it has been specified by law or custom, must be protected by whatever AI company acts as its steward. In principle, this means that there is some clean translation between existing social norms and the RL specification of states, actions, and rewards, although there may be uncertainty about how to achieve this technically. Interoperability means that if the normative structure of the domain requires further specification, the various interfaces at stake in a given system at least remain coherent, interpretable, and subject to evaluation by a third party. In this case, the RL specification must be subject to external oversight, and particular approaches to optimization must be backed up by public documentation.

Together, these values serve to protect the norms we have from encroachment by AI systems and defer the choice of underspecified norms until stakeholders are given the chance to articulate and affirm them. A truly deliberative approach must include both and be reflected in the institutional relationships among engineers, corporate managers, and external regulators. This approach is necessary to avoid conceiving of specification in a "solutionist" manner, as the problem of how to structure feedback is governmental rather than technical in nature. Below, I briefly present what this might look like in the context of AV development.

As AV fleets impinge upon more and more streets, they stand to inherit the responsibilities and commitments that cities have already made to infrastructure, safety, and road equity. In this case, municipal bodies could require companies to bear some of the cost for infrastructure repairs and any future megaprojects resulting from the provision of AV ride services. Companies could also be required to share data so that public commissions could better determine needed repairs. For inspiration on possible standards, we can look to recent work on contextual integrity (Nissenbaum 2009). Contextual Integrity (CI) interprets normative social domains in terms of their ends (the goals of participating agents) and information flow (the medium through which facts or data are permitted to move, with some degree of asymmetry, between agents). This helps specify limits for evaluating reward and information shaping, serving as a potential optimization standard that regulators could apply to companies.

In-house engineers would then have to define states, actions, and rewards to meet the standard specified by CI. For example, an AV that couldn't recognize road debris could still be deployed to those streets and be "street legal" as long as collisions remained infrequent or caused minimal damage according to thresholds specified by a third party. However, if that AV were found to generate unanticipated externalities in the form of traffic congestion, it could be violating commitments to public safety and equitable mobility and be found liable for harm. In this way, CI helps to distinguish the toleration of suboptimal behaviors from the evaluation of direct harms: while some states and actions demand strict enforcement and prevention, technical standards for object detection and accident avoidance have not yet been exhaustively specified.

Meanwhile, we must ensure that alternative forms of mobility are not excluded as roadways are brought online through new forms of "smart" infrastructure. One path forward is open application programming interfaces (APIs), a possible standard for AV companies to follow. Consider a given city in which a single AV fleet is dominant, effectively serving as a gatekeeper for public mobility itself. In this case, an open API could support public-private data sharing and structured competition with smaller services. This would prevent the fleet from leveraging its market dominance into redefining road access and would set limits on the vertical integration of service

provision. The firm's own definition of states, actions, and rewards would be less important than transparency about those definitions between engineering teams, as well as public-private coordination between corporate managers and municipal bodies. In this way, multimodal transportation concerns could be continuously addressed while preventing unilateral control over mobility partnerships.

Beyond regulatory oversight, open APIs also make it possible to set up markets for fair service provision. This could be achieved through a mix of service auctions (e.g., for neighborhood access) and administrative licensing, ensuring that pedestrians, smaller mobility services, and other stakeholders maintain road access. Following the path of telecommunications (Illing and Klüh 2003), AV companies could compete for access to protocols for interoperability as they pertain to particular roadways, within parameters that are acceptable to current road users. These parameters in turn would remain subject to revision as emergent traffic dynamics were observed and interpreted. Crucially, such tools would incentivize companies to care about and monitor the reward function their AVs are optimizing, helping to ensure that service provision is respectful of social welfare as well as technically optimal.

3.4 Conclusion

Both aspects of roads — their legacy status as a public good and their continued ability to accommodate structurally diverse means of use — are necessary conditions for the responsible development and deployment of AVs. Yet the computational governance problems discussed above are general and will be relevant for any RL system whose development entails reward or information shaping in domains of unprecedented scale and normative complexity. Ongoing technical and policy work pertaining to integrity and interoperability will help light a path for investigating the limits of the reward hypothesis in particular contexts, and by extension the emerging political economy of RL systems. The position of normative cybernetics I have outlined here is that the structural integrity of a given domain is in scope for AI development, that its preservation amounts to how feedback is structured between the system and the domain, and that at least some criteria for this may be found in how that domain has been managed by economic and political activity. This adapts an insight that scholars of political economy have long appreciated (Fligstein and Vogel): rather than starting from a stylized view of how the world ought to work and then leveraging data to minimize model bias, we ought to first look at how different institutions (individuals, firms, markets, governments) have approached the sort of problem we face, respecify it accordingly, and maintain interfaces with those institutions so that stakeholder concerns can be effectively addressed.

4. HARD CHOICES IN ARTIFICIAL INTELLIGENCE

The rapid adoption of AI systems is reshaping many public, professional, and personal domains, providing opportunities for innovation while also generating new forms of harm. These harms are diverse, ranging from physical dangers related to new robotic systems (e.g. autonomous vehicles), to economic losses related to welfare systems, to forms of racism and discrimination in systems that engage with biometrical data in public spaces or with personal data on social media platforms. These cases reveal emerging gaps between the promised beneficial outcomes of AI applications and the actual consequences of deployed systems. Like any technology, ongoing risks and harms due to AI are thus a product of the *sociotechnical gap*, "the great divide between what we know we must support socially and what we can support technically" (Ackerman 2000).

In response, a broad spectrum of civil society initiatives have emerged to safeguard human domains from the effects of AI systems. Debates about the sociotechnical gap have taken two forms. One is the proposal of normative principles to determine how the gap should be filled or who should do it. This has led to a plethora of reports and statements about how AI should be governed to respect fundamental rights, alongside a growing need to operationalize these principles (Schiff et al. 2021, Andersen 2018, Mittelstadt 2019). For example, the OECD Principles on Artificial Intelligence "promote artificial intelligence (AI) that is innovative and trustworthy and that respects human rights and democratic values," and are signed by governments. The European Commission recently proposed a regulatory framework to translate higher-level principles into concrete technical and legal solutions through "harmonized standards". However, it is unclear how these standards could reconcile the diverse needs of users in the context of particular systems and domains. Second is the proposal of technical tools to better fill the gap. While these efforts have generated many technical approaches related to mathematical criteria for "safety" or "fairness", their systematic organization and prioritization remains unclear and contested (Gebru et al. 2018, Mitchell et al. 2019, Raji and Buolamwini 2019, Green and Viljoen 2020).

Missing from both debates is a sustained interrogation of *what it means* to identify, diagnose, and ultimately fill the distinctive sociotechnical gaps generated by AI systems. This entails asking deeper questions about how a given system may restructure human values and social practices, whether technical and governance criteria may be reconciled in design choices, and when or where gaps emerge across the system's development lifecycle. Put differently, we lack a presentation of AI development in terms of what we call *machine politics*: interrogating how the choices that structure a system's design and implementation bear on the wholesale reorganization of human domains and communities. In this sense, AI development can be conceived as a deliberative practice comprising how human constituencies make political choices based on their sense of the good life.

Concretely, every AI system requires a consensus definition of what it would mean for it to be safe. But present proposals for the technical safety and governance of AI systems tend to focus on safety either as a criterion of technical design, operational conditions, or the experience of end users. This means safety criteria are marred by an underlying *vagueness*, the absence of unifying categories to establish whether a system's capabilities are safe or not.

This chapter makes two key claims. First, AI development must be reconceived in terms of the multiple points of encounter between system capabilities and sociotechnical gaps. This requires a new vocabulary and framework to make sense of salient gaps in the context of technical design decisions, constituting a reciprocal relationship between system development and governance. Second, developers must take on new roles that are sensitive to feedback about how to manage these gaps. This requires communicative channels so that stakeholders are empowered to help shape the criteria for design decisions.

My contributions flow from these two claims. In Section 4.1 I supply a lexicon of terms for the problems at stake in sociotechnical gaps. In Section 4.2 I analyze the present landscape of proposed technical and normative solutions to particular gaps in terms of piecemeal responses to vagueness. In Section 4.3 I present Hard Choices in Artificial Intelligence (HCAI) as a systematic framework that maps possible gaps to particular feedback channels for designers and stakeholders to use. In Section 4.4 I present this framework's implications for designers and advocates when evaluating the technical performance and governance standards of actual systems. Section 4.5 concludes.

I emphasize that my concerns, while responding to more recent iterations of AI and computer systems, are not new. The research agenda of situated design (Greenbaum 1992) and Agre's call for a "critical technical practice" (Agre 1997) comprise classic phenomenological critiques of "good old-fashioned" symbolic and expert systems, in particular the need to become critical about certain formal assumptions behind intelligence and to reassess problematic metaphors for perception and action (Dreyfus 2014). Yet much technical research today has moved beyond these critiques. Reinforcement learning (RL), for example, satisfactorily incorporates Dreyfus' exposition of intelligence as a learned, situated, dynamic activity developed from coping with one's surrounding environment and embodying different strategies for action. The question is no longer what computers can or cannot do, but **how to structure computation in ways that support human values and concerns**. To support this aim, I propose AI practitioners will need new *cybernetic practices* that guide how feedback may be solicited from existing and emerging political orders.

I thus apply an insight to AI development that scholars in Science and Technology Studies (STS) have appreciated for over four decades: any and every technological system is political, requiring collective agency and a corresponding form of deliberation to ensure its safety for everyone affected by it (Winner 1980).

4.1 Towards a Sociotechnical Lexicon for AI

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

As was stated in the original Dartmouth summer project proposal, research on artificial intelligence is meant to pursue "the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it" (McCarthy

2006). Beneath specific efforts to simulate language, brain models, and intellectual creativity, AI theorists were most interested in precision: adequately specifying the mechanisms underpinning intelligence such that they would be possible to replicate via computation and symbolic reasoning. This quest for exactness has continued to underpin many technical and conceptual interventions on how to model the intelligent behavior of agents within some environment, including the problem of specification in reinforcement learning (Milli 2017, Hadfield-Menell 2017).

- agency--the capacity of some agent (human or artificial) to act in order to achieve a particular outcome or result.
- intelligent agent (IA)--an autonomous entity which acts, directing its activity towards achieving goals.
- environment--a domain in which an IA can perceive through sensors and act using actuators, in pursuit of a goal.
- AI model--a mathematical representation of the environment, constructed through either simple rules, a model, or a combination thereof, the parameters of which may be learned from and updated with observed data.
- objective function--a mathematical representation capturing the goals of the IA.
- specification--the definitions of the environment, the IA's sensors and actuators, and the internal model and objective function necessary to operate and (learn to) perform a particular task.
- artificial intelligence--the study of how to design IAs that simulate, approximate, or surpass the precise capabilities of human intelligence.

In recent years, the rapid advent of AI functionality across societal domains has motivated the formulation of principles and definitions that consider such artifacts in their system setting. Here we include definitions adopted by the OECD in 2019 (OECD 2021).

- AI system--a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy.
- AI system lifecycle--involves: i) 'design, data and models'; which is a context-dependent sequence encompassing planning and design, data collection and processing, as well as model building; ii) 'verification and validation'; iii) 'deployment'; and iv) 'operation and monitoring'. These phases often take place in an iterative manner and are not necessarily sequential. The decision to retire an AI system from operation may occur at any point during the operation and monitoring phase.
- AI knowledge--the skills and resources, such as data, code, algorithms, models, research, know-how, training programs, governance, processes and best practices, required to understand and participate in the AI system lifecycle.
- AI actors--those who play an active role in the AI system lifecycle, including organizations and individuals that deploy or operate AI.
- stakeholders--all organizations and individuals involved in, or affected by, AI systems, directly or indirectly. AI actors are a subset of stakeholders.
- stakeholder--a person or entity with a vested interest in the AI system's performance and operation.

Today, this system lens to AI is largely inspired by the field of "AI Safety" and the associated technical project of "value alignment", which aims to build "provably beneficial" systems that learn the precise preference structures of humans (Russell 2017). Value alignment assumes that such a deterministic description already exists or is discoverable by artificial agents, and if we create precise mechanisms for learning it, then it could be modeled under mathematical conditions of uncertainty.

- AI Safety--the interdisciplinary study of how to build systems that are aligned with the structure of human values, in particular those of stakeholders whom the system is meant to serve.
- value alignment--the creation of systems whose specification is sufficient to learn the structure of human values.

In practice, AI research is as much about redefining philosophical concepts in the context of AI as it is about solving particular engineering and computer science challenges. But there is a fundamental gap between the idea of value alignment and managing the actual consequences of deployed systems. Decades of research in systems engineering for safety-critical systems has shown that values, such as safety or fairness, are an *emergent* property that "arise from the interactions among the system components" (Leveson 2012). Here, the system boundary and its components entail both technical elements or intelligent agents, as well as human agents, processes and supporting infrastructure.

> The emergent properties are controlled by imposing constraints on the behavior of and interactions among the components. Safety then becomes a control problem where the goal of the control is to enforce the safety constraints. Accidents result from inadequate control or enforcement of safety-related constraints on the development, design, and operation of the system. (Leveson 2012).

The emergent and dynamic nature of values, and the inability to "discover" or formalize these in the technical logics of a system, is corroborated by a long tradition of research in computer-supported cooperative work (Greenbaum 1992), human-computer interaction (Shilton 2018), and participatory design. As Halloran et al. conclude, "values emerge, whether you look for them or not" (Halloran et al. 2009). This problem resonates with the classic problem space of cybernetics: continuously interrogating and elaborating the relationship between actions and goals through forms of feedback rather than a deterministic problem formulation or static representation of value (Wiener 1988, Von Foerster 2007). In cybernetics, performance thresholds are determined through the concrete outcomes of actions taken, rather than precisely-defined capacities of the agent that reflect some stylized view of intelligence. This entails looking at the level of systems, composed of integrated components, and how values such as safety are instantiated and maintained through conditions of stability.

- feedback--information about the results of an agent's or system's actions which can then be taken as inputs for future actions, serving as a basis for improvement or stability.
- cybernetics--the interdisciplinary study of how systems behave in response to feedback.

As pragmatist philosophers and sociotechnical scholars have long emphasized, bridging the gap between design principles and real-world system performance requires specifying the normativity of the problem domain in terms of acceptable behaviors and outcomes (Pask 1976, Dewey 1896). On this interpretation, cybernetic feedback is needed to bridge the gap between problem formulation and defining the system's interface with reality. The question is: are norms something that can be passively learned by an agent, or something enacted through new forms of feedback? The former implies uncertainty about norms that in principle could be modeled by e.g. learning a reward function that represents human preferences. The latter however suggests indeterminacy that cannot be resolved without a broader system lens to instantiate design or governance norms.

Recent work in AI Governance suggests the latter. As argued by Wallach and Marchant, the most pressing regulatory questions will require new institutional entities tasked with articulating metrics, standards, or new forms of domain expertise to determine acceptable performance thresholds for particular AI systems (Wallach 2019). This may include governance coordination committees (Cihon 2019), an International Artificial Intelligence Organization (Erdelyi 2018), the Facebook Oversight Board (Klonick 2019), judicial oversight in the spirit of the EU General Data Protection Regulation (Voigt 2017), issues studied by the National Institute of Standards and Technology (Smuha 2021), "arms race" scenario modeling (Zwetsloot 2018), and many others. This emerging literature seeks to resolve situations of indeterminate system performance at various levels of normative abstraction, ranging from individual privacy to global security concerns.

- sociotechnics--the relationship between a system and real-world conditions, whose specification requires active engagement with the concerns of stakeholders.
- normativity--the reciprocal expectations of agents to conform to particular agreed-upon standards of behavior in a given domain.
- normative uncertainty--the unknown features of an environment that the agent must learn in order to behave optimally.
- normative indeterminacy--the lack of prior standards or forms of consensus for the sociotechnical context of a given system, rendering the specification problematic or incomplete.

Contemporary sociotechnical concerns about the development of AI systems share a common theme: data accumulation, increasing computational capacity, and new algorithmic learning procedures are reconstituting the normative systems in which humans live (Yeung 2017, Seaver 2019, Gillespie 2014). In this sense, the problem space of AI Safety is rediscovering cybernetics on new ground. There is an emerging need for sociotechnical specifications that are able to diagnose and resolve undesirable system performance, semantic equivocations, and political conflicts. This requires a principled elaboration through which an AI system's technical specifications (i.e. its model, objective function, sensors, actuators) are interpreted in light of salient normative considerations and real-world performance thresholds that stem from the social and situated context in which the system operates. Without clarifying this landscape, it will not be possible to evaluate whether particular governance mechanisms at different institutional scales are more or less appropriate for addressing the indeterminacies at stake.

- featurization--the system's capacity to represent features of the environment in order to achieve a specified goal.

- optimization--the designer's capacity to articulate how to more efficiently (e.g. cost minimization) or appropriately complete a task.
- integration--the capacity of users, managers, regulators and stakeholders to oversee and incorporate the system's real-world performance.
- sociotechnical specification--the proposed normativity of an AI system in terms of its featurization, optimization, and integration, defining who it is meant to serve, its purpose, and how it is to be evaluated and held accountable.

As Erdelyi and Goldsmith note, "the choice between harder and softer types of legalization [of AI systems] involves a context-dependent tradeoff, which actors should carefully consider on a case-by-case basis" (Erdelyi 2018). To weigh such tradeoffs, it must first be possible to index values and norms in terms of technical decisions about the system specification. This means that normative concerns of comparable significance and scope must be rendered commensurable in order for a responsible tradeoff to be struck and translated to a system's specification. Ruth Chang has highlighted the related philosophical notion of parity (Chang 1997, 2002), which holds that humans are able to articulate evaluative differences to make comparisons between incommensurable values or options (Chang 2017). This permits deliberation regarding an agent's overarching goals. Parity is constitutive of what Chang calls *hard choices*: when different alternatives are on a par, "it may matter very much which you choose, but one alternative isn't better than the other [...] alternatives are in the same neighborhood of value [in terms of how much we care] while at the same time being very different in kind of value". Note that while Chang developed the notion of parity and hard choices for an individual agent or authority weighing different options or values, we reinterpret these concepts in a setting comprising different stakeholders, denoting a distinct form of agency. This renders the weighing of options or values, and thereby notions of parity and hard choices, as inherently *political* as different stakeholders will have different interests, varying political power and potentially diverging ideas about evaluating different problem formulations, solution directions and associated values or principles (de Haan and de Heer 2015). We acknowledge recent empirical insights from Van der Voort et al., who debunk the rational view typically assumed for decision-making. They show how algorithms and big data analytics encounter political and managerial institutions in practice, leading to a spectrum of possible outcomes or theses for how the technology is specified and used (van der Voort 2019).

- comparability--the evaluation of an AI system's technical capacities (e.g. learnable features) as similar to each other in their magnitude, relevance, or problem stakes.
- incommensurability--the evaluation of an AI system's normative capacities (e.g. relationship with users or designers) as not able to be measured by the same standard.
- parity--A relation between values that are comparable in significance but unable to be directly measured as better, worse, or equal to each other.
- hard choices--Situations of value parity in AI system development, which require deliberation in order to make the options technically commensurable.
- machine politics--a mode of deliberation associated with the stakes involved in hard choices, in particular between or among parties that have power over their resolution.

The possibility of hard choices when designing AI systems suggests the need for a principled diagnostic approach, folded into development practices. This approach would specify commitments that match appropriate modalities of algorithmic governance with the potential

harms faced by stakeholders. The goal would not be for developers to make choices on stakeholders' behalf, but for developers to adopt diagnostic practices so that choices can be proactively anticipated and resolved through feedback. As argued by Elizabeth Anderson (Anderson 2006), the form of feedback particular to modern democracies is dissent, indicating that the current specification (e.g. of a law) is problematic and must be amended or rejected. Accommodating dissent as a type of feedback particular to machine politics is thus a path to enacting appropriate features as well as proportional mechanisms for democratic governance, denoting a possible alternative form of design practice.

- commitment--a pledge made by developers to stakeholders about the sociotechnical specification of an AI system, in terms of how it is intended to operate.
- dissent--purposive feedback that lies outside the distribution of previous inputs, serving to challenge the grounds for consensus on system specification.
- cybernetic practice--active attention to the types of feedback needed to address normative indeterminacies and refine the sociotechnical specification of a particular AI system.

Revealingly, the field of cybernetics also applied feedback to cybernetic practices themselves, which culminated in so-called "second-order cybernetics" (Glanville 2004). We embrace the spirit of this tradition, as well as later work proposing such reflective inquiry on technical practices in AI (Agre 1997).

From this lexicon, we conclude that recent work in AI governance and AI Safety reveals a need for:

1. a sociotechnical reframing of classic problem domains in AI (agency, models, representation, learning), in terms of how human behaviors and institutions will be indeterminately reshaped by designed systems.
2. a shared language to diagnose different kinds of normative indeterminacy, both between intended vs. actual system behavior and across communities of stakeholders.
3. the specification of requisite feedback modalities, in order for the system to achieve appropriate stability in the face of operational indeterminacies.

4.2 The Problem of Vagueness

As AI systems are applied to more sensitive contexts and safety-critical infrastructure, normative indeterminacies are becoming more visible. Identifying the missing feedback in a given specification requires interrogating the functions of an AI system in a principled manner. This includes examining what task the AI system is trying to complete and how the system is meant to work in support of human contexts, as well as which normative standards would be appropriate to fulfill these needs.

Here I compare prominent technical and policy standards that have been proposed, revealing each as a partial response to the underlying problem of *vagueness*. The vagueness of a system specification is the ultimate source of the normative indeterminacies at stake. Vagueness is a central topic in metaphysics and the philosophy of logic and language that has important application in system engineering and artificial intelligence (Agre 1997). It is about the

fundamental lack of clarity in our relationship with the world, either in terms of the ways we are able to perceive it, the language we use to describe it, or in the world itself. It is addressed through the drawing of boundaries--forms of classification, demonstration, analogy, and other rhetorical strategies that sort phenomena into particular qualities and quantities or draw distinctions of form and content (Williamson 2002). A classic example is the Sorites paradox: which grain of sand removed from a heap turns the heap into a non-heap? Such situations may yield existential uncertainty, which, if not resolvable through agreed upon standards, may lead to arbitrary tradeoffs, compromise, or restrictions. We thus propose vagueness as a general descriptor for situations in which developers' attempts to model some domain via technical uncertainty fall short and give way to specific forms of indeterminacy.

For each approach to indeterminacy present in the current AI policy and governance literature, we first organize and present the corresponding classical interpretation of vagueness, namely either *epistemicism*, *ontic incomparabilism*, or *semantic indeterminacy* (Chang 2002). We then isolate the respective standards that have been unreflectively derived from these schools of thought, namely *metanormativism*, *value pluralism*, and *fuzziness*. Finally, we identify the stylized form of feedback that each school of thought prioritizes over others to enact these standards, namely *preference learning*, *refusal*, and *equitable outcomes*. I concisely summarize these relationships in Table 1. This exercise motivates the need for sustained engagement with the actual context of system development.

Epistemicism - resolving vagueness through model uncertainty

Epistemicism claims bivalence as a basic condition for an object's existence (Schiffer 1999). This is to say that for any given property of an object, there is in principle some sharp boundary by which the object either does or does not have that property. Illustrated through the Sorites paradox, epistemicists believe that there is an objective fact of the matter about the precise number of sand grains necessary to constitute a heap vs. non-heap, even though we may be ignorant of that cutoff point. The position thus holds that every object property or attribute must terminate at some boundary, no matter how inappreciable this boundary may be at present.

This implies that acquiring more information may help reveal where the boundary actually is or could be drawn. Pure epistemicism is counterintuitive and is philosophically controversial in comparison with the claim that boundaries are semantic constructions (Gomez 1997). But the essence of the position is simply that if distinct communities (or even the same person) claim the same property applies to the same object in different ways, then they are either ignorant about the property's actual boundary or are describing distinct objects.

Epistemicism has a powerful affinity with *metanormativism*, the notion that the criteria for rational decision-making are not fully known or confidently expressed because sufficient information about the environment, other agents, or oneself is absent. Because epistemicists believe that no comparable options are fundamentally "apples and oranges", as there must be some degree to which one is preferable over the other, metanormativism asserts the existence of a clear, positive value relation between available ethical actions: one must be unambiguously better, worse, or equal to the other for a given choice to be demonstrably rational. For example, William MacAskill has sought to articulate "second-order norms" that guide how one should act when multiple

appealing moral doctrines are available (MacAskill 2019). MacAskill, whose work has been cited in support of technical work on AI value alignment and value learning (Soares 2014, 2015), has also proposed a "choice worthiness function" that would generate reward functions in an "appropriate" manner, where appropriateness is defined as "the degree to which the decision-maker ought to choose that option, in the sense of 'ought' that is relevant to decision-making under normative uncertainty" (MacAskill 2016). As such, metanormativism is a natural ally of expected utility theory and in particular the first axiom of the Von Neumann-Morgenstern utility theorem, specifying the completeness of an agent's well-defined preferences (von Neumann and Morgenstern 2007).

Distinct approaches to AI Safety have emerged to define the uncertain scale at which AI systems may cause social harm. At one end of this continuum is *existential risk* (hereafter referred to as x-risk), i.e. the effort to mathematically formalize control strategies that help avoid the creation of systems whose deployment would result in irreparable harm to human civilization. The x-risk literature has focused on the "value alignment problem" in order to ensure that learned reward functions in fact correspond with the values of relevant stakeholders (such as designers, users, or others affected by the agent's actions) (Soares 2015). Here the reward function serves as a representation of stakeholder preferences rather than the AI agent's own objective function, an assumption common in inverse reinforcement learning (Hadfield-Menell 2016). This position is also practically adopted by software engineers and tech enthusiasts for whom the uncertain specification of human preferences comprise an investment opportunity for new AI systems. The following quote from Mark Zuckerberg is illustrative: "I'm also curious about whether there is a fundamental mathematical law underlying human social relationships that governs the balance of who and what we all care about [...] I bet there is" (Hildebrandt 2019).

The promise of such a function continues to provide guidance for designers and AI researchers about what decision procedures are acceptable or unacceptable for the system to follow, specifically when the goal state and risk scale are difficult to define (Hadfield-Menell 2019, Irving 2019). This research agenda prioritizes *preference learning*, the systematic observation of user behavior and choices to learn an underlying reward function, as the most salient form of design feedback for filling the gaps in system specification (Hadfield-Menell 2016). As stated by Stuart Russell (Russell 2019):

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

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

Ontic incomparabilism - respecting value pluralism

Meanwhile, *ontic incomparabilism* holds that there are fundamental limits to what our predicates or semantics can make of the world because there is no objective basis to prefer one definition of a concept to another (Barnes 2011). More concretely, even if we knew everything about the

universe, there would still be no way to argue that a pile of sand "should be considered a heap" after exactly n+1 grains as opposed to after n grains. Ontic incomparabilism therefore claims that we cannot ever fully model the world by discovering additional criteria or accumulating sufficient information about it as its dynamics may be fundamentally unsuited to model specification. Note that this position is distinct from views that the world is impossible for human minds to comprehend completely (as has been argued for specific physical phenomena, e.g. quantum mechanics) or that the world is impossible to describe accurately. Instead, the claim is that any finite number of descriptions or representations cannot exhaust the world's richness because its basic features are not readily discernible, and that there are in principle as many different ways of representing the world as there are agents capable of realizing their agency in that world. This means that modeling the world robustly would require securing the world's total cooperation with the boundaries being drawn over it.

Ontic incomparabilism has found expression in terms of *value pluralism*, i.e. that there cannot or will never be an ultimate scheme for delineating human values because humans exist in the world in a way that cannot be exhaustively represented. This transcends sociological fact (i.e. that people hold different beliefs about values, and value beliefs differently) to make an axiological, anti-monist claim: values are indeterminately varied and incommensurable, and no ethical scheme could ever account for the range of values or concerns held by all humans for all time (MacAskill 2013). Value pluralism is widely adopted by queer theorists who highlight how formal value specifications typically exclude certain subpopulations in favor of others (Keyes 2019). For example, Kate Crawford has endorsed Mouffe's (1999) concept of "agonistic pluralism" (Mouffe 1999) as a design ideal for engineers (Crawford 2016), while Hoffmann argues that abstract metrics of system fairness fail to address the hierarchical logic that produces advantaged and disadvantaged subjects and thereby disproportionately put safety harms on already vulnerable populations (Hoffmann 2019). Mireille Hildebrandt has taken these perspectives to their logical extreme and advocates for "agonistic machine learning", suggesting that the human self should be treated as fundamentally incomputible (Hildebrandt 2019).

These conclusions have found support in the field of Computer Supported Cooperative Work (CSCW). Presenting them as a central challenge, Ackerman has described the inevitability of the "social-technical gap" of computer systems; the inherent divide between what we know we must support socially and what we can support technically (Ackerman 2000). This frames the central danger in terms of software engineers who neglect certain value hierarchies, either by failing to interrogate the context of historical data or external cost biases through design choices that moralize existing structural inequalities (Eubanks 2018). The call to value pluralism, as such, is not opposed to pragmatism in the form of external mechanisms that regulate how our diverse commitments may be reconciled (James 1896), nor to the creation of systems that make use of necessarily limited models in pursuit of stable behavior. Rather, as designers compromise the public interest through incomplete model specifications that create external costs for society, they have merely reframed the central problems of modern political theory (Dewey 1954) and inherited the hallmarks of structural inequality. The history of social technology, from the modern census to the invention of writing, is saturated with ways in which forms of human identity were problematically obfuscated or delimited rather than protected or left undetermined (Benjamin 2019). This phenomenon underpins foundational concepts of twentieth century social theory

(Krais 1993) and deconstructionist critiques of Western philosophy as a "metaphysics of presence" (Heidegger 1962).

On this view, any model design requires fundamental political choices about how values of relevant stakeholders, including those indirectly affected by the system, result in some value hierarchy that may have undesirable consequences for how the benefits and harms of system behavior are distributed across society. Correspondingly, the type of feedback most readily endorsed by ontic incomparabilists has been *refusal*, i.e. the explicit rejection of a model specification as unsuitable. This has been expressed recently through comparisons of facial recognition systems with plutonium (Stark 2019), algorithmic classification with a new form of "Jim Code" (Benjamin 2020), and refusal itself with the notion of feminist data practice (Garcia 2020). However, a major open question is how or whether refusal itself can lead to the articulation of a more just and equitable society in the absence of alternative forms of feedback.

Semantic indeterminism - declaring things fuzzy by nature

Finally, *semantic indeterminism* asserts that the extent to which we can determine the definition of a concept is the extent to which the members of a given community agree on that definition. Commonly associated with Wittgenstein (Wittgestein 1953), this position emphasizes the rules of language-games as defining how we refer to the world and the specific boundaries of a given community's concerns, social tastes, and modes of valuation. To again illustrate this via the Sorites paradox: Persians, Romans, and even distinct Greek city-states may use alternative definitions of "heap" and thus confidently draw different cutoff points without ontological disagreement. Semantic indeterminism does not argue for a radical version of social constructivism according to which any claim to describe reality is arbitrary or fictional, e.g. the notion that claims about the objective world are impossible. Rather, such claims simply cannot be interpreted outside the rules that particular language communities have adopted and refined over time.

Semantic indeterminism can be further illustrated through the formal assumptions of *fuzziness*, which hold that our ways of talking about "the world" admit non-binary variations (e.g. the variable *age* including the values "somewhat young", "nearly middle-aged", "centenarian", or "newborn") that are regulated and modified within distinct language communities or modes of expertise. Fuzziness deals with the contingencies at stake in conventional approaches to set membership and truth-value. As a mode of reasoning that addresses uncertainty and vagueness, it can refer either to the membership of an event in a vaguely defined set--the purview of fuzzy logic (Gerla 2016)--or to the indeterminate features of the world itself, to which linguistic terms make limited (although meaningful) reference. In other words, fuzziness captures how the imprecision of language can be due either to a given community's epistemic limitations (which result in a form of uncertainty and partial knowledge intractable through other forms of logic), or how any semantics is a (necessarily) limited picture of the inherently complex and continuous nature of reality.

For my purposes, fuzziness makes semantic indeterminism institutionally tractable and amenable to elaboration by stakeholders. For example, Lessig's famous modalities of regulation (laws vs. norms vs. markets vs. architecture) show how fuzziness can be distinctively enacted or revisited by human forms of infrastructure and decision-making (Lessig 2009). In the context of AI Safety,

an exemplary discourse has formed within the Fairness, Accountability and Transparency in Computing Systems (FAccT) literature. FAccT research has harvested a multitude of definitions and tools aiming to address safety risks by diagnosing and reducing biases across various subgroups defined along lines of race, gender or social class (Narayanan 2018).

While scholars have pointed out the critical and mathematical shortcomings of abstract definitions for bias mitigation, these are still instrumented in practice as means to resolve fuzziness in particular application domains. For example, industry efforts have embraced bias tools to generate feedback for *equitable outcomes* as a means to engender trust in a given system, while global efforts aim to codify algorithmic bias considerations into certifiable standards "to address and eliminate issues of negative bias in the creation of [...] algorithms" (IEEE 2021).

Still, the tension between eliminating bias and winning social trust reveals the inconsistent determinations of what safety means throughout the entire lifecycle, including which norms should guide design and use decisions. As some argue, "it is important to acknowledge the semantic differences that 'fairness' has inside and outside of ML communities, and the ways in which those differences have been used to abstract from and oversimplify social and historical contexts" (Rea 2020). Scholars have also emphasized important semantic differences and connections between "individual" and "social" fairness that could help clarify and procedurally reshape the way formal fairness criteria are reconciled with policy objectives (Corbett-Davies and Goel 2018, Binns 2018). However, incorporating these semantic differences would mean accommodating additional types of feedback, such as *preference learning* to represent what people actually seem to want as well as *refusal* to serve as a check on the system's tendency to occlude or suppress neglected values. Thus, semantic indeterminism does not resolve the normative indeterminacies raised by epistemicism and ontic incomparabilism. Instead, fuzziness interprets the limits of language as conditioned either on the complexity of the world or our epistemic limitations. Consequently, ambiguities of language must be deferred for designers and stakeholders to deal with, rather than decided in advance of inquiry. This is clearly exemplified in the EU's recent proposal for regulating AI systems in high-stakes domains, in which the need for "harmonised standards" is advocated, stating that "[t]he precise technical solutions to achieve compliance with those requirements may be provided by standards or by other technical specifications or otherwise be developed in accordance with general engineering or scientific knowledge at the discretion of the provider of the AI system."

Instead, I propose fuzziness as a *sociotechnical* commitment to AI development as an unavoidably iterative, interactive, and above all deliberative process of inquiry. This captures the reality that systems' "core interface consists of the relations between a nonhuman system and a human system" (Trist 1981), with various dimensions (e.g. users, citizens, operators, regulators), whose construction is hindered by limited knowledge, subject to error, of how key technical innovations bear on human contexts. Even carefully-designed formalisms that are sensitive to the implicit concerns of human agents are not guaranteed to learn the right preference structures in the right way without new forms of surveillance, control, and assigned roles for both humans and the systems themselves (Eckersley 2018, Agre 1994). Such system setups are limited in three ways: (1) they can never formalize everything, and require subsequent developers to organize around them; (2) they attempt to resolve (and thereby confuse) content and procedure from the get-go,

rather than treat the sociotechnical development of AI systems as a dynamic problem; and (3) they are limited in addressing wider spectra of values across distinct peoples and cultures.

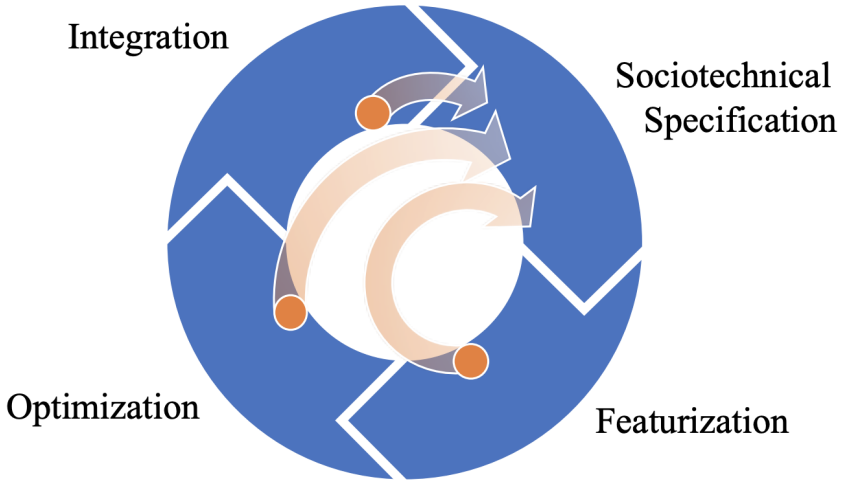
Type of Vagueness	Normative Standard	Mode of Feedback
Epistemicism	metanormativism	preference learning
Ontic Incomparabilism	value pluralism	refusal
Semantic Indeterminism	fuzziness	equitable outcomes

Relationship between types of vagueness, the corresponding normative standard, and the types of feedback each prioritizes.

4.3 A Framework of Commitments for AI Development

As outlined in the previous section, matching safety principles with technical development procedures is fraught with hard choices. There are inherent sources of vagueness about what safety means, how it is formalized, and how it is enacted in an AI system. As a result, indeterminacies are encountered through possible design interventions that are technically comparable but normatively incommensurable. If left unaddressed or underconsidered, these may lead to harms, reinforcement of structural inequalities, or unresolved conflict across different stakeholders. The section on vagueness thus analyzed a broad spectrum of technical, governance and critical scholarship efforts to address the safety of AI systems, and how these fall in three canonical approaches to vagueness. For each lens, I determined the affordances and limitations of their associated cybernetic feedback modalities and the interventions that can be done with these to safeguard an AI system or improve the practices that design or govern it.

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



The cyclical practices in AI system development. Orange circles denote the occurrence of "hard choices", moments where normative indeterminacy arises and which provide opportunities for deliberation in the form of machine politics. This may include revisiting and altering sociotechnical specification.

The resulting *Hard Choices in AI (HCAI) Framework*, presented in Figure 1, contains four cybernetic practices: sociotechnical specification, featurization, optimization, and integration. These activities and corresponding commitments will be introduced and discussed in the following subsections. I stress that this framework is a conceptual depiction of how to deliberate critically and constructively about normative indeterminacy. The framework may however help to identify concrete design approaches that can put commitments in action. In many instances, regulatory measures may form either an existing source of constraints and requirements in the development process, or be informed by it. I do not advocate for particular law or policy interpretations, as these are just as contextual as design approaches, but see such translation work as a natural extension of this chapter. My framework naturally connects with and further concretizes the 'AI system lifecycle' as introduced in the OECD AI Principles (OECD 2021).

Sociotechnical Commitments

Developers must diagnose situations of normative indeterminacy while remaining attentive to the fundamental limitations of technical logics to resolve them. This necessitates an "alertness" to all the factors responsible for the situation, including social, affective, corporeal, and political components (Amrute 2019). AI systems are not merely situated in some pre-existing sociotechnical environment. Rather, the development of the system itself creates novel situations that intervene on social life, reflected in the distinction between pre-existing, technical, and emergent bias (Friedman 1996). These require their own formal treatment (Dobbe 2018).

Furthermore, major stages of AI system development require feedback channels for stakeholders to assign appropriate meaning to possible specifications. In particular, we emphasize the need for dissent mechanisms to help surface parity of different design options and their related value hierarchies. This permits indeterminacies to be diagnosed and resolved via machine politics rather than the whims of designers, systems engineers, or some other narrowly technical constituency.

As Anderson emphasizes, in contexts where a policy is set by a majority or powerful player, "[s]uch dissent is needed not simply to keep the majority in check, but to ensure that decision-making is deliberative—undertaken in an experimental spirit—rather than simply imposed" (Anderson 2006). These channels make AI development an opportunity for communities to reimagine their own moral boundaries.

Developers must also acquire practical reasoning to navigate across sociotechnical approaches to a problem and determine specifications accordingly. A specification that might make sense in one context may not make sense for another, either in terms of feature detection (e.g. facial vs. handwriting) or integration scale (municipal oversight vs. nationwide surveillance). Developers must recognize the differences between these and internalize standards that guide the indeterminate application of abstract principles to the concrete needs and demands of the situation, in a manner responsive to stakeholder feedback. These comprise distinct forms of judgment: formulating the problem, evaluating system criteria, and articulating the performance thresholds that the system must meet in order to be safe. We agree with Philip Agre that this engagement requires "reflexive inquiry [that] places all of its concepts and methods at risk [...] not as a threat to rationality but as a promise of a better way of doing things" (Agre 1997).

At distinct moments of formal specification, we ask: (1) at what development stages and associated cybernetic practice might indeterminacies manifest and what forms may parity take? (2) in what concrete ways are feedback mechanisms/interventions needed to address these issues? (3) what form does the associated canonical dilemma take? (4) what forms of judgment are needed to interpret stakeholder feedback and effectively manage the indeterminacies and dilemmas that the system generates? These points are presented in Table 2.

Cybernetic Practice	Intervention	Dilemma	Sociotechnical Judgment
Sociotechnical Specification	integral	inclusion vs. resolution	solidarity
Featurization	epistemic	underfeaturized vs. misfeaturized	context discernment
Optimization	semantic	verification vs. validation	stewardship
Integration	ontic	exit vs. voice	public accountability

Relationship between cybernetic practices, normative interventions, hard choice moments requiring feedback, and forms of sociotechnical judgment needed to interpret feedback.

Sociotechnical Specification (engaging the "stakes" and forms of agency)

The HCAI Framework does not define a determinate start of AI development, but it does require the initial determination of how the problem is to be formulated and tackled, mechanisms for improving this determination through feedback and dissent, and what stakeholders are already implicated or should be involved in problem formulation. Moreover, not all normative dimensions can be foreseen upfront, as hard choices may surface in subsequent development considerations. Aware of these historical, critical, and empirical complexities, we center the need for *sociotechnical specification*, i.e. the process of facilitating the different interests relevant in understanding a situation that may benefit from a technological intervention. Developers must clarify what the system is actually for--whose agency it is intended to serve, who will administer it, and what mechanisms are necessary to ensure its operational integrity. The sociotechnical specification facilitates *integral* interventions to determine and resolve what safety means

(semantic), how it is formalized (epistemic), and how it is enacted in a system (ontic). This facilitation cannot fall exclusively on the plate of designers or developers.

To appropriately surface parity throughout sociotechnical specification, the following challenges must be taken up: (1) negotiate a program of requirements and conditions on both process and outcomes; (2) determine roles and responsibilities across stakeholders; (3) agree on ethics and modes of inquiry, deliberation, and decision-making. In sociotechnical specification, one needs to understand the context of integration. This includes the positions of different stakeholders with their reasoning and how these relate to each other. It requires an understanding or anticipation of the impacts on social behavior, broader societal implications, and how different solutions would sit within existing legal frameworks. This yields the following dilemma:

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

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

The key hard choice for a successful AI system is to include sufficient perspectives and distribute decision-making power broadly enough in development to cultivate trust and reach a legitimate consensus, while resolving the situation in a set of requirements and a process with roles and responsibilities that are feasible. While we propose these diagnostic and procedural questions for AI system applications broadly (and prospectively for more computationally intensive systems in the future), here we focus our attention on contexts that are safety-critical by nature or play an important public infrastructural role. This includes systems that integrate on a global scale, interacting with a wide spectrum of local and cultural contexts.

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

Solidarity should not be understood as conflating the interests of designers and stakeholders. Rather it motivates the former to create channels for stakeholders to actively determine, rather than

passively accept, the system specification (Unger 1983). Here we endorse Irani et al's vision for postcolonial computing, which "acknowledge[s] stakeholders as active participants and partners rather than passive repositories of 'lore' to be mined" (Irani 2010).

Featurization (epistemic uncertainty)

AI systems generally represent a predictive, causal or rule-based model, or a combination thereof, that is then optimized and integrated in the decision making capabilities of some human agent or automated control system. As such, it has to answer the question 'what information it needs to "know" to make adequate decisions or predictions about its subjects and notions of safety?'. As the model represents an abstraction of the phenomenon about which it makes predictions, the chosen model parameterization and the data used to determine parameter values delimit the possible features and value hierarchies that may be encoded. If not anticipated and accounted for, this may deny stakeholders the opportunity to evaluate design alternatives and force potentially harmful and unsafe hard choices. In this way, featurization is an *epistemic* intervention on the indeterminacies that may be present or latent in the context that precedes or follows system operation.

To surface the parity at stake in featurization, the following challenges must be taken up: (1) make explicit and negotiate what can and cannot be modeled and inferred, crystallized in the underfeaturized/misfeaturized hard choice; (2) engage stakeholders to challenge and inspire modeling assumptions to ensure application aligns with contextual expectations; (3) validate the design with stakeholders to anticipate possible value conflicts that can arise due to the gap between model and world and plurality of values during deployment, preparing to revisit the modeling tools and methodology. Featurization specifies the computational powers of the system: how the limits of what it can model determine its assumptions about people and the broader environment, and what kinds of objects or classes are recognizable to it. At a minimum, stakeholders must resolve the following dilemma:

Underfeaturized: What possible input variables or model parameterizations do we choose not to include? What features will the model not be able to learn that may in fact be open to deliberative re-evaluation?

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

The danger lies in failing to adopt model parameters that are both computationally tractable and normatively defensible. Given finite time and material resources as well as the vested interests of specific stakeholders, this may err towards under- or mis-specification in ways that developers cannot perfectly anticipate. The spirit of the hard choice is crystallized differently in distinct algorithmic learning procedures. For example, the division between model-based and model-free reinforcement learning essentially bears on what kind of control system is being designed and, respectively, whether this specification establishes a permissible space in which a given problem can be formulated and represented causally or merely defines permissible predictive signals (e.g. rewards, elements, qualities) within the environment. At least some corresponding domain features

may be made computationally tractable and suited to optimization despite being experienced by stakeholders as incommensurable. Or some features may be technically obfuscated despite their mutual comparability and integrity in lived experience. An often returning example of this dilemma is the need to interpret or explain the decision-logic of an AI model. While deep learning models may offer a higher performance, this need may lead to opting for a lower complexity model that has more potential for forms of accountability.

The model must be capacious enough to represent the nature of the environment in a way that safeguards stakeholders' interests. But its training must also be constrained enough to be tractable, guarantee performance (Achiam 2017), and preserve privacy boundaries. Imposing modeling constraints necessarily creates technical bias, which may take away space for stakeholders to express or protect their own specific values in terms of the phenomena permitted or excluded by the model's system boundaries (Dobbe 2018). There is already some technical work acknowledging this as a formal dilemma with no optimal solution in the context of reinforcement learning (Choudhury 2019, Yu 2019). But the deeper sociotechnical point is that the criterion for these constraints, which entail a choice of the moment at which a model must remain technically ignorant or intentionally suboptimal, must be specified in terms of a commitment to the self-determination of stakeholders.

Featurization requires *context discernment*, the disqualification of specific features and modeling choices that, while technically proficient, are judged to be sociotechnically inappropriate within the problem space at hand. Here we draw from (Dreyfus 2011): "The task of the craftsman is not to *generate* the meaning, but rather to *cultivate* in himself the skill for *discerning* the meanings that are *already there*." Featurization is about anticipating how the model would interact with the context of deployment, how else it could be (mis)used, what bias issues may arise during training, how to protect vulnerable affected groups, and how learned objective functions may generate externalities. In the event no consensus is reached and dissent persists, the option of not designing the system should be preserved (Baumer 2011).

Optimization (semantic indeterminacy)

The parameters of the system's internal model must be further determined by performing some form of optimization. This determines the input-output behavior of the model and how it will interact with human agents and other systems. Optimization extends across the design stage (e.g. training an algorithm) and implementation (e.g. finetuning parameters) and answers the question 'what criteria and specifications are considered to measure and determine whether a system is safe to integrate?'. Depending on the chosen representation, such optimization can either be performed mathematically, done manually through the use of heuristics and tuning, or some combination thereof. For mathematical optimization, the recruitment of historical and experimental data is needed to either (a) infer causal model parameters (e.g. for system identification, an inference practice common in control engineering (Guo 2018), (b) infer parameters of noncausal representations, or (c) iteratively adjust parameters based on feedback (as in reinforcement learning). The objectives and constraints and the choice of parameters constitute a *semantic* intervention on how the identification of specific objects relates to the forms of meaning inherited by and active in the behavior of stakeholders themselves.

Therefore the following challenges must be taken up: (1) assess the extent and limitations with which the optimization criteria and procedure can translate and respect specifications, crystallized in the validation/verification tradeoff; (2) codify a validation procedure for empirical criteria that conforms to stakeholders' specific concerns, addressing specifications not covered through mathematical optimization; (3) adjudicate and modify verification and validation strategies over time as indeterminacies of featurization and integration continue to be highlighted. To declare a system safe it must go through a process of verifying and validating its functionality, both of itself as an artifact as well as integrated in the context of deployment. This is done with the help of engineers and domain experts who interface between the problem the system is meant to solve and the workings of the system itself. Here, the minimum requirements for safe outcomes are impartial assessments of the following questions/dilemma:

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

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

This hard choice poses several concrete challenges for development. First, systems that are mostly optimized in a design or laboratory environment fall inherently short as their data cannot fully capture the context of integration. In the development of safety-critical systems, this design issue is acknowledged by the need to minimize any remaining errors in practice (through feedback control (Aastrom 2010) and putting in place failsafe procedures and organizational measures as well as promoting a safety culture. Second, accounting for interactions with other systems and human agents is not to be taken lightly and is heavily undervalued in current AI literature (Parasuraman 1997). For example, the overspecification of environments through simulation (as is now popular in the development of autonomous vehicles) may backfire if the optimization scheme overfits the model for features or elements that are not reflective of the context of integration. Third, a lack of validation and safeguarding systems in practice can result in disparate impacts (Barocas 2016) and failures. This is especially pertinent for underrepresented (and undersampled) groups that are often not properly represented on AI design teams (West 2019). For systems that are "optimized in the wild" with reinforcement and online learning techniques, these considerations are even more acute, although recent efforts have proposed hybrid methods that can switch from learning to safety-control to prevent disasters (Fisac 2018). This technical point, which mirrors the well known bias-variance tradeoff, becomes *sociotechnical* at the moment when the choice of optimization procedure is interpreted from the standpoint of jurisprudence applicable to the domain.

The cultivation of *stewardship* is needed to reconcile the technical problematics of value alignment with optimization procedures capable of providing qualitative assurances to the particular sociotechnical stakes of the domain, whether physical, psychological, social, or environmental. System engineers must internalize an understanding of how the finitude of their teams' tools and procedures bears on the urgency felt by stakeholders towards objects of sociotechnical concern, compelling attention to how sparse team resources should be allocated and complemented, rather

than to abstract notions of accuracy or efficiency. Only in this way can under- or mis-featurization risks be managed and mitigated without perverting intended stakeholders' semantic and moral commitments. The team must decide: what internal verification strategies might we need in order to safeguard the validations already endorsed by legal inquiry? Here "quality management" must be elevated to the contestation and adjudication of how (possibly pluralist) values are operationalized without compromising parity.

Integration (ontic incomparabilism)

Finally, as AI systems are rapidly introduced into new contexts, new forms of harm emerge that do not always meet standard definitions. In addition, the diversity of stakeholder expectations, as well as of environmental contexts, may challenge specifying safety for systems that are deployed across different jurisdictions. At a minimum, those developing and/or managing the system must specify mechanisms to identify, contest, and mitigate safety risks across all affected communities, as well as who is responsible for mitigating harms in the event of accidents. This can be done via general rules and use cases of safety hazards that identify terms of consent, ensure interpretive understanding without coercion, and outline failsafe mechanisms and responsibilities. Hence, such conditions should spell out both the technical mechanisms as well as the processes, organizational measures, responsibilities, and cultural norms required to prevent failures and minimize damage and harm in the event of accidents. Here we appropriate tradeoffs already identified by social theorists regarding the moral authority and political powers of social institutions (Flew 2009). This dimension serves as a decisive *ontic intervention* of what kind(s) of agency stakeholders possess as far as the system is concerned.

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

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

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

To the extent that proposed value hierarchies remain indeterminate beyond featurization and optimization, sociotechnical integration challenges systems to handle the multiple objectives, values, and priorities of diverse stakeholders. At stake here are the unexpressed moral relationships of subpopulations not originally considered part of the potential user base, who must bear the "cost function" of specification, as well as other forms of agency (animal, environmental, cybernetic) alien to yet implicated in system specification and creation. At a minimum, system administrators

must acknowledge that users will interpret the system agreement both as economic (acting as a *consumer*) as well as political (acting as a *citizen*). The developments on social media in recent years have taught us that these roles cannot be seen as mutually exclusive. The increasing dependence of the public on these platforms and their AI systems strengthens the need for voice (as exit options have become increasingly difficult or unlikely).

Administrators must cultivate *public accountability* to deal with these challenges, ensuring both Voice and Exit remain possible for stakeholders such that some criterion of trustworthiness is maintained. That is, anyone can leave the service contract if they want, but enough people choose to remain because they believe in their ability to express concerns as needed. Trustworthiness lies in supporting stakeholders' belief in their ability to exert different kinds of agency as they see fit, either within the system (by dissenting to its current mode of operation) or outside it (by choosing it through active use). This sociotechnical balance must hold regardless of the specific commitment being made. For example, service providers may specify some channel by which vulnerable groups can opt out of a publicly-operated facial recognition system (preserving Exit), or supply private contractors with a default user agreement that must be relayed to anyone whose data will be used by the system (preserving Voice). Either way, administrators must ensure they treat people both as respected consumers (a customer, client, or operator treated more or less as a black box) as well as citizens (a subject with guaranteed rights, among them the right to dissent to relevant forms of political power) in the context of the terms for system integration. Failure to have meaningful exit or voice can motivate collective action to reshape power relationships (Hirschman 1970), a phenomenon that has recently manifested when pushing back against harmful AI systems (Crawford 2019).

4.4 Implications and Discussion

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

Expand the boundary of analysis to include relevant sociotechnics - systems, organizations and institutions

Engineering and computer science disciplines have long tradition of working with "control volumes", which are mathematical abstractions employed to render problems and their solutions in terms of technical terms (Li 2007). In doing so, they allow a designer to decontextualize, depoliticize and ignore the history of a problem (Kadir 2021).

While often done in a more controlled context, the sociotechnical complexity and normative stakes of AI systems engaging in sensitive social and safety-critical domains requires a more comprehensive lens. An algorithm or AI system alone cannot engage with its inherent normativity. In contrast, studies in systems safety have shown that safety is inherently an *emergent property* that "arises from the interactions among the system components" (Leveson 2012). This requires a

system perspective that includes the human agents interacting with a technology (Green 2019), as well as how it is situated with respect to organizational processes (von Krogh 2018) and cultural and institutional norms (Gasser and Schmitt 2020). Such a systems lens also provides a more comprehensive starting point for controlling for safety, which is done by "imposing constraints on the behavior of and interactions among the components" of a system (Leveson 2012). This lens also explains how vulnerabilities of AI systems originate from across these components and system interactions, which corroborates insights from computer security that systems cannot be secured by addressing technical/mathematical vulnerabilities alone (Crawford et al. 2019, Carlini et al.). Lastly, a broader systems lens will be vital in understanding to what extent intended standards for AI systems can lean on general principles versus contextual needs and stakes specific to the domain of application. AI developers can lean on a long history in systems engineering of analyzing, modeling and designing sociotechnical systems, which should go hand-in-hand with a multi-actor approach (de Bruijn and Herder 2009).

Confront the choices and assumptions behind the AI system

Rather than addressing the limitations of a formalization itself, an honest encounter with normative indeterminacy deserves an account of the normative assumptions behind it and their implications. Recently, various scholars have advocated about the dangers of abstraction in AI systems (Selbst 2019), and pointed to the dangers of how imposing such abstractions can reify inequities resulting from institutional racism (Benjamin 2019) and harm marginalized communities (Bender 2021). Put bluntly, the choice of capturing everything in terms of an AI model and objective function is political. Cybernetic practices should make explicit where the boundary for acceptable formalization lies and what forms of feedback and evaluation are needed to safeguard their integration. Apart from their role in formalizing, modeling choices come with their own externalities. An inverse reinforcement learning procedure inherently requires the observation of detailed human behavior which might violate privacy norms (Raji and Dobbe 2020). And deep learning architectures are synonymous with extensive data gathering, which challenges privacy as well as environmental norms (Dobbe et al. 2019).

Orient development and governance towards a multi-actor approach with "problem and solution spaces"

As De Bruijn and Herder argue, in addition to a more techno-rational systems lens, one needs to take into account the effects of different intentions of actors involved in or affected by a system (De Bruijn and Herder 2009). As we saw in our vagueness analysis, actors may have different lived realities, languages or epistemic perspectives and, as a result, conflicting interests or incommensurable demands. Generally speaking, the actor perspective acknowledges and conceptualizes the dependencies between actors, sometimes captured in an "issue network" (Borzel 1998), and develops the "rules of the game" or governance mechanisms needed to satisfy all actors and manage the system adequately. While it is obvious that the system and actor perspective should at least happen in parallel and be in conversation, opinions vary on how integrated they should be, which in itself is a matter of normative indeterminacy. In this paper we argue that diagnosing and grappling with normative indeterminacy (or providing design space for parity) requires cybernetic feedback, which we specify both at the system (dynamical feedback) and at the actor level (feedback to renegotiate what abstractions and procedures are necessary to

safeguard a system). The iterative nature of dealing with emergent hard choices in AI system development, requires an iterative approach that also revisits the stakes and consensus reached among actors. As such, AI safety can only emerge through a *reciprocal relationship between system development and governance*.

Complex multi-actor problems are often called *wicked problems*, especially when they are subject to normative indeterminacy: "Wicked problems have incomplete, contradictory, and changing requirements, and solutions to them are often difficult to recognize as such because of complex dependencies" (De Bruijn 2009). Put differently, "they rely upon elusive political judgment for resolution. (Not "solution." Social problems are never solved. At best they are only re-solved--over and over again.) [..] The formulation of a wicked problem is the problem!" (Rittel 1973). At a minimum, a safety-critical context requires an honest account of the *problem and solution spaces*, which elaborate the different perspectives on the problem and its solution by various actors, as a basis for trying to reach broad consensus (De Haan 2015).

Acknowledge the connections between specification and political interests

Because the value hierarchy specified for and designed into a system will determine the space of actions available to it (as well as those that the system forecloses), it is crucial to acknowledge and account for the power and elevated status of design work (Irani 2016). This means recognizing developers' tendencies to prioritize certain actors and networks over others. Haraway (Haraway 1988), Harding (1986), and other critical scholars would argue that we cannot escape having some agenda: researchers are themselves situated in the social world they study. As such, technology development is inherently political and requires forms of accountability (Wagner 2020). Pioneers in participatory design argue that conflicts should be expected and that "[i]t's not the IT designer's job to cover up or try to solve political conflicts that surface [...] it is their job to develop different design visions and assess their consequences for the affected parties" (Bodker 2009). However, there are recent concerns that design methods using participation as a form of accountability are increasingly co-opted and stripped of their essence (Bodker 2018, Bannon 2018).

However, reducing political reflection to the role of the "developer" is too narrow to adequately capture the implications for specification. Just like other actors, developers are embedded in a network and subject to power differentials themselves. Understanding how broader hierarchies of power both promote and constrain certain problem formulations is necessary to determine viable strategies for promoting system safeguards. Today, much AI research and development, system implementation and management, as well as computational and software infrastructure is in the hands of a small number of technology companies. As Gürses and Van Hoboken argue, the move of tech companies to offer software engineering tools and data provision in service libraries and APIs has made the development of "values by design" an elusive task, and enabled new economic feedback loops that, when implemented at scale, drive new forms of inequality across social groups (Gürses and Van Hoboken 2017, Kostova 2020). We believe that real success in safeguarding high-stakes systems will require forms of oversight and dissent that support machine politics and respond to emergent safety hazards through citizen deliberation, especially for AI systems developed and deployed by the private sector and state actors.

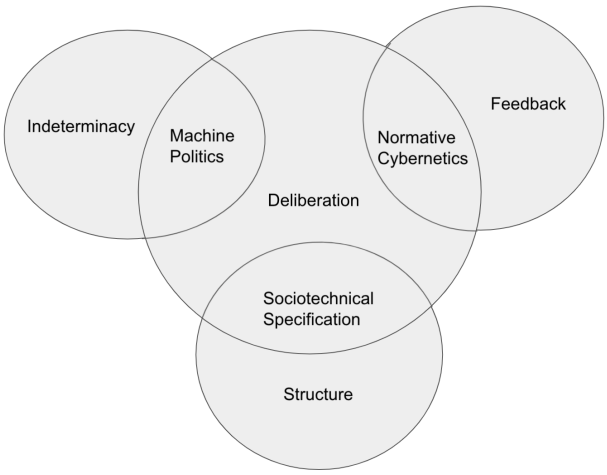
4.5 Conclusion

My framework is strongly influenced by the classic work of Philip Agre, which aimed to have AI practitioners and designers build better AI systems by requiring "a split identity - one foot planted in the craft work of design and the other foot planted in the reflexive work of critique". While I embrace the spirit of Agre's work, I also believe that the critical applications of today's AI systems require a new lens that can see beyond technical practices, and reframes the inherently interdisciplinary practice of AI development as critical in its own right. Apart from reflexivity, such a critical practice includes the forms of feedback that the domain of application asks for. The technical work done by AI practitioners plays a necessary but not sufficient part in development. It must be compensated by efforts to facilitate stakeholders' ability to be "full and active participants," while "the tools and techniques for doing this are dependent on the situations within the workplace...steer[ing] toward understanding different, pluralistic perspectives of how we think and act" (Greenbaum 1992). As such, I prioritize and label the centering of stakeholder safety concerns and hard choices to guide and inform AI development as *cybernetic practices*. I view this paper as a preliminary for what forms these practices might take in particular development domains, and will pursue this effort in future work.

My lodestar in this project is the intuition that clarifying the sociotechnical foundations of safety requirements will lay the groundwork for developers to take part in distinct dissent channels proactively, before the risks posed by AI systems become technically or politically insurmountable. In this way, incorporating dissent within development pipelines will provide opportunities for machine politics, i.e. deliberation about how citizens and public tribunals may want to reform the domain itself, rather than merely guarantee the AI system conforms to pre-existing normative criteria. I anticipate that cybernetic practices will need to be included within the training of engineers, data scientists, and designers as qualifications for the operation and management of advanced AI systems in the wild. Ultimately, the public itself must be educated about the assumptions, abilities, and limitations of these systems so that informed dissent will be made desirable and attainable as systems are being deployed. Deliberation is thus the goal of AI Safety, not just the procedure by which it is ensured. I endorse this approach due to the computationally underdetermined, semantically indeterminate, and politically obfuscated value hierarchies that will continue to define diverse social orders both now and in the future. Democratic dissent is necessary for such systems to safeguard the possibility of parity throughout their development and permit deliberation about the contours of our own values. To paraphrase Reinhold Niebuhr (Niebuhr 1986), AI's capacity for alignment makes machine politics possible, but its inclination to misalignment makes machine politics necessary.

5. CONCLUSION: THE (RE)BIRTH OF THE CLINIC?

It is one thing to justify and analytically present distinct modes of deliberation about how AI systems could be designed, built, and used to reorganize social domains. It is another to ask: how could these modes' maturation be facilitated, nurtured, and distinguished from each other in practice? In what kind of institutional space could designers, planners, commissioners, and citizens work together to reflect on and decide what AVs should be able to do and not do? While these questions are not tackled in the preceding chapters empirically or theoretically, I briefly address them here as a natural extension that outlines a direction I foresee for my work post-PhD.



The problem space of machine ethics.

At present, the spaces in which AI systems are intellectually theorized, mathematically modeled, technically developed, and publicly evaluated are either scattered or neglected. While activists have achieved limited success highlighting urgent risks present in deployed systems, we lack a systematic means of appraising future systems as well as the resources to imagine them differently. This status quo leaves designers without the disciplinary tools to change them, those familiar with the proper tools without access to the affected systems, activists without the support to take full advantage of either, and regulators without the criteria to oversee decisions.

Other fields like law and medicine have grappled with defining good outcomes in the context of domain expertise, and struggled to resolve them in institutional practice. Historically, this practice took the form of a clinic: a dedicated social space in which students learned to interface directly with the stakes and impacts of real-world problems. We might contrast this with a "best practices" approach that tries to make a system conform to unquestioned metrics, rather than deliberating to define and fine-tune metrics that make sense in context. An AI clinic would offer a shared institutional space where those who know the systems, the tools, and the domain stakes could come together. Through this new social space, the practitioners, the researchers, the representatives, the tools, and the systems would be given the resources to deliberate as needed.

Tech clinics would also serve a second goal of cross-pollinating agendas at the frontier of technical AI Safety and fairness research. At present, these subfields remain autonomous specializations for graduate students, despite their goal of building systems that reflect common human values like

autonomy, self-determination, and a shared sense of dignity. Moreover, their grasp of sociotechnical issues remains a work-in-progress, resting on problematic metaphors for agency and social order. By serving a translational role, the clinic could bridge the messiness of actual social domains with cutting-edge research, in the process professionalizing the next generation of practitioners into deliberating about actual (rather than toy) problems.

Different fields arrived at the "clinic" model for their own reasons. They learned distinct lessons about what kind(s) of social spaces are necessary for resolving certain kinds of problems in practice. They also had to translate between different forms of knowledge to diagnose these problems in the first place (Bonner 2000). Above all, the interface between technical tools and social problems cannot be assumed: it must be questioned, examined, and evaluated with both expert oversight and practical sense. But across fields, the role of the clinic is to: 1) make sense of and come to a decision about a particular social problem in practice; 2) translate between distinct knowledge representations and actual problems, in order to better articulate their stakes.

At present, what a dedicated clinic space for AI treatment might look like remains unclear. While there is a growing tech ethics curricula drawing on real-world datasets (Skirpan et al. 2018), most computer science graduate students gain practical experience through competitive internships at for-profit companies. This talent pipeline is nurtured by companies in pursuit of meritocratic solutions and short-term gains rather than extended reflection or societal investment. As a result, the requisite deliberation on domains or prospective types of feedback that design work would need to function well is hindered. AI clinics would need to interface between graduate students, non-profits, regulators, tech activists, and startups to specify the problem statement of a given AI system—its purpose, who it is for, and the risks entailed in its optimization. Through simultaneous support in a dedicated social space, these distinct positions could jointly confront misspecification problems and index technical choices within the wider context of development. This would include relevant business models, legal standards, and graduate certifications in fair, safe, or accountable system design.

What would it mean to include clinical engagement and some form of residency as an option for graduate students interested to specialize in applied machine learning or industry work after their PhD? How is the public interest involved? How might clinics change, subvert, or augment the mission of AI research as a whole? These are the questions raised by my dissertation research, and which I intend to investigate with collaborators in the coming years.

6. BIBLIOGRAPHY

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

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

Agre, P.E. Surveillance and capture: Two models of privacy, The Information Society 10 (2) (1994) 101–127.

Agre, P. Computation and human experience, Cambridge University Press, 1997.

Agre, P. Toward a critical technical practice: Lessons learned in trying to reform AI, Social Science, Technical Systems and Cooperative Work: Beyond the Great Divide. Erlbaum.

Al-Qadi, I. L.; Yang, R.; Kang, S.; Ozer, H.; Ferrebee, E.; Roesler, J. R.; Salinas, A.; Meijer, J.; Vavrik, W. R.; and Gillen, S. L. 2015. Scenarios developed for improved sustainability of Illinois Tollway: Life-cycle assessment approach. *Transportation Research Record* 2523(1): 11–18.

Alam, A.; Besselink, B.; Turri, V.; Martensson, J.; and Johansson, K. H. 2015. Heavy-duty vehicle platooning for sustainable freight transportation: A cooperative method to enhance safety and efficiency. *IEEE Control Systems Mag- azine* 35(6): 34–56.

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

Anand, S.; and Sen, A. 1994. Human Development Index: Methodology and Measurement.

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

Andersen, L. Human Rights in the Age of Artificial Intelligence, Tech. rep., Access Now (Nov. 2018).

Apte, J. S.; Messier, K. P.; Gani, S.; Brauer, M.; Kirchstetter, T. W.; Lunden, M. M.; Marshall, J. D.; Portier, C. J.; Vermeulen, R. C.; and Hamburg, S. P. 2017. High-resolution air pollution mapping with Google street view cars: exploiting big data. *Environmental science & technology* 51(12): 6999–7008.

Aristotle, *Politics*. Hackett Publishing, 1998.

Asbahan, R. E.; McCracken, J. K.; and Vandenbossche, J. M. 2008. Evaluating the Cracking Predicted by the MEPDG Using Results from the SR 22 Smart Pavement Study.

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

Azmat, M.; Kummer, S.; Moura Trigueiro, L.; Gennaro Di, F.; and Moser, R. 2018. Impact of innovative technologies on highway operators: Tolling organizations' perspective.

Banihashemi, M. 2011. Highway Safety Manual, new model parameters vs. calibration of crash prediction models. In *Moving Toward Zero. 2011 ITE Technical Conference and ExhibitInstitute of Transportation Engineers (ITE)*.

Bannon, L. and J. Bardzell, S. Bødker, Reimagining participatory design, Interactions 26 (1) (2018) 26–32. doi:10.1145/3292015.

Barnes, E. and J. R. G. Williams, A theory of metaphysical indeterminacy.

Barocas, Solon, and Andrew D. Selbst. "Big Data's Disparate Impact." *California Law Review* 104, no. 3 (2016): 671-732.

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

Beaulac, J.; Kristjansson, E.; and Cummins, S. 2009. Peer reviewed: A systematic review of food deserts, 1966-2007. *Preventing chronic disease* 6(3).

Bender, E.M. and T. Gebru, A. McMillan-Major, S. Shmitchell, On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21, Association for Computing Machinery, New York, NY, USA, 2021, pp. 610–623.

Benjamin, R. Race after technology: Abolitionist tools for the new jim code, John Wiley & Sons, 2019.

Benjamin, Ruha. "Race after Technology: Abolitionist Tools for the New Jim Code." *Social Forces* 98, no. 4 (2020): 1-3.

Bergenhem, C.; Shladover, S.; Coelingh, E.; Englund, C.; and Tsugawa, S. 2012. Overview of platooning systems. In *Proceedings of the 19th ITS World Congress, Oct 22-26, Vienna, Austria (2012)*.

Besselink, B.; Turri, V.; Van De Hoef, S. H.; Liang, K.- Y.; Alam, A.; Martensson, J.; and Johansson, K. H. 2016. Cyber–physical control of road freight transport. *Proceedings of the IEEE* 104(5): 1128–1141.

Binns, R. Fairness in machine learning: Lessons from political philosophy, in: Conference on Fairness, Accountability and Transparency, PMLR, 2018, pp. 149–159.

Bolukbasi, Tolga, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. "Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings." arXiv preprint arXiv:1607.06520 (2016)

Bonnefon, J.-F.; Shariff, A.; and Rahwan, I. 2016. The social dilemma of autonomous vehicles. *Science* 352(6293): 1573– 1576.

Bonner, Thomas Neville. *Becoming a physician: medical education in Britain, France, Germany, and the United States, 1750-1945*. JHU Press, 2000.

Borzel, T.A. Organizing Babylon-On the different conceptions of policy networks, Public administration 76 (2) (1998) 253–273, publisher: Wiley Online Library.

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

Bødker, S. and M. Kyng, Participatory Design that Matters—Facing the Big Issues, ACM Transactions on Computer-Human Interaction 25 (1) (2018) 4:1–4:31. doi:10.1145/3152421.

Brandao, M.; Jirotka, M.; Webb, H.; and Luff, P. 2020. Fair navigation planning: A resource for characterizing and designing fairness in mobile robots. *Artificial Intelligence* 282: 103259.

Brooks, Rodney. 2018. "Bothersome Bystanders and Self Driving Cars". Available online: https://rodneybrooks.com/bothersome-bystanders-and-self-driving-cars/

Cammack, Daniela. "Aristotle's denial of deliberation about ends." *Polis: The Journal for Ancient Greek and Roman Political Thought* 30.2 (2013): 228-250.

Campello-Vicente, H.; Peral-Orts, R.; Campillo-Davo, N.; and Velasco-Sanchez, E. 2017. The effect of electric vehicles on urban noise maps. *Applied Acoustics* 116: 59–64.

Carlini, N. and A. Athalye, N. Papernot, W. Brendel, J. Rauber, D. Tsipras, I. Goodfellow, A. Madry, A. Kurakin, On Evaluating Adversarial Robustness, arXiv:1902.06705 [cs, stat]ArXiv: 1902.06705.

Caubel, J. J.; Cados, T. E.; Preble, C. V.; and Kirchstetter, T. W. 2019. A distributed network of 100 black carbon sensors for 100 days of air quality monitoring in West Oakland, California. *Environmental science & technology* 53(13): 7564–7573.

Chang, K. S.; Karl Hedrick, J.; Zhang, W.-B.; Varaiya, P.; Tomizuka, M.; and Shladover, S. E. 1993. Automated highway system experiments in the path program. *Journal of Intelligent Transportation Systems* 1(1): 63–87.

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

Chang, R. The Possibility of Parity, Ethics 112 (4) (2002) 659–688.

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

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

Cihon, P. Standards for AI governance: international standards to enable global coordination in ai research & development, Future of Humanity Institute. University of Oxford.

Coase, Ronald Harry. "The Nature of the Firm." *Essential Readings in Economics*. London: Palgrave, 1995. 37-54.

Coase, Ronald H. "The Problem of Social Cost." *Classic Papers in Natural Resource Economics*. London: Palgrave Macmillan, 1960. 87-137.

Corbett-Davies, Sam, and Sharad Goel. "The measure and mismeasure of fairness: A critical review of fair machine learning." *arXiv preprint arXiv:1808.00023* (2018).

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

Crawford, K. and R. Dobbe, T. Dryer, G. Fried, B. Green, E. Kaziunas, A. Kak, V. Mathur, E. McElroy, A. N. S´anchez, AI Now Report 2019, Tech. rep., AI Now Institute, New York University, New York, NY, USA (Dec. 2019).

Crawford, Matthew B. *Why We Drive: Toward a Philosophy of the Open Road*. HarperCollins, 2020.

Dal Bó, Ernesto. "Regulatory Capture: A Review." *Oxford Review of Economic Policy* 22, no. 2 (2006): 203-225.

de Bruijn, H. and P. M. Herder, System and Actor Perspectives on Sociotechnical Systems, IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans 39 (5) (2009) 981–992, conference Name: IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans.

de Haan, A. and P. de Heer, Solving Complex Problems, 1, Auflage, 2015.

de Oña, J.; de Ona, R.; and Calvo, F.J. 2012. A classification tree approach to identify key factors of transit service quality. *Expert Systems with Applications* 39(12): 11164–11171.

DeCorla-Souza, P. 1994. Applying the cashing out approach to congestion pricing. *Transportation Research Record* 34– 34.

DeCorla-Souza, P. 2007. High-performance highways. *Public roads* 70(6).

Dean, Sarah, et al. "Axes for Sociotechnical Inquiry in AI Research." *IEEE Transactions on Technology and Society* (2021).

Dewey, John. The reflex arc concept in psychology., Psychological review 3 (4) (1896) 357.

Dewey, J. Public & its problems.

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

Dobbe, R. and M. Whittaker, AI and Climate Change: How they're connected, and what we can do about it, Tech. rep., AI Now Institute, New York City (Oct. 2019).

Dreyfus, Hubert L., and L. Hubert. *What computers still can't do: A critique of artificial reason*. MIT press, 1992.

Dreyfus, Hubert L. *Skillful coping: Essays on the phenomenology of everyday perception and action*. OUP Oxford, 2014.

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

Eckersley, P. Impossibility and uncertainty theorems in AI value alignment (or why your AGI should not have a utility function), arXiv preprint arXiv:1901.00064.

Elefteriadou, L.; et al. 2012. Expanded transportation performance measures to supplement level of service (los) for growth management and transportation impact analysis.

Erdélyi, O.J. and J. Goldsmith, Regulating artificial intelligence: Proposal for a global solution, in: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, 2018, pp. 95–101.

Eubanks, Virginia. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York: St. Martin's Press, 2018.

Ewing, R.; Schieber, R. A.; and Zegeer, C. V. 2003. Urban sprawl as a risk factor in motor vehicle occupant and pedestrian fatalities. *American journal of public health* 93(9): 1541–1545.

Fagnant, D. J.; and Kockelman, K. 2015. Preparing a nation for autonomous vehicles: opportunities, barriers and policy recommendations. *Transportation Research Part A: Policy and Practice* 77: 167–181.

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

Fisher, G.; Rolfe, K.; Kjellstrom, T.; Woodward, A.; Hales, S.; Sturman, A.; Kingham, S.; Petersen, J.; Shrestha, R.; and King, D. 2002. Health effects due to motor vehicle air pollution in New Zealand. *Wellington: Ministry of Transport*.

Flannery, A.; Anderson, A.; and Martin, A. 2004. Highway Capacity Manual and Highway Capacity software 2000 and advanced transportation modeling tools: Focus group findings. *Transportation research record* 1883(1): 176–184.

Fleetwood, J. 2017. Public health, ethics, and autonomous vehicles. *American journal of public health* 107(4): 532– 537.

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

Fligstein, Neil and Vogel, Steven. "Political Economy After Neoliberalism," available here: http://bostonreview.net/class-inequality/neil-fligstein-steven-vogel-political-economy-after-neoliberalism.

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

Fukuyama, Francis, Barak Richman, and Ashish Goel. "How to Save Democracy from Technology: Ending Big Tech's Information Monopoly." *Foreign Aff.* 100 (2021): 98.

Galster, G.; Hanson, R.; Ratcliffe, M. R.; Wolman, H.; Coleman, S.; and Freihage, J. 2001. Wrestling sprawl to the ground: defining and measuring an elusive concept. *Housing policy debate* 12(4): 681–717.

Garcia, P. and T. Sutherland, M. Cifor, A. S. Chan, L. Klein, C. D'Ignazio, N. Salehi, No: Critical refusal as feminist data practice, in: Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing, 2020, pp. 199–202.

Gasser, U. and C. Schmitt, The role of professional norms in the governance of artificial intelligence, The oxford handbook of ethics of AI (2020) 141Publisher: Oxford University Press.

Gatersleben, B.; and Uzzell, D. 2007. Affective appraisals of the daily commute: Comparing perceptions of drivers, cyclists, walkers, and users of public transport. *Environment and behavior* 39(3): 416–431.

Gebru, Timnit, et al. "Datasheets for datasets." *arXiv preprint arXiv:1803.09010* (2018).

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

Gillespie, T. The relevance of algorithms, Media technologies: Essays on communication, materiality, and society 167 (2014) (2014) 167.

Glanville, R. The purpose of second-order cybernetics, Kybernetes 33 (9/10) (2004) 1379–1386, publisher: Emerald Group Publishing Limited.

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

Green, Ben. *The smart enough city: putting technology in its place to reclaim our urban future*. MIT Press, 2019.

Green, B. and Y. Chen, The Principles and Limits of Algorithm-in-the-Loop Decision Making, Proceedings of the ACM on Human-Computer Interaction 3 (CSCW) (2019) 50:1–50:24.

Green, B. and S. Viljoen, Algorithmic realism: expanding the boundaries of algorithmic thought, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, FAT* '20, Association for Computing Machinery, New York, NY, USA, 2020, pp. 19–31.

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

Greenblatt, J. B.; and Saxena, S. 2015. Autonomous taxis could greatly reduce greenhouse-gas emissions of US light- duty vehicles. *Nature Climate Change* 5(9): 860–863.

Greenblatt, J. B.; and Shaheen, S. 2015. Automated vehicles, on-demand mobility, and environmental impacts. *Current sustainable/renewable energy reports* 2(3): 74–81.

Griffith, Shane, et al. "Policy Shaping: Integrating Human Feedback with Reinforcement Learning." *Advances in Neural Information Processing Systems* 26 (2013): 2625-2633.

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

Gurses, S. and J. v. Hoboken, Privacy after the Agile Turn, Tech. rep., SocArXiv, type: article (May 2017). doi:10.31235/osf.io/9gy73.

Hadfield-Menell, Dylan, et al. "Cooperative inverse reinforcement learning." *arXiv preprint arXiv:1606.03137* (2016).

Hadfield-Menell, D. and S. Milli, P. Abbeel, S. J. Russell, A. Dragan, Inverse reward design, in: Advances in Neural Information Processing Systems, 2017, pp. 6765–6774.

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

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

Hancock, P. A.; Nourbakhsh, I.; and Stewart, J. 2019. On the future of transportation in an era of automated and autonomous vehicles. *Proceedings of the National Academy of Sciences* 116(16): 7684–7691.

Haq, K.; and Ponzio, R. 2008. *Pioneering the human development revolution: an intellectual biography of Mahbub ul Haq*. Oxford University Press.

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

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

Head, L.; Shladover, S.; and Wilkey, A. 2012. Multi-Modal Intelligent Traffic Signal System. *University of Arizona* 32– 36.

Hedrick, J. K.; Tomizuka, M.; and Varaiya, P. 1994. Control issues in automated highway systems. *IEEE Control Systems Magazine* 14(6): 21–32.

Heidegger, M. and J. Macquarrie, E. Robinson, Being and time.

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

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

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

Hui, Yuk. "Cosmotechnics as Cosmopolitics." E-Flux 86, no. November (2017).

Hui, Yuk. 2019. *Recursivity and contingency*. Rowman & Littlefield International.

Hymel, K. 2019. If you build it, they will drive: Measuring induced demand for vehicle travel in urban areas. *Transport policy* 76: 57–66.

IEEE P7003 - Algorithmic Bias Working Group (2021).

Illing, Gerhard, and Ulrich Klüh, eds. *Spectrum Auctions and Competition in Telecommunications*. Cambridge, MA: MIT Press, 2003.

Ioannou, P. A.; and Stefanovic, M. 2005. Evaluation of ACC vehicles in mixed traffic: Lane change effects and sensitivity analysis. *IEEE Transactions on Intelligent Transportation Systems* 6(1): 79–89.

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

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

Irving, G. and A. Askell, AI safety needs social scientists, Distill 4 (2) (2019) e14.

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

Jiang, D.; Zhang, P.; Lv, Z.; and Song, H. 2016. Energy-efficient multi-constraint routing algorithm with load balancing for smart city applications. *IEEE Internet of Things Journal* 3(6): 1437–1447.

Kadir, K. Engineering justice? Rethinking engineering and positions as engineers to make the world a better place. (Feb. 2021).

Kalra, N.; and Paddock, S. M. 2016. Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability? *Transportation Research Part A: Policy and Practice* 94: 182–193.

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

Klein, Jacob. *Greek mathematical thought and the origin of algebra*. MIT Press, 1968.

Klonick, K. The facebook oversight board: Creating an independent institution to adjudicate online free expression, Yale LJ 129 (2019) 2418.

Knibbs, L. D.; Cole-Hunter, T.; and Morawska, L. 2011. A review of commuter exposure to ultrafine particles and its health effects. *Atmospheric Environment* 45(16): 2611– 2622.

Konrardy, B.; Christensen, S. T.; Hayward, G.; and Farris, S. 2018. Autonomous vehicle routing during emergencies. US Patent 10,156,848.

Koopman, P.; and Wagner, M. 2017. Autonomous vehicle safety: An interdisciplinary challenge. *IEEE Intelligent Transportation Systems Magazine* 9(1): 90–96.

Kostova, B. and S. Gurses, C. Troncoso, Privacy Engineering Meets Software Engineering. On the Challenges of Engineering Privacy ByDesign, arXiv preprint arXiv:2007.08613.

Krais, B. Gender and symbolic violence: Female oppression in the light of pierre bourdieu's theory of 985 social practice, Bourdieu: critical perspectives (1993) 156–177.

Krzyzanowski, M.; Kuna-Dibbert, B.; and Schneider, J. 2005. *Health effects of transport-related air pollution*. WHO Regional Office Europe.

Lee, S.; Kim, Y.; Kahng, H.; Lee, S.-K.; Chung, S.; Cheong, T.; Shin, K.; Park, J.; and Kim, S. B. 2020. Intelligent traffic control for autonomous vehicle systems based on machine learning. *Expert Systems with Applications* 144: 113074.

Lee Jr, D. B.; Klein, L. A.; and Camus, G. 1999. Induced traffic and induced demand. *Transportation Research Record* 1659(1): 68–75.

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

Leveson, N.G. and J. Moses, Engineering a Safer World: Systems Thinking Applied to Safety, MIT Press, Cambridge, UNITED STATES, 2012.

Levin, M. W. 2017. Congestion-aware system optimal route choice for shared autonomous vehicles. *Transportation Research Part C: Emerging Technologies* 82: 229–247.

Levy, K. E. 2015. The contexts of control: Information, power, and truck-driving work. *The Information Society* 31(2): 160–174.

Li, T.M. The Will to Improve: Governmentality, Development, and the Practice of Politics, Duke University Press, 2007.

Lin, S.-H.; and Ho, T.-Y. 2019. Autonomous vehicle routing in multiple intersections. In *Proceedings of the 24th Asia and South Pacific Design Automation Conference*, 585–590.

Lipfert, F.; Wyzga, R.; Baty, J.; and Miller, J. 2006. Traffic density as a surrogate measure of environmental exposures in studies of air pollution health effects: long-term mortality in a cohort of US veterans. *Atmospheric Environment* 40(1): 154–169.

Litman, T. 2017. *Autonomous vehicle implementation predictions*. Victoria Transport Policy Institute Victoria, Canada.

Lyman, K.; and Bertini, R. L. 2008. Using travel time reliability measures to improve regional transportation planning and operations. *Transportation Research Record* 2046(1): 1–10.

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

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

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

Manual, H. C. 2000. Highway capacity manual. *Washington, DC* 2: 1.

McAllister, R.; Gal, Y.; Kendall, A.; Van Der Wilk, M.; Shah, A.; Cipolla, R.; and Weller, A. 2017. Concrete problems for autonomous vehicle safety: Advantages of bayesian deep learning. International Joint Conferences on Artificial Intelligence, Inc.

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

Messier, K. P.; Chambliss, S. E.; Gani, S.; Alvarez, R.; Brauer, M.; Choi, J. J.; Hamburg, S. P.; Kerckhoffs, J.; LaFranchi, B.; Lunden, M. M.; et al. 2018. Mapping air pollution with google street view cars: Efficient Approaches with mobile monitoring and land use regression. *Environ- mental science & technology* 52(21): 12563–12572.

Metz, D. 2018. Developing policy for urban autonomous vehicles: Impact on congestion. *Urban Science* 2(2): 33.

Millard-Ball, A. 2018. Pedestrians, autonomous vehicles, and cities. *Journal of planning education and research* 38(1): 6–12.

Miller, J.; and How, J. P. 2017. Predictive positioning and quality of service ridesharing for campus mobility on demand systems. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, 1402–1408. IEEE.

Miller, S. A.; and Heard, B. R. 2016. The environmental impact of autonomous vehicles depends on adoption patterns.

Milli, S. and D. Hadfield-Menell, A. Dragan, S. Russell, Should robots be obedient?, arXiv preprint arXiv:1705.09990.

Mittelstadt, B. Principles alone cannot guarantee ethical AI, Nature Machine Intelligence 1 (11) (2019) 501–507, number: 11 Publisher: Nature Publishing Group.

Mishra, S.; Welch, T. F.; and Jha, M. K. 2012. Performance indicators for public transit connectivity in multi-modal transportation networks. *Transportation Research Part A: Policy and Practice* 46(7): 1066–1085.

Mitchell, M. and S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, T. Gebru, Model Cards for Model Reporting, Proceedings of the Conference on Fairness, Accountability, and 880 Transparency (2019) 220–229.

Montello, D. R.; and Sas, C. 2006. Human factors of wayfinding in navigation.

Montgomery, W. D.; Mudge, R.; Groshen, E. L.; Helper, S.; MacDuffie, J. P.; and Carson, C. 2018. America's workforce and the self-driving future: Realizing productivity gains and spurring economic growth.

Morrow, W. R.; Greenblatt, J. B.; Sturges, A.; Saxena, S.; Gopal, A.; Millstein, D.; Shah, N.; and Gilmore, E. A. 2014. Key factors influencing autonomous vehicles' energy and environmental outcome. In *Road vehicle automation*, 127–135. Springer.

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

Narayanan, A. FAT* tutorial: 21 fairness definitions and their politics (Feb. 2018).

Ng, Andrew Y., and Stuart J. Russell. "Algorithms for Inverse Reinforcement Learning." *Proceedings of the Seventeenth International Conference on Machine Learning*. Vol. 1. 2000.

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

Nissenbaum, Helen. *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford, CA: Stanford University Press, 2009.

Noble, Safiya Umoja. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: New York University Press, 2018.

Noland, R. B.; and Quddus, M. A. 2006. Flow improve- ments and vehicle emissions: effects of trip generation and emission control technology. *Transportation Research Part D: Transport and Environment* 11(1): 1–14.

Noë, Alva. 2015. *Strange tools: Art and human nature*. Hill and Wang.

Norton, P. D. 2011. *Fighting traffic: the dawn of the motor age in the American city*. Mit Press.

O'Donnell, B.; Corey, E.; and Podowski, M. 2017. 2017 DRIVE CLEAN SEATTLE Implementation Strategy.

Ohnemus, M.; and Perl, A. 2016. Shared autonomous vehi- cles: Catalyst of new mobility for the last mile? *Built Environment* 42(4): 589–602.

Omstedt, G.; Bringfelt, B.; and Johansson, C. 2005. A model for vehicle-induced non-tailpipe emissions of particles along Swedish roads. *Atmospheric Environment* 39(33): 6088– 6097.

O'Neil, Cathy. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Broadway Books, 2016.

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

Part, D. 2010. Highway safety manual.

Pask, Gordon. Conversation theory, Applications in Education and Epistemology.

Pasquale, Frank. *The Black Box Society*. Cambridge, MA: Harvard University Press, 2015.

Poole, R. 2020. The Impact of HOV and HOT Lanes on Congestion in the United States.

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

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

Queiroz, C.; Vajdic, N.; and Mladenovic, G. 2013. Public–private partnerships in roads and government support: trends in transition and developing economies. *Transportation Planning and Technology* 36(3): 231–243.

Raji, I.D. and R. Dobbe, Concrete Problems in AI Safety, Revisited, in: Workshop on Machine Learning In Real Life, Addis Abeba, Ethiopia, 2020.

Reddy, S. G. 2003. The 'Index Number Problem' and Poverty Monitoring: The Unique Advantages of a Capability Based Approach.

Rittel, H.W. and M. M. Webber, Dilemmas in a general theory of planning, Policy sciences 4 (2) (1973) 155–169, publisher: Springer.

Rossi, F.; Zhang, R.; Hindy, Y.; and Pavone, M. 2018. Routing autonomous vehicles in congested transportation networks: Structural properties and coordination algorithms. *Autonomous Robots* 42(7): 1427–1442.

Russell, Stuart. Provably beneficial artificial intelligence, Exponential Life, The Next Step.

Russell, Stuart. *Human compatible: Artificial intelligence and the problem of control*. Penguin, 2019.

Russell, Stuart, and Peter Norvig. 2002. *Artificial intelligence: a modern approach*. Pearson.

Schelling, T. C. 1969. Models of segregation. *The American Economic Review* 59(2): 488–493.

Schiff, D. and J. Borenstein, J. Biddle, K. Laas, AI Ethics in the Public, Private, and NGO Sectors: A Review of a Global Document Collection, IEEE Transactions on Technology and Society 2 (1) (2021) 31–42, conference Name: IEEE Transactions on Technology and Society.

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

Seaver, N. and J. Vertesi, D. Ribes, Knowing algorithms, in: digitalSTS, Princeton University Press, 2019, pp. 412–422.

Selbst, A.D. and D. Boyd, S. A. Friedler, S. Venkatasubramanian, J. Vertesi, Fairness and Abstraction in Sociotechnical Systems, in: Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* '19, Association for Computing Machinery, New York, NY, USA, 2019, pp. 59–68.

Sen, A. 1983. Development: Which way now? *The economic journal* 93(372): 745–762.

Sen, A. 1988. The concept of development. *Handbook of development economics* 1: 9–26.

Sendak, Mark, et al. "'The Human Body Is a Black Box': Supporting Clinical Decision- Making With Deep Learning." *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 2020.

Sgouridis, S.; Bonnefoy, P. A.; and Hansman, R. J. 2011. Air transportation in a carbon constrained world: Long-term dynamics of policies and strategies for mitigating the carbon footprint of commercial aviation. *Transportation Research Part A: Policy and Practice* 45(10): 1077–1091.

Shaheen, S.; and Chan, N. 2016. Mobility and the sharing economy: Potential to facilitate the first- and last-mile public transit connections. *Built Environment* 42(4): 573–588.

Shen, Q. 1998. Location characteristics of inner-city neighborhoods and employment accessibility of low-wage workers. *Environment and planning B: Planning and Design* 25(3): 345–365.

Shilton, K. Values and Ethics in Human-Computer Interaction, Foundations and TrendsR Human–Computer Interaction 12 (2) (2018) 107–171.

Shladover, S. E.; Su, D.; and Lu, X.-Y. 2012. Impacts of cooperative adaptive cruise control on freeway traffic flow. *Transportation Research Record* 2324(1): 63–70.

Silver, D., Singh, S., Precup, D. and Sutton, R.S., 2021. Reward is enough. *Artificial Intelligence*, p.103535.

Simoni, M. D.; Kockelman, K. M.; Gurumurthy, K. M.; and Bischoff, J. 2019. Congestion pricing in a world of self-driving vehicles: An analysis of different strategies in alternative future scenarios. *Transportation Research Part C: Emerging Technologies* 98: 167–185.

Skirpan, M. and 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, SIGCSE '18, (New York, NY, USA), p. 940–945, Association for Computing Machinery, 2018.

Smuha, N.A. From a 'race to AI'to a 'race to AI regulation': regulatory competition for artificial 925 intelligence, Law, Innovation and Technology 13 (1) (2021) 57–84.

Snaith, M.; and Burrow, J. 1984. Priority assessment. *Transportation Research Record* 951: 87–95.

Snchez, O. A. 2000. The legacy of human development: a tribute to Mahbub ul Haq.

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

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

Sousa, N.; Almeida, A.; Coutinho-Rodrigues, J.; and Natividade-Jesus, E. 2018. Dawn of autonomous vehicles: review and challenges ahead. In *Proceedings of the Insti- tution of Civil Engineers-Municipal Engineer*, volume 171, 3–14. Thomas Telford Ltd.

Stark, L. Facial recognition is the plutonium of AI, XRDS: Crossroads, The ACM Magazine for Students 25 (3) (2019) 50–55, publisher: ACM New York, NY, USA.

Steinbaum, Marshall, and Maurice E. Stucke. "The Effective Competition Standard: A New Standard for Antitrust." *University of Chicago Law Review,* forthcoming (2018).

Stiegler, Bernard. *Technics and time: The fault of Epimetheus*. Vol. 1. Stanford University Press, 1998.

Stocker, A.; and Shaheen, S. 2018. Shared automated mobility: early exploration and potential impacts. *Road Vehicle Automation 4* 125–139.

Sullivan,H.T.;andHa¨kkinen,M.T.2011. Preparedness and warning systems for populations with special needs: Ensuring everyone gets the message (and knows what to do). *Geotechnical and Geological Engineering* 29(3): 225–236.

Theberge, P. E. 1987. Development of mathematical models to assess highway maintenance needs and establish rehabilitation threshold levels. *Transportation Research Record* 1109: 27–35.

Tiba, K.; Parizi, R. M.; Zhang, Q.; Dehghantanha, A.; Karimipour, H.; and Choo, K.-K. R. 2020. Secure blockchain-based traffic load balancing using edge computing and reinforcement learning. In *Blockchain Cybersecurity, Trust and Privacy*, 99–128. Springer.

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

Ul Haq, M. 1995. *Reflections on human development*. Oxford University Press.

Unger, R.M. The critical legal studies movement, Harvard Law Review (1983) 561–675.

Van den Berg, V. A.; and Verhoef, E. T. 2016. Autonomous cars and dynamic bottleneck congestion: The effects on capacity, value of time and preference heterogeneity. *Transportation Research Part B: Methodological* 94: 43–60.

van der Voort, H.G. and A. J. Klievink, M. Arnaboldi, A. J. Meijer, Rationality and politics of algorithms. 940 Will the promise of big data survive the dynamics of public decision making?, Government Information Quarterly 36 (1) (2019) 27–38.

Vinitsky, E.; Kreidieh, A.; Le Flem, L.; Kheterpal, N.; Jang, K.; Wu, C.; Wu, F.; Liaw, R.; Liang, E.; and Bayen, A. M. 2018. Benchmarks for reinforcement learning in mixed-autonomy traffic. In *Conference on Robot Learning*, 399– 409.

Vinitsky, Eugene, et al. "Lagrangian control through deep-rl: Applications to bottleneck decongestion." *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2018.

Voigt, P. and A. Von dem Bussche, The EU general data protection regulation (gdpr), A Practical Guide, 1st Ed., Cham: Springer International Publishing 10 (2017) 3152676.

Von Foerster, H. Understanding understanding: Essays on cybernetics and cognition, Springer Science & Business Media, 2007.

von Krogh, G. Artificial Intelligence in Organizations: New Opportunities for Phenomenon-Based Theorizing, Academy of Management Discoveries 4 (4) (2018) 404–409, publisher: Academy of Management. doi:10.5465/amd.2018.0084.

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

Wagner, B. Accountability by design in technology research, Computer Law & Security Review 37 (2020) 105398.

Walker, Jarrett. *Human transit: How clearer thinking about public transit can enrich our communities and our lives*. Island Press, 2012.

Wallach, W. and G. Marchant, Toward the agile and comprehensive international governance of AI and 915 robotics [point of view], Proceedings of the IEEE 107 (3) (2019) 505–508.

Wang, P.; Motamedi, S.; Canas Bajo, T.; Zhou, X.; Zhang, T.; Whitney, D.; and Chan, C.-Y. 2019. Safety Implications of Automated Vehicles Providing External Communication to Pedestrians.

West, S. E. 2004. Distributional effects of alternative vehicle pollution control policies. *Journal of public Economics* 88(3-4): 735–757.

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

Whittaker, Meredith, et al. *AI now report 2018*. New York: AI Now Institute at New York University, 2018.

Wiener, N. The human use of human beings: Cybernetics and society, no. 320, Da Capo Press, 1988.

Wiener, N., 2019. *Cybernetics or Control and Communication in the Animal and the Machine*. MIT press.

Wittgenstein, L. Philosophical Investigations, Basil Blackwell, Oxford, 1953.

Williamson, T. Vagueness, Routledge, 2002.

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

Winner, Langdon. *The whale and the reactor: A search for limits in an age of high technology*. University of Chicago Press, 2010.

Wong, S.; Shaheen, S.; and Walker, J. 2018. Understanding evacuee behavior: a case study of hurricane Irma.

Wu, C.; Kreidieh, A.; Parvate, K.; Vinitsky, E.; and Bayen, A. M. 2017a. Flow: Architecture and benchmarking for reinforcement learning in traffic control. *arXiv preprint arXiv:1710.05465* 10.

Wu, C.; Kreidieh, A.; Parvate, K.; Vinitsky, E.; and Bayen, A. M. 2017b. Flow: Architecture and Benchmarking for Reinforcement Learning in Traffic Control. *CoRR* abs/1710.05465. URL http://arxiv.org/abs/1710.05465.

Wu, C.; Kreidieh, A.; Vinitsky, E.; and Bayen, A. M. 2017c. Emergent behaviors in mixed-autonomy traffic. In *Conference on Robot Learning*, 398–407.

Yeung, K. 'hypernudge': Big data as a mode of regulation by design, Information, Communication & Society 20 (1) (2017) 118–136.

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

Zhang, K.; and Batterman, S. 2013. Air pollution and health risks due to vehicle traffic. *Science of the total Environment* 450: 307–316.

Zuboff, Shoshana. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. London: Profile Books, 2019.

Zwetsloot, Remco, and Allan Dafoe. "Thinking about risks from AI: accidents, misuse and structure." *Lawfare. February* 11 (2019): 2019.

Zwetsloot, R. and H. Toner, J. Ding, Beyond the AI arms race: America, china, and the dangers of zero-sum thinking, Foreign Affairs 16.