Fairness and Sequential Decision Making: Limits, Lessons, and Opportunities

Anonymized for Submission

ABSTRACT

As automated decision making and decision assistance systems become common in everyday life, research on the prevention or mitigation of potential harms that arise from decisions made by these systems has proliferated. However, various research communities have independently conceptualized these harms and proposed interventions. The result is a somewhat fractured landscape of literature focused generally on ensuring decision-making algorithms "do the right thing." In this paper, we compare and discuss work across two major subsets of this literature: algorithmic fairness, which focuses primarily on predictive systems, and ethical decision making, which focuses primarily on sequential decision making and planning. We explore how each of these settings has articulated its normative concerns, the viability of different techniques for these different settings, and how ideas from each setting may have utility for the other.

ACM Reference Format:

Anonymized for Submission. 2023. Fairness and Sequential Decision Making: Limits, Lessons, and Opportunities. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The social and ethical implications of different technologies have long been the object of study for scholars outside of computer science, and recently many computer scientists have taken up this broader agenda under a variety of names. In particular, two largely independent communities have evolved from established fields of computer science. The study of algorithmic fairness that has emerged at the FAccT conference and its predecessors is heavily influenced by the field of machine learning and focuses on predictive systems, while the study of ethical decision making[1] has attracted primarily researchers from classical artificial intelligence and focuses on sequential decision making. Nominally, these groups have similar goals: to produce predictive or decision-making systems that "do the right thing." However, many key ideas from ethical decision-making have not yet percolated into the fairness literature, and many important concepts from fair prediction are not yet common in ethical decision making. This paper is an effort to bridge this gap.

Unlike predictive systems, which consider decisions independently and one at a time (known as myopic decision making), sequential decision-making systems consider sequences of potential actions, allowing them to evaluate the long-term effects of taking a particular set of actions before they are made. Many real-world problems, such as autonomous driving, power grid management, wildfire fighting, military engagement, disaster relief, and inventory logistics, both fundamentally affect people's safety and access to resources and require sequential reasoning as they cannot be solved adequately via myopic decision making. However, although problems such as autonomous driving sometimes motivate the fairness literature [87, 88, 108, 140, 166, 233, 240], fairness conceptualizations and methods have largely been developed for predictive rather than sequential decision-making systems. Moreover, despite the fairness literature's acknowledgement of the long-term effects and sequential nature of many high-stakes decisions [41, 62, 74, 81, 82, 107, 110, 119, 162, 172, 177, 194, 212, 213, 221], including education and college admissions [7, 113, 182], recidivism risk prediction [67, 167], predictive policing [54], child and homeless welfare [75, 208], clinical trials [61], and hiring [37, 163], work on these settings rarely engages problem formulations or approaches developed for sequential decision making, or attempts to conceptualize and address ethical concerns beyond fairness, such as those emerging from the ethical decision making literature.

Our paper makes the following contributions. We begin by introducing a foundational and widely-used sequential decision-making model, the Markov decision process (MDP), from which many special-case models are derived. We cover problem formulation, solution methods, and key assumptions and properties (§2). We then examine how ethical concerns have been conceptualized within the ethical decision-making and fairness literatures (§3), examine the sequential decision-making model pipeline (§4), introduce some of the measurements (§5) and mitigations (§6) common in the ethical decision-making literature, and discuss some current challenges and state-of-the-art techniques for ethical decision making.

Throughout, we offer observations following three general themes. First, we draw comparisons between conceptualizations, measurements, and mitigations proposed in the fairness and ethical decision making literatures to highlight where insights and methods from fairness may or may not be appropriate for ethical decision making, and vice versa. We draw two conclusions. 1) some techniques and methods do not or will not work and are not transferable for fundamental reasons; 2) other concepts have potential for adoption, and fairness researchers would benefit from considering a broader array of solutions, including some sequential decision-making techniques. Second, inspired by the fairness literature's analyses of machine learning pipelines, we draw attention to aspects of sequential decision-making pipelines that represent opportunities for future analysis. We argue that fairness researchers may be uniquely positioned to study sequential decision-making systems in their native

[1] The term "ethical decision making" has (unsurprisingly) been used to describe a variety of research, including symbolic planning and system verification. Here, we use it to refer to work on ethical concerns arising from sequential decision making systems.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM. . . $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

deployments and the accompanying sociotechnical nuance and contribute to ethical decision making research. Finally, we highlight some problem formulations and techniques developed for ethical decision making that may offer advantages for fairness research.

2 BACKGROUND ON SEQUENTIAL DECISION MAKING

A **Markov decision process (MDP)** is a general sequential decision-making model[2] that enables an agent[3] to make a sequence of decisions in fully observable, stochastic environments [21] and has been used in many decision-making problems, such as search and rescue [99, 201], extraterrestrial exploration [91, 183], and autonomous driving [20, 262, 263]. An MDP describes a decision-making problem using four attributes: (1) a set of **states** that represent different possible scenarios, (2) a set of **actions** that can be performed by the agent, (3) a **transition function** that gives the probability of reaching a given state when the agent performs a particular action in its current state, and (4) a **reward function** that gives the immediate utility of performing a particular action in its current state. At each time step the agent performs an action in a state, receives a reward based on the reward function, and transitions to a successor state based on the transition function. MDPs satisfy a key property, called the **Markov property**, that holds that the outcome of any action only depends on the current state. That is, the agent's prior states and actions do not matter. The solution to an MDP is the **optimal policy**, the mapping from states actions that maximizes the value function. The **value function** is defined over all states and represents the expected cumulative reward the agent would earn if it executed the optimal policy from each state.

Formal Definition: An MDP is a tuple, $\langle S, A, T, R\rangle$, where: S is a finite set of states; A is a finite set of actions; $T(s, a, s')$ is a transition function that represents the probability of reaching state s' after performing action a in state s; and $R(s, a)$ is a reward function that represents the immediate reward gained by performing action a in state s. At each time step, the agent performs an action a in a state s, experiences reward $R(s, a)$, and transitions to a successor state s' with probability $T(s, a, s')$. The agent either repeats these steps forever (infinite horizon) or until a deadline (finite horizon).

A solution to an MDP is a policy $\pi : S \rightarrow A$, where $\pi(s) = a$ indicates that the agent should perform action a when in state s. For a given policy $\pi(s)$, its value function $V^\pi(s)$ describes the value of each state s with respect to the policy $\pi(s)$. In particular, the value function $V^\pi(s)$ describes the expected cumulative reward that the agent would earn starting in state s and executing policy $\pi(s)$, until reaching the horizon:

$$V^\pi(s) = R(s, \pi(s)) + \gamma \sum_{s' \in S} T(s, \pi(s), s') V(s').$$

Typically, the expected cumulative reward is discounted to balance the value of immediate rewards with the value of future rewards: that is, the discount factor is often $\gamma \in [0, 1)$ in infinite horizon MDPs and $\gamma = 1$ in finite horizon MDPs. Along with balancing rewards gained in the present and rewards gained in the future, a discount factor $\gamma < 1$ provides guarantees that the value function of an infinite horizon MDP converges to finite values. The goal of the agent is to find the optimal policy $\pi^*(s)$ that maximizes the value—the expected cumulative reward—of each state s until reaching the horizon:

$$V^*(s) = \max_{a \in A} \left[R(s, a) + \gamma \sum_{s' \in S} T(s, a, s') V^*(s') \right].$$

Finally, given the optimal value function $V^*(s)$, the optimal policy $\pi^*(s)$ can be calculated in the following way:

$$\pi^*(s) = \arg\max_{a \in A} V^*(s).$$

There are two main approaches to solving MDPs, depending on whether or not the reward function and transition function of the MDP are known. In problems in which both functions are available, an agent can use *planning* methods to directly calculate an effective policy by computing the optimal value of each state and then the optimal action [21]. More specifically, these methods typically involve calculating the optimal value function and then the optimal policy by using dynamic programming or linear programming. However, in problems in which either or both of these functions are unavailable, an agent can use *reinforcement learning* methods to gradually learn an effective policy by performing actions and observing rewards to estimate the optimal value of each state and then the optimal action [235]. That is, all reinforcement learning is built upon MDPs and their variants. In particular, these methods usually involve estimating the optimal value function and then the optimal policy by interleaving greedy actions with exploratory actions.[4]

Example: Consider a power plant that supplies power to several neighborhoods. The goal of the power plant is to balance three potentially competing objectives: it must (1) supply power to each neighborhood as cheaply as possible, (2) avoid outages, and (3) reduce excess power that is stored in its battery network and dissipates gradually. Armed with the MDP framework, we can formally represent the decision-making problem of the power plant as an MDP $\langle S, A, T, R\rangle$.

In particular, suppose the plant can supply a maximum of R kilowatts (kW) to a set of neighborhoods N where each neighborhood N_i demands D_i kW. The plant incurs a cost of $C \geq 0$ per kW generated and charges each neighborhood a price $P \geq C$ per kW. It also incurs a cost $E \propto R - D$ for generating excess power. We assume the power plant either meets all or none of the power demand D_i kW for a given neighborhood N_i: that is, the plant supplies either D_i or 0 kW to neighborhood N_i. Thus, our set of states $S = E \times P \times D_1 \times \cdots \times D_{|N|} \times F_1 \cdots \times F_{|N|}$ where $P = \{\textsc{Low}, \textsc{Normal}, \textsc{High}\}$ is the current price of power, D_i is the current power demand for the neighborhood N_i, and $F_i = \{\textsc{Fulfilled}, \textsc{Unfulfilled}\}$ is the current fulfillment status of the neighborhood N_i, reflecting whether or not the current power demand D_i is met.

[2] MDPs and their variants occupy the vast majority of the AI and planning literature that uses the term "sequential decision making".

[3] We use the terms "agent" (the preferred term in classical AI research) and "system" (a more general, catchall term) interchangeably to describe collections of processes which can take actions in the world, such as a robot. We use the term "model" to describe a decision-making or predictive model specifically, removed from the larger system in which it operates.

[4] Although we do not discuss many solution methods in this work, many, such as value iteration [21], RTDP [19], Monte Carlo tree search [44], Q-learning [256], and SARSA [56] have proven to be effective across a variety of applications, including Atari [179], chess [226], and StarCraft [250].

The plant has two ways to control load: it can increase or decrease the price P in order to keep total demand $D = D_1 + D_2 + \cdots + D_{|N|}$ close to, but below, the maximum rate R. However, if $D > R$, the power plant can also terminate supply to neighborhoods $\bar{N} \subset N$. The set of actions is thus $A = \{\oplus, \ominus\} \times \mathcal{P}(N)$, where $\oplus$ and $\ominus$ increase and decrease the current price of power P and the powerset $\mathcal{P}(N)$ is every combination of neighborhoods for which power can be shut down.

The transition function $T(s, a, s')$ represents how the probability of the power demand of each neighborhood varies with the current price of power. The reward function $R(s, a)$ represents the relative cost of service interruptions, charging a price of power higher than the cost of power generation, and having to store excess power in a battery network. A non-myopic model like an MDP is obviously preferable to a classifier in this decision-making scenario since the outcome of a given action has both some uncertainty as well as some impact on possible subsequent actions.

Frontiers of Sequential Decision Making: A substantial body of work focuses on solving MDPs efficiently given that the computational complexity of solving them "blows up" with the size of their states and actions. This problem is colloquially referred to as the *curse of dimensionality* in AI literature. To provide some background, we highlight three common approaches to solving MDPs approximately. *Approximate programs* estimate the optimal value function and then calculate the optimal policy for that estimated optimal value function by using approximate forms of dynamic programming [27, 204] or linear programming [103, 171, 200, 203]. *Replanning methods* generate a policy for a subset of states (called a partial policy) and then generate a new partial policy whenever a state is encountered for which the partial policy is undefined [202, 230, 267], enabling the solver to reason only about the most likely states. Finally, *abstraction methods* build an abstracted MDP to reduce the size of its state and action spaces and then solve for the optimal policy of the abstracted MDP [34, 70, 85, 96, 158, 188, 189, 211, 216, 268], retaining relevant details and condensing those less important. In practice, approximate MDP solvers may employ various combinations of these approaches.

In addition to work on solving MDPs efficiently, there are many MDP extensions that can represent different classes of decision-making problems. Here we focus on MDPs, a model for decision-making problems in which the current state can be directly observed. However, there are many decision-making models with different forms of expressiveness for decision-making problems with different properties. For example, for problems in which the current state is not directly observed by the agent, requiring the agent to manage a belief over the current state, we can use a partially observable MDP (POMDP) [130]. For problems in which the agent must find the shortest path from a start state to a goal state, we can use a stochastic shortest path problem (SSP) [145]. For problems in which multiple, decentralized agents must coordinate, we can use a decentralized MDP (DecMDP) [25]. There are many other MDP flavors, but they also suffer from the curse of dimensionality, often so much so that they require specialized approaches to solve efficiently.

Fairness in Sequential Decision-Making Systems: Recently, there have been arguments for using MDPs to model decisions traditionally handled by supervised learning [113, 272]. However, there are relatively few efforts at producing fairness definitions consistent with the definition of an MDP, such as Wen et al. [258], who use constrained MDPs to express fairness constraints for a subclass of MDPs with separable reward and transition functions. Here, expected reward (value) plays an analogous role to the loss function in supervised learning. Some surveys also highlight the temporal nature of the many decisions AI systems make, but focus primarily on allocative tasks, stopping short of expanding these problems to encompass the types of sequential decision-making models most often deployed by embodied AI systems [274].

One class of MDPs that is relatively well-studied, however, is the multi-armed bandit problem [40]. In this problem there is a set of arms (actions), each yielding a different reward according to an unknown distribution. The objective is to determine the arm to pull that maximizes expected cumulative reward. Formally, the multi-armed bandit problem is a class of MDPs in which the agent performs a single action instead of a sequence of actions. Here, recent work has offered models and algorithms for introducing different notions of fairness [175]. Joseph et al. [126, 127, 128] initiated this line of research by introducing a meritocratic definition of fairness that ensures that a better arm is always favored over a worse arm despite uncertainty over each arm's expected reward. Then, extending this work to each arm's reward distribution instead of its expected reward, Liu et al. [164] offered a method for ensuring that two arms are pulled roughly the same number of times if they satisfy a notion of similarity based on these reward distributions. Moreover, in the context of fairness constraints, there has been a range of methods for ensuring that each arm is selected a minimum number of times [57, 58, 64, 157, 197]. Finally, as a way to reason about group fairness, Schumann et al. [220] offered a method for partitioning the arms into different sensitive groups based on protected features, such as race, age, and socio-economic status, that are in turn picked from according to a given definition of fairness. However, while these works examine fairness in the context of multi-armed bandits and have led to encouraging results, it remains challenging to extend these ideas to MDPs because MDPs are a strict generalization of multi-armed bandits in which the agent must optimize over a sequence of actions instead of a single action, and each action affects state transitions.

3 CONCEPTUALIZATIONS

In this section, we briefly examine how the fairness and ethical decision-making literatures have conceptualized their work. Fairness is an essentially contested construct [120, 123], and fairness in predictive systems, though generally centered on unequal exposure of certain people to potential system failures, has been conceptualized in a variety of ways, among them individual and group fairness. Individual fairness requires that similar individuals be treated similarly [76], whereas group fairness requires that different groups be treated similarly. As Jacobs and Wallach [123] explain, debates about individual versus group fairness, as well as debates about the right definitions of individual and group fairness, reflect an array of "different theoretical understandings" of what constitutes fairness. For example, some definitions of group fairness reflect concerns that predicted scores should have the same meaning for people from different demographic groups, while others reflect concerns that errors experienced by people of different groups should be comparable [123].

Despite these differences and the substantial debate they have engendered, conceptualizations of algorithmic fairness reflect some general patterns. For example, most fairness operationalizations formally represent only decision subjects; other stakeholders impacted by system predictions, such as business owners trying to make hiring decisions or the dependents of someone eligible for parole, are rarely represented. Much of the fairness literature has focused on settings where systems might withhold resources or opportunities from some people, such as recidivism [12], hiring [163, 217], and education [113, 173, 182], rather than settings where systems might represent some people unfavorably.[5][6] These focuses reflect, as Hoffmann [116] writes, "liberal anti-discrimination discourses in the law, which have historically sought to address injustices in the distribution and exercise of important rights, opportunities, and resources in domains like voting, housing, and employment." In addition to anti-discrimination law, conceptualizations of fairness often draw on political philosophy, particularly theories of distributive justice [30, 114, 135]. Critiques of the fairness literature have observed that it may leave assumptions about what constitutes fairness unexamined [123], treat fairness as a self-evidently appropriate framing [23], or place too much faith in a fairness framing's ability to address structural concerns [116]. Work addressing the ethical implications of predictive systems from perspectives other than fairness has also emerged, including analyses of systems' underlying logics and sociohistorical contexts [232], how systems reproduce power relations [139, 180], the values and incentives of the disciplines producing systems (e.g., machine learning) [32], and the labor upon which systems rely [122].

By contrast, within the ethical decision-making literature, concerns about system outputs manifest from two distinct scenarios. First, how might a sequential decision-making system cause harm due to an inadequate decision-making model? For example, if the decision-making model from §2 had only two options for price, HIGH and LOW, customers may be charged more than necessary in scenarios where the optimal action is to set the price to NORMAL instead of HIGH, since the decision-making model does not recognize this as a possibility. Similarly, if the number of neighborhoods used to model an area decreases, then some people may be left without power even when not strictly necessary since the agent cannot make more fine-grained decisions. Note that the fix for both of these examples is to make a more complex, and therefore more expensive, decision-making model. Second, how should the system behave when faced with a decision for which there is no obviously good outcome, or a high degree of risk? For example, if the power management agent must cut power to a neighborhood, how should it decide which neighborhood to cut? Some of the first ethical decision-making proof-of-concept systems were focused on military applications where there was potential for lethal use of force [13–15]. This application provides a context of rules for the moral conduct of warfare, studied extensively by ethicists and moral philosophers, which influence the way many scholars view ethical decision making today.

Ethical decision-making systems are considered ethical when their behavior aligns with a set of rules for acting, either learned or prescribed. These sets of rules are devised under the assumption that systems which follow those rules will minimize harm. This conceptualization sits in stark contrast to that of the fairness literature, where many fairness definitions are expressed in terms of relative failure rates or unevenly distributed system error.[7] In some sense, strict adherence to a set of rules for acting sets a much higher standard for agent behavior, though in practice expressive, effective, and general rule sets are exceedingly difficult to generate. Justification for this strategy often comes from moral philosophy, where ethical theories are broadly understood to provide such rules for acting. Several major ethical theories have been used to motivate autonomous systems where an agent is, morally speaking, required, permitted, or prohibited from taking specific actions in specific states depending on whether that action in that scenario violates the rules of the ethical theory. These theories include Act Utilitarianism [9, 100, 143], Kantianism [117, 206, 259], Virtue Ethics [148, 190, 236], Norm-based systems [89, 134], The Veil of Ignorance [156, 186], Divine Command Theory [42], The Golden Rule [186], and Prima Facie Duties [10, 236] among others. In addition to these applied works, there have been many more theoretical pieces examining when and why particular ethical frameworks ought to be used [83, 98, 115, 161, 198, 205, 206, 247, 266, 278]. However, these systems are still largely imagined, and we are not aware of any real-world systems yet in operation, in contrast to "fair" predictive systems which we know operate in a variety of public settings already.

3.1) Conceptualizations of ethical behavior for sequential decision-making systems are shaped by models' increased capacity for reasoning. Unlike predictive models, sequential decision-making models allow a system to reason explicitly about the effects of its actions, including long-term consequences. Thus, sequential decision-making systems are generally conceptualized as systems that act—and enact change—in the world, shaping what it means for a these systems to behave ethically. This represents a fundamentally different perspective on a system's role within its sociotechnical context compared to predictive systems [5, 265]. For example, many discussions of ethical decision making focus on long-term behavior [186, 236], whereas many conceptualizations of fairness incorporate no notion of either a system's future decisions or the downstream consequences of those decisions. Recently, *welfare*, which generally measures holistic outcome effects, rather than error rates, has been proposed as an alternative measure [87, 118, 131], bringing evaluation ethea in ethical decision making and fairness slightly closer. Specifically, framings using welfare allow detailed longitudinal analyses previously scarce in fairness literature, and suggest conceptualizations of predictive systems in their contexts of use rather than as standalone systems.

3.2) While concerns in ethical decision making are rarely articulated in terms of fairness, fairness may nevertheless offer a useful lens for evaluating the outcomes of deployed sequential decision-making systems. For example, the MDP given in §2 may

[5] Barocas et al. [17] and Crawford [66] refer to these as allocative and representational harms, respectively.

[6] This is not to say, of course, that there has not been ample work examining the latter, for example Sweeney [238] on discrimination in online ads and Noble [193] on search engines' reproduction of racial and gender stereotypes.

[7] Raji et al. [209] observe that fairness research may often neglect to ask whether systems function in the first place.

be more likely to interrupt service to neighborhoods with a particular demographic. We will discuss potential underlying causes and intervention strategies in later sections, but here we simply highlight that at a high level, fairness-type audits of such a system could potentially detect this type of behavior and that research on how to do these types of audits for sequential decision-making systems is absent from the literature. Similarly, research addressing *if and when* fairness is the right construct for analyzing sequential decision-making system outcomes is also absent.

4 SYSTEM DEVELOPMENT

In this section, we examine common prediction and sequential decision-making pipelines, cover some important differences—including system inputs and outputs, how expert knowledge is encoded, and informatic assumptions—and discuss what these differences suggest as objects of analysis for future research. Predictive systems have two components: data and a function approximator. The goal is to learn a function that can predict some hidden variable using data. Ideally, the data is accurately labeled, representative of the deployment setting, and plentiful enough to train a model. These assumptions of course may not all be met by development or deployment conditions, and while these are common topics of research in the broader machine learning community, fairness researchers have also proposed methods for handling flawed data [16, 47, 55, 170, 255], out of training distribution data [227, 241], and efficient learning [228].

4.1) By contrast, the product of a sequential decision-making system is a policy that when executed results in a sequence of actions taken in the world. Instead of function approximators, sequential decision-making systems use planners. Often, these planners produce *provably optimal* policies, meaning that, with respect to its model, the agent maximizes its cumulative expected reward. The existence of a policy instead of a set of i.i.d. decisions means that understanding system behavior is more contextual and not always possible with the same statistical measures. Because policies represent situation-dependent prescriptions for actions, analyzing a policy requires inspecting the action that would be prescribed by the policy for every state. This represents a major departure from aggregate measures of monitoring behavior.

4.2) When designing a decision-making model, developers hypothesize about the structure and value of unseen data, rather than extracting patterns from existing data. In other words, developers write down what they think is (or will be) true about the world [39, 218]. For example, when specifying the reward function, they decide the relative utility of outcomes; when enumerating the state space, they forecast the importance and availability of different data. Therefore, the most important and most fallible assumption in sequential decision making is that the model is faithful enough to the dynamics of the real world to support effective decision making. For example, if the transition function representing changes in demand is not perfect, then there may be scenarios when the agent's actions are optimal with respect to the model, but not the real world.[8] Problems may also arise from under-specified state spaces, as in the power grid example from §3. Adding more state factors or increasing their domains creates exponentially and polynomially more states, respectively. This tradeoff between model descriptiveness and computational tractability is not present in predictive systems since typically defining the desired classes or the meaning of the numerical output is not a fundamentally intractable problem.[9]

While tricky to get right, the practice of crafting models has several advantages compared to learning from data alone. The ability to provide an initial model of the world, even if refined using data, is useful as it allows developers to encode knowledge and model feedback effects that otherwise may be difficult to learn. Thus, developers spend considerable effort in creating decision-making models that are as compact and descriptive as possible, and experts are highly valued for their ability to design tractable, accurate models. Often, these models also require specific domain expertise, and while there are some cases in which data may be gathered to enhance model building or, most often, to improve the transition function, generally data is not available for the task at hand, and in some cases relevant data may be unavailable altogether.

4.3) The designs of states, actions, and rewards are obvious objects of analysis for decision-making models. The process of defining the state and action spaces and reward functions can be thought of as a structured way of encoding expert domain knowledge. Through implicit assumptions about how the world works, the purpose of the system, the source of data, or the responsibility of the agent for certain outcomes, developers encode expert knowledge about decision-making scenarios. This knowledge is not only important, but required, in order to make problems manageable computationally. However, it is also through these mechanisms that decisions are made which may cause harm.

For predictive systems, expert knowledge is often exploited via careful curation and selection of data, or through choosing the spaces of inputs and outputs (labels, classes, or target variables). The space of outputs is typically driven by the task. It may map directly from a description of the decision to be made, such as whether a manufactured part passes a quality control test, or it may correspond to a component of a larger decision-making task, such as calculating recidivism risk for determining bail. The set of inputs is often more difficult to determine and has historically attracted more attention in the fairness literature [141] than in the ethical decision-making literature. First, we note that all features are proxy data for the variable of interest; all inputs to such predictive models are pieces of data that modelers believe either cause, or at the very least correlate with, the variable of interest. Choices of inputs and features create two concerns. First, from an engineering perspective, the choice of inputs can be somewhat of a "dark art," and even with the advent of deep learning it is still often challenging to understand if a function is challenging to learn because of uninformative features or some other phenomenon. Second, including data that is unlikely to be causal and may be correlated with data protected under non-discrimination law can lead to concerns that labels are being assigned due to protected attributes. This is especially difficult when there are proxy variables that correlate with both the variable of interest and protected variables.

[8] It is possible to learn or refine the transition and reward functions from data. Reinforcement learning is used to learn the transition function by sampling actions and inverse reinforcement learning is used to learn the reward function by observing sequences of state-action pairs from an agent running a policy.

[9] While increasing the number of classes does increase the data required to learn robust models, a larger space of class labels alone does not make the problem computationally intractable.

Sequential decision-making systems offer analogous choices, corresponding to the choice of state and action spaces and the reward function. The action space, like the space of labels, is typically more straightforward to conceive and less controversial, since it simply represents the capabilities of the agent. For example, there are a limited number of commands an autonomous car can give its motors and thus it is usually clear what is appropriate and what is not.[10] The choice of state factors, on the other hand, creates two concerns. First, from an engineering perspective, the tradeoff between computational tractability and expressive decision-making power is hard to get right. While the complexity of solving an MDP is polynomial in the size of the state space, the size of the state space scales exponentially with respect to the number of state factors. This often severely limits the number of state factors that an agent can use for decision making. The creation of state spaces that are small enough to solve policies for, but also do not obfuscate important nuance by abstracting away important details of the situation, is also something of a "dark art."

From an ethical perspective, both the state space as well as the reward function may be poorly designed and cause harm. For example, given a transition function that describes the *true* probability of a power shortage in a certain neighborhood, the MDP may make a decision to cut power or raise rates in order to balance the load of the entire network. This true probability may be accurate, and the MDP may be taking the optimal action with respect to both its model and the real-world application, but this probability may also be influenced by societal forces, like neglect of local infrastructure, that correlate with a sensitive attribute, such as race. Behaving in accordance with certain ethical norms may create need for additional state factors beyond those required to complete the task. If this additional detail is not encoded in the state space, the resultant policy will not be able to distinguish between scenarios where multiple actions are roughly equivalent with respect to the task but have drastically different ethical implications. An important corollary to this is that fairness is usually not useful as an optimization constraint since protected attributes are typically not encoded in the state space. Moreover, while well-designed MDPs only use state factors that are important to making decisions for the task and omit data not related to the decision, such as sensitive attributes, there may still be insidious correlations.

Even given a descriptive and compact state space, designing a decision-making agent that behaves ethically requires a reward function. The reward function presents another unique challenge in that it implicitly combines and homogenizes all possible outcomes onto the same numerical scale. Thus in a regular MDP, no matter how large and descriptive the state space, all good and bad outcomes, regardless of how they might be measured and who or what they might affect, are converted into the same "unit." Setting aside debates about whether this is even possible to do in a principled manner [144, 222, 257], this still challenges developers and creates a potential source of error. Finally, although the true transition function is fixed given the state and action spaces, and thus not usually considered a design choice, most MDPs do not have perfectly accurate transition functions and inaccurate transition estimates can also lead to undesired behavior.

[10] This is less clear in systems that use hierarchies of sequential decision-making systems [105], but the vast majority of real-world applications do not abstract sequential decision making to more than one level.

4.4) Design processes for decision-making models often lack scrutiny. As with prediction, the "dark art" of the design process, particularly regarding choice of state factors, means it is generally not a topic of discussion in the research community, let alone available for public scrutiny. Much of the design process is done by instinct, relying on domain experts, and is not well-codified or written down. Reasons behind decisions are often not available via publications or other documents, and the final models, if believed to be interpretable in their own right, may be taken as self-evidently appropriate and therefore under-inspected. This pattern is exemplified here [68, 69, 138], where the relevant variables are simply stated without further explanation or justification. Because of the proof-of-concept nature of most existing work on ethical decision making, the pattern of implicit justification extends to this research as well. Most works do not use metrics based on specific attributes and instead examine how to specify high-level abstract rules. In these cases, the justification for these omissions is that the particulars of the state factors or rewards are simply placeholders used to study the effect of the rule, for example [71, 246, 248, 261]. However, as researchers begin designing systems for more specific applications and with intention to deploy them, this justification will need to be critically examined.

These observations suggest a number of research questions. For example, for a particular decision-making model, which stakeholders are explicitly modeled? What kinds of approximations of the world are common, and what assumptions underpin them? Whose domain expertise is solicited? Are model aspects borrowed from one application to another, or developed afresh? How does model design take into account the larger systems that models participate in? We are not aware of research that explicitly studies processes related to the design and development of *sequential decision-making systems*, although there are substantial bodies of adjacent literature on interpretability [43, 178, 199, 243], explainability [26, 129, 187], and on participatory design generally [159, 168, 174].

5 SYSTEM EVALUATION AND MEASUREMENT

Ideals regarding how a system ought to behave in the abstract are only as good as our ability to show, either theoretically or empirically, that they will adhere to these ideals when deployed. Research on operationalizing fair or ethical behavior makes up a significant percentage of papers published at FAccT and similar conferences. Here, we examine different metrics, tools, and strategies employed by researchers in fairness and ethical decision making.

For predictive systems, measuring fairness means operationalizing some conceptualization of fairness, typically by statistically analyzing a system's performance over one or more groups of users [120]. These measurements might look for predictive parity [2, 53, 73], error rate balance [121, 276], or anti-classification [125, 132, 270], and are often focused on counterfactual analysis, either at the group [77, 150, 231] or individual level [125, 185]. Evaluating predictive systems with these measures requires access to the predictive model, and either real-world data or high-quality simulated data. When data is readily available, only a computer capable of running the model is required for evaluation. Whether these measures ultimately track the quality of the outcomes for users, however, is still an open question [97].

5.1) Sequential decision-making systems must be deployed to fully evaluate them. While predictive systems are similar in that some harms may only be evident in deployment, when the system is viewed holistically in its social context, researchers can at least take measures of fairness absent a deployed system. However, shortcomings of decision-making models are nearly impossible to uncover without an agent operating in the real world, taking actions, affecting its environment, and encountering real-world data from the resultant states. Thus, the only way to systematically understand harmful outcomes is to deploy, which is expensive, time-consuming, and often unsafe. To emphasize, most policies are optimal with respect to their model and thus abide by constraints of their model. However, assumptions made by the model may not reflect the real world and thus lead to unintended behavior.

5.2) Sequential decision-making models are often embedded in larger systems. Often, MDPs are part of a larger system, such as a robot [6, 45, 154, 234], and it may be challenging to write a reward function that represents its high-level task, which may be a mixture of several objectives. Thus, we often evaluate these decision-making models using a task-based metric [6, 50, 151, 219, 229]. For example, consider a robot running a policy for loading boxes into a truck. We can compare policies generated by different decision-making models, regardless of how their reward functions represent the task, simply by counting the number of boxes loaded into the truck. This seems simple, but is often the most costly and tedious experiment to run since it requires a fully functioning system, and makes predicting the ethical impact of different interventions even more challenging. Given the high costs and risks associated with the process, how might we safely approach system evaluation for ethical concerns, and how might we incentivize doing so?

5.3) Though imperfect, many proxy measures for policy quality exist. Here, we understand policy quality as a holistic measure of the fitness of a policy for a given application, which includes not only a policy's efficacy in completing a task, but also whether or not the actions taken by the agent are appropriate normatively or morally. A policy's fitness may be affected by many factors, including the implicit incentives indicated by the reward function, the accuracy of the transition function with respect to the real-world dynamics being modeled, and any additional constraints enforced by the planner.

Exact planners are optimal so we rarely evaluate the planning algorithm, but rather the decision-making model and its ability to produce accurate real-world decisions. One common technique is to simply spot-check the policy at different states where the agent is balancing competing reward signals to verify it behaves as expected. A more methodical strategy is to calculate the probability of reaching a certain bad state if the agent follows the optimal policy and begins in a given state. For example, we could compute the probability that neighborhood N_i experiences a service interruption during the next year. This type of analysis can uncover some policy errors, but is limited due to (1) the difficulty in enumerating all bad states or outcomes and (2) humans' poor intuition for likelihoods of different events. In short, sanity checking policies in this way is time-consuming and prone to error.

Moving towards in situ evaluation, there are several methods for evaluating policies that rely on the ability to simulate deployments. One basic method is to simulate the agent executing a policy many times and calculate the variance in performance in an effort to understand its reliability. A more principled method, with a prediction analog, is to test the performance of the policy under a variety of transition functions, or "possible worlds," typically in terms of the total reward earned. For example, we may be uncertain about whether our transition function correctly represents the probability of change in demand following a change in price action—that is, whether our model correctly captures the relative probabilities of different events—and the robustness of our policy to potential errors in the model. This is similar to some predictive settings where training, testing, and validation data sets represent different distributions of data [184]. MDPs which are solved assuming a distribution over transition functions are called robust MDPs [22, 192]. There is also a large body of related work on "safe" policies or "safe" learning, where "safety" has been defined in terms of behavioral constraints [215, 252], policy ergodicity [181], risk metrics [86, 214], and the probability of improving a policy [242].

Finally, one important difference between prediction problems and sequential decision problems is that when deployed, predictive models generally cannot know whether the result of their inference is correct. By contrast, sequential decision-making systems always know that the chosen action was optimal in expectation with respect to the model. Moreover, they can immediately observe the reward for a given outcome, even if it is unexpected. What is unknown is whether the optimal decision with respect to the model is also the optimal decision for those affected by the decision. Of course, the reward may not be perfectly aligned with preventing harms, but the ability to examine performance longitudinally is a powerful benefit nonetheless.[11]

5.4) Methodical evaluation of sequential decision-making systems for ethical behavior is an open problem. We are not aware of rigorous empirical research on harms produced by deployed sequential decision-making systems, including basic questions such as "Who is harmed?" The question is more often framed in terms of rules violations, but even these studies are not common due to the more theoretical nature of most existing research and lack of access to sequential decision-making systems. Although there are many methods for evaluating policies with respect to their decision-making models, ethical decision-making researchers are almost completely in the dark when it comes to understanding the impact of their agents on the world outside the model. We view this as a critical shortcoming in existing research.

5.5) Currently, auditing sequential decision-making systems poses a serious logistical challenge to researchers. In order to audit most sequential decision-making systems, an auditor would

[11] We should note for completeness that there are many approximate techniques for solving sequential decision-making problems. When evaluating these techniques, for a fixed decision-making model, directly measuring the value function can often provide a signal as to the quality of the resultant policy with respect to the task since all value functions are upper bounded by the value function induced by the optimal policy—the policy we would get if we used an exact planner. Even better is measuring the actual cumulative reward experienced by the agent using simulation or data collected from a deployment. However, if the model is changed in any way, including the discount factor, these comparisons cannot be run across models. This is because changes to rewards, transition probabilities, or the discount factor can change the scale of the optimal value function, producing different upper bounds and therefore preventing fair comparison.

require the physical agent, the agent's policy, and any supporting software that connects the two, such as algorithms for determining states from data and controllers for executing actions specified by the policy. For example, auditing decision making in an autonomous car would require the car and its entire software stack in order to verify how it behaves in different scenarios. This is a significant obstacle for researchers interested in transparency and accountability. Moreover, since these systems are often functioning as part of a larger system, it can be difficult for the public or regulators to even know where systems are deployed. Academic [260] or community audits, such as audits of commercial image cropping algorithms [33, 63], are challenging if not impossible.

6 MITIGATIONS

Often, the goal is to modify, augment, or in some other way intervene in the decision-making process of an existing system in order to ensure that it behaves fairly or ethically. In this section, we examine common interventions, including data augmentation, reward modification, optimization formulation, and system integration. For predictive models one of the simplest interventions is to collect or generate more data (data augmentation), under the assumption that as more samples are acquired the training set will improve its approximation of the true distribution and the model will learn a more accurate, representative function [196, 210, 223, 249, 264, 275]. This practice works well if data deficiency is the *only* reason for poor performance. In many cases, the problem is not the amount of data but the quality of the data. Specifically, there are often artifacts in the data, such as correlations between attributes like race or gender and the target variable, that we do not want our predictive model to learn. One way to mitigate this is to try to balance the data set to remove these correlations within the data by adding new data points or editing existing ones (data curation) [49, 51, 155].

Data-based interventions present challenging tradeoffs. For example, gathering sensitive data may be required to ensure a balanced data set with respect to certain attributes, or verify a model meets certain fairness criteria. However, doing so raises privacy concerns [60, 79, 207] as well as accuracy concerns when sensitive attributes are not readily available or easily identifiable [95]. Moreover, many researchers have rightly pointed out that common operationalizations of race, gender, and other socially constructed concepts may in fact be more harmful than helpful [101, 136]. Nonetheless, this is often the only data available to these systems. There is somewhat of a paradox in wanting to avoid sensitive attributes influencing predictions and reifying certain categories, but requiring these attributes in order to verify these criteria [11].

Beyond data augmentation and curation, often collectively referred to as "pre-processing," fairness researchers have also introduced methods for mitigation known as in-processing [52, 254] and post-processing [52]. In-processing methods generally involve either modifying the predictive loss function, such as by using constraints [76, 93, 94, 106, 110, 137, 150, 152, 224, 269, 270] or regularization [3, 4, 24, 28, 29, 72, 125, 133], or adaptively re-weighting training examples [124, 147]. They may also apply constraints to latent representations within the classifier via disentanglement [142, 165, 195], contrastive learning [59, 146, 244, 277], or adversarial learning [38, 78, 80, 169, 237, 253, 271]. These methods can work well, but often can have complications arising from multiple or unknown sensitive attributes, conflicting or differential desired definitions of fairness, and decrease in interpretability.

Post-processing techniques post-hoc transform or calibrate the outputs of a model to fit a definition of fairness. For example, by calibrating outputs across different sub-groups [109]. One advantage of post-processing is that it only requires the predictions and sensitive attributes and not underlying model, making them applicable to a wider variety of scenarios.

6.1) Data augmentation and curation are often impractical or unsafe for sequential decision-making systems. While it is possible to use data from actual deployments, potentially along with reinforcement learning, to improve the accuracy of the transition function with respect to the real world, this is often very costly and occasionally unsafe. For example, letting the power management system from §2 take suboptimal actions in order to gather data about the true distribution of outcomes (successor states) of raising or lowering the price of power or shutting off service to different subsets of neighborhoods jeopardizes the general public's access to reliable, fairly priced power. This is clearly not acceptable, even if the end result is a more accurate transition function and thus a better decision-making model.

Generally speaking, MDP agents cannot be developed in isolation with data from elsewhere. The agent itself is required in order to generate data by interacting with the world—taking actions, recording state, and experiencing reward. Many researchers get around this problem using simulation, but again, many types of failures may occur in the real-world that do not show up in simulation, especially those related to ethical behavior. Thus, real-world deployments are often a bottleneck for gold-standard evaluation — so much so that issues of safety when gathering data for MDPs form the primary motivation for the field of safe reinforcement learning [45, 92, 102].

6.2) Computation constraints limit performance of sequential decision-making systems. While there is no direct analog of data augmentation or curation in MDPs, the decision-making models themselves are often augmented by expanding the state space. This is done by either adding new state factors or expanding the domains of existing ones, for example by adding new neighborhood factors which represent subsets of the original neighborhoods in the power management problem. This delineates some scenarios that were previously considered identical, allowing the agent to choose different actions under those conditions. By adding new neighborhoods, the agent has more fine-grained control over service interruptions and can maintain power to more homes. We may also add completely new state factors to the MDP, such as the existence of backup generators in some neighborhoods, that can be used to modify the reward function so that it reduces penalties for outages in these neighborhoods since the ultimate impact is reduced. Thus, larger, more descriptive state spaces often produce more nuanced, performant policies at the cost of time required to generate a policy.

Generally, in prediction, more computation cannot improve the estimate of a target variable. This is not so for sequential decision making. While MDP[12] solvers have polynomial complexity in the

[12] We should emphasize the tremendous volume of work on extending MDPs to other informatic settings. For example, state factors may not be directly observable [130], such as when a pedestrian becomes temporarily occluded from the view of an autonomous vehicle. Their position is unobservable, so the vehicle maintains a belief over the

number of states, the number of states often grows exponentially with the number state factors. This presents a challenge as improvements via state space augmentation are limited since adding state factors orthogonal to the task adds exponential cost [103, 104, 160]. Moreover, additional states which do not map to different actions increase computational cost without increasing performance. In practice, decision-making models are often *necessarily* approximations, which marginalize (in the computational sense) some variables or compress different scenarios into the same state representation to reduce model size and therefore compute load.

6.3) There are no existing, principled methods that can predict how design decisions regarding the state space or reward function affect stakeholders *in the general case*. Although understanding how changes to a decision-making model may affect resulting policies is an important part of designing MDPs, even for specific applications there is essentially no systematic method for predicting the impact of model changes on all stakeholders. For example, there is little research on whether decision-making models exhibit the same problems as predictive models, such as when removing protected attributes explicitly from the reasoning process does not prevent differential treatment with respect to those attributes. In the absence of a general account we therefore see many important research questions: What resources might be developed to help practitioners understand how to augment their state spaces or modify their reward functions? How might the research community contribute to making this process more systematic, such as via checklists or other design processes? How can practitioners anticipate what kinds of ethical scenarios must be delineated by the model beyond what is necessary for the task? By what processes can we reliably uncover and anticipate such scenarios—without risking stakeholders?

One complicating property for both developers and auditors is that state factors are not restricted to the data at hand—they may represent any information the agent can measure or sense once deployed. In some cases, this prevents models from reproducing historically biased patterns since unwelcome correlations simply are not present in the model. However, in other cases it makes correcting issues identified as disparities between protected classes more difficult as this data is not directly reasoned about within the model and thus cannot be directly constrained as it might be with, for example, in-processing techniques from classification. Simply adding protected attributes as state factors seems ill-advised since unless these factors affect the reward or transition functions they will not affect the reasoning process and will add exponential computational burden.

There remain other questions about what it means for protected attributes to be part of the reasoning process in sequential decision making. Even if protected attributes are not represented as state factors, might some factors still implicitly encode them, or might reward functions encode harmful patterns? For example, consider the power grid management agent from §2. The agent, in its effort to minimize outages, price, and wasted power, may disproportionately restrict access to some neighborhoods. Because neighborhood boundaries according to the utility infrastructure often correlate strongly with some demographic attribute, such as race or income, this policy may therefore disproportionately impact members of those groups. More generally, we still know little about how sequential decision-making outcomes may or may not reproduce patterns of discrimination within different applications.

pedestrian's location and thus a belief over the state of the world. Other specialized models have been developed for decentralized behavior [25], adversarial scenarios [90], and hierarchical decision processes [111] among many, many others. These models vary greatly in their assumptions and complexity, and understanding the feasibility of different interventions across different models is an open question.

6.4) As with predictive systems, the most straightforward interventions come with considerable drawbacks. One of the simplest ways developers modify the behavior of MDP agents is by modifying the reward function, which we call reward modification. [13] Reward modification has no direct analog in prediction, but is similar in spirit to tweaking a loss function in an asymmetric way that affects the model's penalty for incorrect labels on a subset of cases. The important similarities are that the intervention is local, it targets a specific behavior, the outcome has no formal guarantees since it is unknown how the optimization problem will be re-solved given the new loss or reward function, and that these interventions require a significant level of expertise, since the practitioner needs to understand how a given change affects some intermediate computation which ultimately affects behavior. Thus, reward modification is necessarily non-methodical. There is no theory that describes how to specify reward functions in order to generate a particular policy or behavior for an arbitrary MDP.

While this intervention is often the easiest, it is also the least effective. Not only is the control over the resultant policy indirect, but this method also leaves room for many tacit normative assumptions. In particular, it allows developers to make implicit comparisons between different types of outcomes due to the reward function mapping all possible outcomes onto the same "unit." As the agent maximizes expected cumulative reward, it inherently balances avoiding negative reward states and visiting positive reward states based on their respective reward values and the likelihood of reaching those states. Thus, there is always a future amount of positive reward for which the agent will accept experiencing a negative reward in the short term, no matter what real-world scenario that negative reward represents. This is one reason that decision-making model design is so difficult, and this problem is no simpler when modifying reward functions for ethical reasons. That said, in practice this is still the most popular method for modifying agent behavior.

6.5) Behavioral constraints on decision-making systems have several benefits. Beyond gathering more data, expanding decision-making models, and modifying loss functions or reward functions, there are more principled ways to control the behavior of decision-making systems. Generally, these methods constrain the optimization processes involved in determining behavior, and the similarities between techniques devised to produce "fair" and "ethical" behavior are remarkable. Attempts to train fair predictive models have used constraints [53], regularization [48, 132], and causal and counterfactual analysis [35, 36, 65]. These techniques essentially constrain

[13] In reinforcement learning systems with very sparse reward functions—reward functions where most states have the same value, usually zero—a similar sounding technique known as "reward shaping" is used to add reward signal to states which represent progress towards or away from one of the original, sparse reward signals. The idea is that the agent will learn faster as it has more frequent access to a learning signal. In the reinforcement learning application there is a lot of concern about executing reward shaping in a manner which does not alter the optimal policy one would get if they solved the original MDP using the original, sparse reward function (remember, however, this is not possible since they do not know the transition function). There are some theoretical results regarding how this may be done [191], but they do not apply to our case because we want to *change* a policy for a *known* MDP.

the space of possible *functions* that the model can learn. In ethical decision making we typically constrain the space of *policies* using domain-specific hand-coded rules [225, 261] or constraints [134, 186, 236, 248]. MDP agents that use constraints still compute a policy that maximizes cumulative expected reward, but do so subject to some constraints on, for instance, how often certain state-action pairs can occur. These pairs can be enumerated explicitly or identified via an abstract rule—in the ethical decision-making case, these are rules for acting. This type of MDP is called a constrained MDP (CMDP). However, generating the right constraints is difficult, and is comparable in difficulty to choosing a definition of fairness. There are no clearly best options, and the right choice in terms of satisfying as many stakeholders as possible is often context and deployment specific.

Constraint-based methods are more difficult to design and program but have several advantages. First, this is the only method that guarantees instance-level behavior, although constraints may also be defined in terms of aggregate or expected behavior such that individual decisions may not have guarantees. Second, these methods allow more direct, expressive behavior specification. Instead of changing the reward or loss function, or training using an adversarial agent [253, 273], we can encode precisely how the agent ought or ought not to behave. Limited only by the expressiveness of the model, the act of writing constraints also surfaces many normative assumptions explicitly. Third, these methods offer substantially greater potential for generalization. Constraints may be formulated in abstract terms, such as false negative rates or the probability of violating a norm, allowing their application to many different decision-making problems. Fourth, although mathematically more complex, these interventions often operate at a level of abstraction that can be communicated to non-experts. This is an important benefit since it allows a greater variety of expertise to be consulted in a given application. While theoretically such constraints have many advantages, these methods are not frequently deployed due to their complexity.

6.6) Non-mathematical and auxiliary interventions, such as human-in-the-loop solutions and explainability, are under-studied in the ethical decision-making context. Increasing the ability of users or auditors to understand, interact, or correct automated decision-making systems is likely to increase the effectiveness of many existing interventions and perhaps lead to new methods altogether. Interpretable and explainable AI systems are of course large fields in their own right; however, there is a relative lack of research on explainable sequential decision making compared to predictive systems. Not only is there still foundational algorithmic work to be done, but there are also open conceptual questions such as how ideas of actionable recourse [18, 245] or cross-examination [1] might be applied to this setting. Similarly, human-in-the-loop systems have been proposed and studied in the sequential decision-making literature [84, 153, 262], but outside of military research [8, 46, 176, 239], rarely if ever as problems with explicit ethical consequences.

6.7) Sequential decision-making models are not generally developed with engagement from the full spectrum of stakeholders. The opportunity that decision-making models and explicit constraints allow for leveraging expert knowledge cannot be understated. However, current practices in academia and industry do not take full advantage of these benefits in part because they lack exposure to, and knowledge of, qualitative or participatory processes [159, 168, 174]. By contrast, disciplines such as human-computer interaction have well-developed approaches for engaging with stakeholders, ranging from user-centered design practices [112] to participatory approaches, where stakeholders work with researchers in a process of collective inquiry [251], or where stakeholders participate in system design and development processes [149].

7 DISCUSSION AND CONCLUSION

Algorithmic fairness interventions are developed for only a subset of the algorithms deployed in the world. In this paper, we draw attention to sequential decision-making models, which are the subject of an increasingly rich literature on ethical decision-making, and describe how interventions for fair or ethical behavior are currently conceptualized and operationalized across the fairness and ethical decision-making communities. We further ask: Where might the two communities benefit from one another? And where might the paradigm of sequential decision making demand different interventions than those developed for predictive models?

Towards the first question, we explore how the fairness and ethical decision-making communities may benefit from knowledge, tools, and practices emerging from one field or the other. In one direction, methods for sequential decision making and modes of analysis from ethical decision making have the potential to advance fairness research given recent calls to examine feedback effects of systems on stakeholder welfare—two themes that have been researched extensively by these communities. In the other, the widespread deployment of sequential decision-making systems and the domains in which they operate (e.g., autonomous driving, power grid management) makes urgent analyses of these systems—analyses which the fairness literature is already undertaking for predictive systems. Here, we imagine analyses of the sequential decision-making model and the processes by which models are designed; the outcomes of decision-making systems in terms of which stakeholders are harmed, particularly the ways in which outcomes might reproduce existing patterns of injustice; and how choices regarding the design of decision-making models give rise to particular outcomes. Alongside these analyses, what processes and resources might we develop to help anticipate the outcomes of a given model and policy, and support safe iterative model development, without incurring too much of the risk inherent to deployment?[14] Although sequential decision making may demand new methods for carrying out these analyses, we can draw on the lenses—questions about disparities in outcomes and the processes that produce them—that have emerged from years of rich discussion in the fairness community.

Nevertheless, decision-making models are in many ways fundamentally different from predictive models, and their reasoning capabilities, design, and deployment will make realizing these goals difficult. We have illustrated that some of the interventions— conceptualizations of normative concerns and their accompanying measurements and mitigations—that have entered best practice from the fairness literature may not be applicable to sequential decision making. Moreover, addressing many questions about the design processes, modification, and outcomes of these systems would be prohibitively expensive and likely risky to stakeholders, meaning

[14] Bird et al. [31] raise a similar question about the risks of autonomous experimentation.

that such work is likely to be disincentivized in the private sector. Complicating efforts, decision-making systems often operate below the awareness of the public and many regulatory bodies, because they do not tend to make decisions that directly affect individual people. Practical efforts to realize these efforts will require a realistic account of what sequential decision-making systems look like, and of how well our assumptions about what it takes to make fair, transparent, or accountable predictive systems serve us in this different setting.

REFERENCES

[1] Rediet Abebe, Moritz Hardt, Angela Jin, John Miller, Ludwig Schmidt, and Rebecca Wexler. 2022. Adversarial Scrutiny of Evidentiary Statistical Software. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 1733–1746.

[2] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. 2018. A reductions approach to fair classification. In *Proceedings of the 37th International Conference on Machine Learning*. 60–69.

[3] Alekh Agarwal, Miroslav Dudík, and Zhiwei Steven Wu. 2019. Fair regression: Quantitative definitions and reduction-based algorithms. In *International Conference on Machine Learning*. PMLR, 120–129.

[4] Sina Aghaei, Mohammad Javad Azizi, and Phebe Vayanos. 2019. Learning optimal and fair decision trees for non-discriminative decision-making. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 1418–1426.

[5] Fahad Alaieri and André Vellino. 2016. Ethical decision making in robots: Autonomy, trust and responsibility. In *International conference on social robotics*. Springer, 159–168.

[6] Hub Ali, Dawei Gong, Meng Wang, and Xiaolin Dai. 2020. Path planning of mobile robot with improved ant colony algorithm and MDP to produce smooth trajectory in grid-based environment. *Frontiers in neurorobotics* 14 (2020), 44.

[7] Abdulaziz A Almuzaini, Chidansh A Bhatt, David M Pennock, and Vivek K Singh. 2022. ABCinML: Anticipatory Bias Correction in Machine Learning Applications. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 1552–1560.

[8] Robert St. Amant, Ralph Brewer, and Mary Anne Fields. 2021. Who is responsible for robot behavior?. In *Proceedings of the 1st IROS Workshop on Building and Evaluating Ethical Robotic Systems*.

[9] Michael Anderson, Susan Anderson, and Chris Armen. 2005. Towards machine ethics: Implementing two action-based ethical theories. In *Proceedings of the AAAI 2005 Fall Symposium on Machine Ethics*. AAAI, Pittsburgh, PA, USA, 1–7.

[10] Susan Leigh Anderson and Michael Anderson. 2011. A prima facie duty approach to machine ethics and its application to elder care. In *Proceedings of the AAAI Workshop on Human-Robot Interaction in Elder Care*. AAAI, San Francisco, CA, USA, 2–7.

[11] McKane Andrus and Sarah Villeneuve. 2022. Demographic-Reliant Algorithmic Fairness: Characterizing the Risks of Demographic Data Collection in the Pursuit of Fairness. *arXiv preprint arXiv:2205.01038* (2022).

[12] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias. *ProPublica* (May 2016).

[13] Ronald Arkin. 2008. Governing lethal behavior: Embedding ethics in a hybrid deliberative/reactive robot architecture. In *Proceedings of the 3rd ACM/IEEE International Conference on Human Robot Interaction*. Amsterdam, The Netherlands.

[14] Ronald Arkin. 2009. *Governing lethal behavior in autonomous robots*. Chapman and Hall/CRC.

[15] Ronald C Arkin, Patrick D Ulam, and Brittany Duncan. 2009. *An ethical governor for constraining lethal action in an autonomous system*. Technical Report. Georgia Institute of Technology.

[16] Pranjal Awasthi, Alex Beutel, Matthäus Kleindessner, Jamie Morgenstern, and Xuezhi Wang. 2021. Evaluating fairness of machine learning models under uncertain and incomplete information. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*. 206–214.

[17] Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. 2017. The problem with bias: From allocative to representational harms in machine learning. In *Proceedings of the SIGCIS Conference*.

[18] Solon Barocas, Andrew D Selbst, and Manish Raghavan. 2020. The hidden assumptions behind counterfactual explanations and principal reasons. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 80–89.

[19] Andrew G Barto, Steven J Bradtke, and Satinder P Singh. 1995. Learning to act using real-time dynamic programming. *Artificial Intelligence* 72, 1-2 (1995), 81–138.

[20] Connor Basich, Justin Svegliato, Kyle Hollins Wray, Stefan Witwicki, Joydeep Biswas, and Shlomo Zilberstein. 2020. Learning to optimize autonomy in competence-aware systems. In *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems*.

[21] Richard Bellman. 1966. Dynamic programming. *Science* (1966).

[22] Aharon Ben-Tal and Arkadi Nemirovski. 1998. Robust convex optimization. *Mathematics of Operations Research* 23, 4 (1998), 769–805.

[23] Cynthia L Bennett and Os Keyes. 2020. What is the point of fairness? Disability, AI and the complexity of justice. *ACM SIGACCESS Accessibility and Computing* 125 (2020), 1–1.

[24] Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. 2017. A convex framework for fair regression. *arXiv preprint arXiv:1706.02409* (2017).

[25] Daniel S Bernstein, Robert Givan, Neil Immerman, and Shlomo Zilberstein. 2002. The complexity of decentralized control of Markov decision processes. *Mathematics of Operations Research* 27, 4 (2002), 819–840.

[26] Josh Bertram and Peng Wei. 2018. Explainable deterministic MDPs. *arXiv preprint arXiv:1806.03492* (2018).

[27] Dimitri P Bertsekas. 2011. Approximate policy iteration: A survey and some new methods. *Journal of Control Theory and Applications* 9, 3 (2011), 310–335.

[28] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Li Wei, Yi Wu, Lukasz Heldt, Zhe Zhao, Lichan Hong, Ed H Chi, et al. 2019. Fairness in recommendation ranking through pairwise comparisons. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 2212–2220.

[29] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Allison Woodruff, Christine Luu, Pierre Kreitmann, Jonathan Bischof, and Ed H Chi. 2019. Putting fairness principles into practice: Challenges, metrics, and improvements. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 453–459.

[30] Reuben Binns. 2018. Fairness in machine learning: Lessons from political philosophy. In *Proceedings of the Conference on Fairness, Accountability and Transparency*. 149–159.

[31] Sarah Bird, Solon Barocas, Kate Crawford, Fernando Diaz, and Hanna Wallach. 2016. Exploring or exploiting? Social and ethical implications of autonomous experimentation in AI. In *Proceedings of the Workshop on Fairness, Accountability, and Transparency in Machine Learning*.

[32] Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2022. The Values Encoded in Machine Learning Research. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 173–184.

[33] Abeba Birhane, Vinay Uday Prabhu, and John Whaley. 2022. Auditing Saliency Cropping Algorithms. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 4051–4059.

[34] Ondrej Biza and Robert Platt. 2018. Online abstraction with MDP homomorphisms for deep learning. *arXiv preprint arXiv:1811.12929* (2018).

[35] Emily Black and Matt Fredrikson. 2021. Leave-one-out unfairness. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*. 285–295.

[36] Emily Black, Samuel Yeom, and Matt Fredrikson. 2020. Fliptest: Fairness testing via optimal transport. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 111–121.

[37] Avrim Blum, Kevin Stangl, and Ali Vakilian. 2022. Multi Stage Screening: Enforcing Fairness and Maximizing Efficiency in a Pre-Existing Pipeline. *arXiv preprint arXiv:2203.07513* (2022).

[38] Avishek Bose and William Hamilton. 2019. Compositional fairness constraints for graph embeddings. In *International Conference on Machine Learning*. PMLR, 715–724.

[39] Richard J Boucherie and Nico M Van Dijk. 2017. *Markov decision processes in practice*. Springer.

[40] Djallel Bouneffouf, Irina Rish, and Charu Aggarwal. 2020. Survey on applications of multi-armed and contextual bandits. In *2020 IEEE Congress on Evolutionary Computation*. IEEE, 1–8.

[41] Dimitrios Bountouridis, Jaron Harambam, Mykola Makhortykh, Mónica Marrero, Nava Tintarev, and Claudia Hauff. 2019. SIREN: A simulation framework for understanding the effects of recommender systems in online news environments. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 150–159.

[42] Selmer Bringsjord and Joshua Taylor. 2012. The divine-command approach to robot ethics. *Robot ethics: The ethical and social implications of robotics* (2012), 85–108.

[43] Alexander Brown and Marek Petrik. 2018. Interpretable reinforcement learning with ensemble methods. *arXiv preprint arXiv:1809.06995* (2018).

[44] Cameron B Browne, Edward Powley, Daniel Whitehouse, Simon M Lucas, Peter I Cowling, Philipp Rohlfshagen, Stephen Tavener, Diego Perez, Spyridon Samothrakis, and Simon Colton. 2012. A survey of Monte Carlo tree search methods. *IEEE Transactions on Computational Intelligence and AI in Games* 4, 1 (2012), 1–43.

[45] Lukas Brunke, Melissa Greeff, Adam W Hall, Zhaocong Yuan, Siqi Zhou, Jacopo Panerati, and Angela P Schoellig. 2022. Safe learning in robotics: From learning-based control to safe reinforcement learning. *Annual Review of Control, Robotics, and Autonomous Systems* 5 (2022), 411–444.

[46] Don Brutzman, Curtis L Blais, Duane T Davis, and Robert B McGhee. 2018. Ethical mission definition and execution for maritime robots under human supervision. *IEEE Journal of Oceanic Engineering* 43, 2 (2018), 427–443.

[47] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 77–91.

[48] Robin Burke, Nasim Sonboli, and Aldo Ordonez-Gauger. 2018. Balanced neighborhoods for multi-sided fairness in recommendation. In *Proceedings of the Conference on Fairness, Accountability and Transparency*. 202–214.

[49] William Cai, Ro Encarnacion, Bobbie Chern, Sam Corbett-Davies, Miranda Bogen, Stevie Bergman, and Sharad Goel. 2022. Adaptive Sampling Strategies to Construct Equitable Training Datasets. *arXiv preprint arXiv:2202.01327* (2022).

[50] Gerard Canal, Michael Cashmore, Senka Krivić, Guillem Alenyà, Daniele Magazzeni, and Carme Torras. 2019. Probabilistic planning for robotics with ROSPlan. In *Annual Conference Towards Autonomous Robotic Systems*. Springer, 236–250.

[51] Yushi Cao, David Berend, Palina Tolmach, Guy Amit, Moshe Levy, Yang Liu, Asaf Shabtai, and Yuval Elovici. 2022. Fair and accurate age prediction using distribution aware data curation and augmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 3551–3561.

[52] Simon Caton and Christian Haas. 2020. Fairness in machine learning: A survey. *arXiv preprint arXiv:2010.04053* (2020).

[53] L Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K Vishnoi. 2019. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 319–328.

[54] Adriane Chapman, Philip Grylls, Pamela Ugwudike, David Gammack, and Jacqui Ayling. 2022. A Data-driven analysis of the interplay between Criminological theory and predictive policing algorithms. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 36–45.

[55] Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. 2019. Fairness under unawareness: Assessing disparity when protected class is unobserved. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 339–348.

[56] Sheng-Lei Chen and Yan-Mei Wei. 2008. Least-squares SARSA(Lambda) algorithms for reinforcement learning. In *Proceedings of the 4th International Conference on Natural Computation*, Vol. 2. IEEE, 632–636.

[57] Yifang Chen, Alex Cuellar, Haipeng Luo, Jignesh Modi, Heramb Nemlekar, and Stefanos Nikolaidis. 2020. The fair contextual multi-armed bandit. In *19th International Conference on Autonomous Agents and Multi-Agent Systems*.

[58] Yifang Chen, Alex Cuellar, Haipeng Luo, Jignesh Modi, Heramb Nemlekar, and Stefanos Nikolaidis. 2020. Fair contextual multi-armed bandits: Theory and experiments. In *Uncertainty in Artificial Intelligence*. PMLR, 181–190.

[59] Pengyu Cheng, Weituo Hao, Siyang Yuan, Shijing Si, and Lawrence Carin. 2021. Fairfil: Contrastive neural debiasing method for pretrained text encoders. *arXiv preprint arXiv:2103.06413* (2021).

[60] Victoria Cheng, Vinith M Suriyakumar, Natalie Dullerud, Shalmali Joshi, and Marzyeh Ghassemi. 2021. Can you fake it until you make It? Impacts of differentially private synthetic data on downstream classification fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 149–160.

[61] Isabel Chien, Nina Deliu, Richard Turner, Adrian Weller, Sofia Villar, and Niki Kilbertus. 2022. Multi-disciplinary fairness considerations in machine learning for clinical trials. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 906–924.

[62] Alexandra Chouldechova, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. 2018. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Proceedings of the Conference on Fairness, Accountability and Transparency*. 134–148.

[63] Rumman Chowdhury and Jutta Williams. 2021. Introducing Twitter's first algorithmic bias bounty challenge. *URl: https://blog. twitter. com/engineering/en_us/topics/insights/2021/algorithmic-bias-bountychallenge* (2021).

[64] Houston Claure, Yifang Chen, Jignesh Modi, Malte Jung, and Stefanos Nikolaidis. 2020. Multi-armed bandits with fairness constraints for distributing resources to human teammates. In *2020 ACM/IEEE International Conference on Human-Robot Interaction*. 299–308.

[65] Amanda Coston, Alan Mishler, Edward H Kennedy, and Alexandra Chouldechova. 2020. Counterfactual risk assessments, evaluation, and fairness. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 582–593.

[66] Kate Crawford. 2017. The Trouble with Bias. Keynote at the Conference on Neural Information Processing Systems.

[67] Alexander D'Amour, Hansa Srinivasan, James Atwood, Pallavi Baljekar, D Sculley, and Yoni Halpern. 2020. Fairness is not static: Deeper understanding of long term fairness via simulation studies. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 525–534.

[68] Lucas de Azevedo Fernandes, Thadeu Brito, Luis Piardi, José Lima, and Paulo Leitão. 2020. A real framework to apply collaborative robots in upper limb rehabilitation. In *Proceedings of the 13th International Joint Conference on Biomedical Engineering Systems and Technologies*. 176–183.

[69] Nelson De Moura, Raja Chatila, Katherine Evans, Stéphane Chauvier, and Ebru Dogan. 2020. Ethical decision making for autonomous vehicles. In *Proceedings of the 2020 IEEE Intelligent Vehicles Symposium*. IEEE, 2006–2013.

[70] Thomas L Dean, Robert Givan, and Sonia Leach. 1997. Model reduction techniques for computing approximately optimal solutions for Markov decision processes. *arXiv preprint arXiv:1302.1533* (1997).

[71] Louise Dennis, Michael Fisher, Marija Slavkovik, and Matt Webster. 2016. Formal verification of ethical choices in autonomous systems. *Robotics and Autonomous Systems* 77 (2016), 1–14.

[72] Pietro G Di Stefano, James M Hickey, and Vlasios Vasileiou. 2020. Counterfactual fairness: removing direct effects through regularization. *arXiv preprint arXiv:2002.10774* (2020).

[73] William Dieterich, Christina Mendoza, and Tim Brennan. 2016. *COMPAS risk scales: Demonstrating accuracy equity and predictive parity*. Technical Report.

[74] Kate Donahue and Solon Barocas. 2021. Better together? How externalities of size complicate notions of solidarity and actuarial fairness. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*. 185–195.

[75] Yuhao Du, Stefania Ionescu, Melanie Sage, and Kenneth Joseph. 2022. A Data-Driven Simulation of the New York State Foster Care System. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 1028–1038.

[76] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through Awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. 214–226.

[77] Cynthia Dwork, Nicole Immorlica, Adam Tauman Kalai, and Max Leiserson. 2018. Decoupled classifiers for group-fair and efficient machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 119–133.

[78] Harrison Edwards and Amos Storkey. 2015. Censoring representations with an adversary. *arXiv preprint arXiv:1511.05897* (2015).

[79] Michael D Ekstrand, Rezvan Joshaghani, and Hoda Mehrpouyan. 2018. Privacy for all: Ensuring fair and equitable privacy protections. In *Proceedings of the Conference on Fairness, Accountability and Transparency*. 35–47.

[80] Yanai Elazar and Yoav Goldberg. 2018. Adversarial removal of demographic attributes from text data. *arXiv preprint arXiv:1808.06640* (2018).

[81] Hadi Elzayn, Shahin Jabbari, Christopher Jung, Michael Kearns, Seth Neel, Aaron Roth, and Zachary Schutzman. 2019. Fair algorithms for learning in allocation problems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 170–179.

[82] Danielle Ensign, Sorelle A Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. 2018. Runaway feedback loops in predictive policing. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 160–171.

[83] Mirko Farina, Petr Zhdanov, Artur Karimov, and Andrea Lavazza. 2022. AI and society: a virtue ethics approach. *AI & SOCIETY* (2022), 1–14.

[84] Lu Feng, Clemens Wiltsche, Laura Humphrey, and Ufuk Topcu. 2016. Synthesis of human-in-the-loop control protocols for autonomous systems. *IEEE Transactions on Automation Science and Engineering* 13, 2 (2016), 450–462. https://doi.org/10.1109/TASE.2016.2530623

[85] Norm Ferns, Prakash Panangaden, and Doina Precup. 2004. Metrics for finite Markov decision processes. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, Vol. 4. 162–169.

[86] Seyedshams Feyzabadi and Stefano Carpin. 2014. Risk-aware path planning using hirerachical constrained markov decision processes. In *2014 IEEE International Conference on Automation Science and Engineering*. IEEE, 297–303.

[87] Jessie Finocchiaro, Roland Maio, Faidra Monachou, Gourab K Patro, Manish Raghavan, Ana-Andreea Stoica, and Stratis Tsirtsis. 2021. Bridging machine learning and mechanism design towards algorithmic fairness. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*. 489–503.

[88] Benjamin Fish, Jeremy Kun, and Ádám D Lelkes. 2016. A confidence-based approach for balancing fairness and accuracy. In *Proceedings of the 2016 SIAM international conference on data mining*. SIAM, 144–152.

[89] Spencer Frazier, Md Sultan Al Nahian, Mark Riedl, and Brent Harrison. 2019. Learning norms from stories: A prior for value aligned agents. *arXiv preprint arXiv:1912.03553* (2019).

[90] Victor Gallego, Roi Naveiro, and David Rios Insua. 2019. Reinforcement learning under threats. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 9939–9940.

[91] Yang Gao and Steve Chien. 2017. Review on space robotics: Toward top-level science through space exploration. *Science Robotics* 2, 7 (2017). https://doi.org/10.1126/scirobotics.aan5074

[92] Javier Garcıa and Fernando Fernández. 2015. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research* 16, 1 (2015), 1437–1480.

[93] David García-Soriano and Francesco Bonchi. 2021. Maxmin-fair ranking: individual fairness under group-fairness constraints. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 436–446.

[94] Sahaj Garg, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed H Chi, and Alex Beutel. 2019. Counterfactual fairness in text classification through robustness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 219–226.

[95] Azin Ghazimatin, Matthaus Kleindessner, Chris Russell, Ziawasch Abedjan, and Jacek Golebiowski. 2022. Measuring fairness of rankings under noisy sensitive information. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 2263–2279.

[96] Robert Givan, Thomas Dean, and Matthew Greig. 2003. Equivalence notions and model minimization in Markov decision processes. *Artificial Intelligence* 147, 1-2 (2003), 163–223.

[97] Bruce Glymour and Jonathan Herington. 2019. Measuring the biases that matter: The ethical and casual foundations for measures of fairness in algorithms. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 269–278.

[98] Noah J Goodall. 2014. Machine ethics and automated vehicles. In *Road vehicle automation*. Springer, 93–102.

[99] Michael A. Goodrich, Bryan S. Morse, Damon Gerhardt, Joseph L. Cooper, Morgan Quigley, Julie A. Adams, and Curtis Humphrey. 2008. Supporting wilderness search and rescue using a camera-equipped mini UAV. *Journal of Field Robotics* (2008).

[100] Christopher Grau. 2006. There is no "I" in "robot": Robots and utilitarianism. *IEEE Intelligent Systems* 21, 4 (2006), 52–55.

[101] Ben Green and Salomé Viljoen. 2020. Algorithmic realism: expanding the boundaries of algorithmic thought. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 19–31.

[102] Shangding Gu, Long Yang, Yali Du, Guang Chen, Florian Walter, Jun Wang, Yaodong Yang, and Alois Knoll. 2022. A review of safe reinforcement learning: Methods, theory and applications. *arXiv preprint arXiv:2205.10330* (2022).

[103] Carlos Guestrin, Daphne Koller, Ronald Parr, and Shobha Venkataraman. 2003. Efficient solution algorithms for factored MDPs. *Journal of Artificial Intelligence Research* 19 (2003), 399–468.

[104] Carlos Guestrin, Daphne Koller, Ronald Parr, and Shobha Venkataraman. 2011. Efficient solution algorithms for factored MDPs. *arXiv e-prints* (2011), arXiv–1106.

[105] Carlos E Guestrin and Geoffrey Gordon. 2012. Distributed planning in hierarchical factored MDPs. *arXiv preprint arXiv:1301.0571* (2012).

[106] Maya Gupta, Andrew Cotter, Mahdi Milani Fard, and Serena Wang. 2018. Proxy fairness. *arXiv preprint arXiv:1806.11212* (2018).

[107] Swati Gupta and Vijay Kamble. 2021. Individual Fairness in Hindsight. *J. Mach. Learn. Res.* 22, 144 (2021), 1–35.

[108] Jacqueline Hannan, Huei-Yen Winnie Chen, and Kenneth Joseph. 2021. Who Gets What, According to Whom? An Analysis of Fairness Perceptions in Service Allocation. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 555–565.

[109] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems* 29 (2016).

[110] Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. 2018. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*. PMLR, 1929–1938.

[111] Milos Hauskrecht, Nicolas Meuleau, Leslie Pack Kaelbling, Thomas L Dean, and Craig Boutilier. 2013. Hierarchical solution of Markov decision processes using macro-actions. *arXiv preprint arXiv:1301.7381* (2013).

[112] Gillian R Hayes. 2014. Knowing by doing: action research as an approach to HCI. In *Ways of Knowing in HCI*. Springer, 49–68.

[113] Hoda Heidari and Jon Kleinberg. 2021. Allocating opportunities in a dynamic model of intergenerational Mobility. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*. 15–25.

[114] Hoda Heidari, Michele Loi, Krishna P. Gummadi, and Andreas Krause. 2019. A Moral Framework for Understanding Fair ML through Economic Models of Equality of Opportunity. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 181–190.

[115] Bill Hibbard. 2012. Avoiding unintended AI behaviors. In *International Conference on Artificial General Intelligence*. Springer, 107–116.

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

[117] John N Hooker and Tae Wan N Kim. 2018. Toward non-intuition-based machine and artificial intelligence ethics: A deontological approach based on modal logic. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 130–136.

[118] Lily Hu and Yiling Chen. 2020. Fair classification and social welfare. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 535–545.

[119] Lily Hu, Nicole Immorlica, and Jennifer Wortman Vaughan. 2019. The disparate effects of strategic manipulation. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 259–268.

[120] Ben Hutchinson and Margaret Mitchell. 2019. 50 years of test (un)fairness: Lessons for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 49–58.

[121] Vasileios Iosifidis and Eirini Ntoutsi. 2019. Adafair: Cumulative fairness adaptive boosting. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 781–790.

[122] Lilly C Irani and M Six Silberman. 2013. Turkopticon: Interrupting worker invisibility in Amazon Mechanical Turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 611–620.

[123] Abigail Z. Jacobs and Hanna Wallach. 2021. Measurement and Fairness. In *Proceedings of the Conference on Fairness, Accountability and Transparency*.

[124] Heinrich Jiang and Ofir Nachum. 2020. Identifying and correcting label bias in machine learning. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 702–712.

[125] Ray Jiang, Aldo Pacchiano, Tom Stepleton, Heinrich Jiang, and Silvia Chiappa. 2020. Wasserstein fair classification. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*. 862–872.

[126] Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. 2016. Fair algorithms for infinite and contextual bandits. *arXiv preprint arXiv:1610.09559* (2016).

[127] Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. 2018. Meritocratic fairness for infinite and contextual bandits. In *2018 AAAI/ACM Conference on AI, Ethics, and Society*. 158–163.

[128] Matthew Joseph, Michael Kearns, Jamie H Morgenstern, and Aaron Roth. 2016. Fairness in learning: Classic and contextual bandits. *Advances in Neural Information Processing Systems* 29 (2016).

[129] Zoe Juozapaitis, Anurag Koul, Alan Fern, Martin Erwig, and Finale Doshi-Velez. 2019. Explainable reinforcement learning via reward decomposition. In *Proceedings of the IJCAI/ECAI Workshop on Explainable Artificial Intelligence*.

[130] Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. 1998. Planning and acting in partially observable stochastic domains. *Artificial Intelligence* 101, 1-2 (1998), 99–134.

[131] Nathan Kallus and Angela Zhou. 2021. Fairness, welfare, and equity in personalized pricing. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*. 296–314.

[132] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2018. Recommendation independence. In *Proceedings of the Conference on Fairness, Accountability and Transparency*. 187–201.

[133] Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. 2011. Fairness-aware learning through regularization approach. In *2011 IEEE 11th International Conference on Data Mining Workshops*. IEEE, 643–650.

[134] Daniel Kasenberg and Matthias Scheutz. 2018. Norm conflict resolution in stochastic domains. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[135] Atoosa Kasirzadeh. 2022. Algorithmic Fairness and Structural Injustice: Insights from Feminist Political Philosophy. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. 349–356.

[136] Atoosa Kasirzadeh and Andrew Smart. 2021. The use and misuse of counterfactuals in ethical machine learning. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*. 228–236.

[137] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International conference on machine learning*. PMLR, 2564–2572.

[138] Simon Keizer, Mary Ellen Foster, Oliver Lemon, Andre Gaschler, and Manuel Giuliani. 2013. Training and evaluation of an MDP model for social multi-user human-robot interaction. In *Proceedings of the 14th SIGdial Meeting on Discourse and Dialogue*.

[139] Os Keyes. 2018. The misgendering machines: Trans/HCI implications of automatic gender recognition. *Proceedings of the ACM Conference on Human-Computer Interaction* 2 (2018).

[140] Utsab Khakurel, Ghada Abdelmoumin, Aakriti Bajracharya, and Danda B Rawat. 2022. Exploring bias and fairness in artificial intelligence and machine learning algorithms. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications IV*, Vol. 12113. SPIE, 629–638.

[141] Fereshte Khani and Percy Liang. 2021. Removing spurious features can hurt accuracy and affect groups disproportionately. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 196–205.

[142] Hyemi Kim, Seungjae Shin, JoonHo Jang, Kyungwoo Song, Weonyoung Joo, Wanmo Kang, and Il-Chul Moon. 2021. Counterfactual fairness with disentangled causal effect variational autoencoder. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 8128–8136.

[143] Richard Kim, Max Kleiman-Weiner, Andrés Abeliuk, Edmond Awad, Sohan Dsouza, Joshua B Tenenbaum, and Iyad Rahwan. 2018. A computational model of commonsense moral decision making. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 197–203.

[144] Serge-Christophe Kolm. 1993. The impossibility of utilitarianism. In *The good and the economical*. Springer, 30–69.
[145] Andrey Kolobov and Daniel Weld. 2012. A theory of goal-oriented MDPs with dead ends. *arXiv preprint arXiv:1210.4875* (2012).
[146] Öykü Deniz Köse and Yanning Shen. 2021. Fairness-aware node representation learning. *arXiv preprint arXiv:2106.05391* (2021).
[147] Emmanouil Krasanakis, Eleftherios Spyromitros-Xioufis, Symeon Papadopoulos, and Yiannis Kompatsiaris. 2018. Adaptive sensitive reweighting to mitigate bias in fairness-aware classification. In *Proceedings of the 2018 world wide web conference*. 853–862.
[148] Piotr Kulicki, Michael P Musielewicz, and Robert Trypuz. 2019. Virtue ethics for autonomous cars. *ResearchGate Preprint* (2019).
[149] Bogdan Kulynych, David Madras, Smitha Milli, Inioluwa Deborah Raji, Angela Zhou, and Richard Zemel. 2020. Participatory Approaches to Machine Learning. Proceedings of the International Conference on Machine Learning Workshop.
[150] Matt J Kusner, Joshua R Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. *arXiv preprint arXiv:1703.06856* (2017).
[151] Bruno Lacerda, Fatma Faruq, David Parker, and Nick Hawes. 2019. Probabilistic planning with formal performance guarantees for mobile service robots. *The International Journal of Robotics Research* 38, 9 (2019), 1098–1123.
[152] Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed Chi. 2020. Fairness without demographics through adversarially reweighted learning. *Advances in neural information processing systems* 33 (2020), 728–740.
[153] Chi-Pang Lam and S Shankar Sastry. 2014. A POMDP framework for human-in-the-loop system. In *Proceedings of the 53rd IEEE Conference on Decision and Control*. IEEE, 6031–6036.
[154] Mikko Lauri, David Hsu, and Joni Pajarinen. 2022. Partially Observable Markov Decision Processes in Robotics: A Survey. *IEEE Transactions on Robotics* (2022).
[155] Susan Leavy, Eugenia Siapera, and Barry O'Sullivan. 2021. Ethical data curation for AI: An approach based on feminist epistemology and critical theories of race. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 695–703.
[156] Derek Leben. 2017. A Rawlsian algorithm for autonomous vehicles. *Ethics and Information Technology* 19, 2 (2017), 107–115.
[157] Fengjiao Li, Jia Liu, and Bo Ji. 2019. Combinatorial sleeping bandits with fairness constraints. *IEEE Transactions on Network Science and Engineering* 7, 3 (2019), 1799–1813.
[158] Lihong Li, Thomas J Walsh, and Michael L Littman. 2006. Towards a unified theory of state abstraction for MDPs. In *Proceedings of the International Symposium on Artificial Intelligence and Mathematics*.
[159] Q Vera Liao and Michael Muller. 2019. Enabling value sensitive AI systems through participatory design fictions. *arXiv preprint arXiv:1912.07381* (2019).
[160] Michael L Littman, Thomas L Dean, and Leslie Pack Kaelbling. 2013. On the complexity of solving Markov decision problems. *arXiv preprint arXiv:1302.4971* (2013).
[161] J Liu. 2017. Confucian robotic ethics. In *International Conference on the Relevance of the Classics under the Conditions of Modernity: Humanity and Science*.
[162] Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. 2018. Delayed impact of fair machine learning. In *International Conference on Machine Learning*. PMLR, 3150–3158.
[163] Lydia T Liu, Ashia Wilson, Nika Haghtalab, Adam Tauman Kalai, Christian Borgs, and Jennifer Chayes. 2020. The disparate equilibria of algorithmic decision making when individuals invest rationally. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 381–391.
[164] Yang Liu, Goran Radanovic, Christos Dimitrakakis, Debmalya Mandal, and David C Parkes. 2017. Calibrated fairness in bandits. *arXiv preprint arXiv:1707.01875* (2017).
[165] Francesco Locatello, Gabriele Abbati, Thomas Rainforth, Stefan Bauer, Bernhard Schölkopf, and Olivier Bachem. 2019. On the fairness of disentangled representations. *Advances in neural information processing systems* 32 (2019).
[166] Laura Londoño, Juana Valeria Hurtado, Nora Hertz, Philipp Kellmeyer, Silja Voeneky, and Abhinav Valada. 2022. Fairness and Bias in Robot Learning. *arXiv preprint arXiv:2207.03444* (2022).
[167] Kristian Lum, Chesa Boudin, and Megan Price. 2020. The impact of overbooking on a pre-trial risk assessment tool. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 482–491.
[168] Michael A Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. Co-designing checklists to understand organizational challenges and opportunities around fairness in AI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.
[169] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. 2018. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*. PMLR, 3384–3393.
[170] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. 2019. Fairness through causal awareness: Learning causal latent-variable models for biased data. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 349–358.
[171] Alan Malek, Yasin Abbasi-Yadkori, and Peter Bartlett. 2014. Linear programming for large-scale Markov decision problems. In *Proceedings of the International Conference on Machine Learning*. 496–504.
[172] Vahideh Manshadi, Rad Niazadeh, and Scott Rodilitz. 2021. Fair dynamic rationing. In *Proceedings of the 22nd ACM Conference on Economics and Computation*. 694–695.
[173] Frank Marcinkowski, Kimon Kieslich, Christopher Starke, and Marco Lünich. 2020. Implications of AI (un-)fairness in higher education admissions: The effects of perceived AI (un-)fairness on exit, voice and organizational reputation. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 122–130.
[174] Donald Martin Jr, Vinodkumar Prabhakaran, Jill Kuhlberg, Andrew Smart, and William S Isaac. 2020. Participatory problem formulation for fairer machine learning through community based system dynamics. *arXiv preprint arXiv:2005.07572* (2020).
[175] Blossom Metevier, Stephen Giguere, Sarah Brockman, Ari Kobren, Yuriy Brun, Emma Brunskill, and Philip S Thomas. 2019. Offline contextual bandits with high probability fairness guarantees. *Neural Information Processing Systems* 32 (2019).
[176] Curtis R Michael. 2019. *Ethical considerations for the use of lethal autonomous weapons systems*. Technical Report. US Army Command and General Staff College Fort Leavenworth United States.
[177] Smitha Milli, John Miller, Anca D Dragan, and Moritz Hardt. 2019. The social cost of strategic classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 230–239.
[178] Shuwa Miura and Shlomo Zilberstein. 2020. Maximizing plan legibility in stochastic environments. In *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems*. 1931–1933.
[179] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. 2015. Human-level control through deep reinforcement learning. *Nature* 518, 7540 (2015), 529–533.
[180] Shakir Mohamed, Marie-Therese Png, and William Isaac. 2020. Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence. *Philosophy & Technology* 33, 4 (2020), 659–684.
[181] Teodor Mihai Moldovan and Pieter Abbeel. 2012. Safe exploration in Markov decision processes. *arXiv preprint arXiv:1205.4810* (2012).
[182] Hussein Mouzannar, Mesrob I Ohannessian, and Nathan Srebro. 2019. From fair decision making to social equality. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 359–368.
[183] John F. Mustard, D. Beaty, and D. Bass. 2013. Mars 2020 science rover: Science goals and mission concept. In *Proceedings of the AAS/Division for Planetary Sciences Meeting Abstracts*, Vol. 45.
[184] Vedant Nanda, Samuel Dooley, Sahil Singla, Soheil Feizi, and John P Dickerson. 2021. Fairness through robustness: Investigating robustness disparity in deep learning. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 466–477.
[185] Harikrishna Narasimhan, Andrew Cotter, Maya Gupta, and Serena Wang. 2020. Pairwise fairness for ranking and regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 5248–5255.
[186] Samer Nashed, Justin Svegliato, and Shlomo Zilberstein. 2021. Ethically compliant planning within moral communities. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 188–198.
[187] Samer B Nashed, Saaduddin Mahmud, Claudia V Goldman, and Shlomo Zilberstein. 2023. Causal Explanation for Sequential Decision Making Under Uncertainty. *arXiv preprint arXiv:2205.15462* (2023).
[188] Samer B Nashed, Justin Svegliato, Abhinav Bhatia, Stuart Russell, and Shlomo Zilberstein. 2022. Selecting the Partial State Abstractions of MDPs: A Metareasoning Approach with Deep Reinforcement Learning. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*.
[189] Samer B Nashed, Justin Svegliato, Matteo Brucato, Connor Basich, Rod Grupen, and Shlomo Zilberstein. 2021. Solving Markov decision processes with partial state abstractions. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 813–819.
[190] Mitchell J Neubert and George D Montañez. 2020. Virtue as a framework for the design and use of artificial intelligence. *Business Horizons* 63, 2 (2020), 195–204.
[191] Andrew Y Ng, Daishi Harada, and Stuart Russell. 1999. Policy invariance under reward transformations: Theory and application to reward shaping. In *Proceedings of the International Conference on Machine Learning*, Vol. 99. 278–287.
[192] Arnab Nilim and Laurent El Ghaoui. 2005. Robust control of Markov decision processes with uncertain transition matrices. *Operations Research* 53, 5 (2005), 780–798.

[193] Safiya Umoja Noble. 2018. *Algorithms of Oppression*. New York University Press.

[194] Alejandro Noriega-Campero, Bernardo Garcia-Bulle, Luis Fernando Cantu, Michiel A Bakker, Luis Tejerina, and Alex Pentland. 2020. Algorithmic targeting of social policies: Fairness, accuracy, and distributed governance. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 241–251.

[195] Sungho Park, Sunhee Hwang, Dohyung Kim, and Hyeran Byun. 2021. Learning disentangled representation for fair facial attribute classification via fairness-aware information alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 2403–2411.

[196] Ioannis Pastaltzidis, Nikolaos Dimitriou, Katherine Quezada-Tavarez, Stergios Aidinlis, Thomas Marquenie, Agata Gurzawska, and Dimitrios Tzovaras. 2022. Data augmentation for fairness-aware machine learning: Preventing algorithmic bias in law enforcement systems. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 2302–2314.

[197] Vishakha Patil, Ganesh Ghalme, Vineet Nair, and Y Narahari. 2021. Achieving fairness in the stochastic multi-armed bandit problem. *Journal of Machine Learning Research* 22 (2021), 174–1.

[198] Anco Peeters and Pim Haselager. 2021. Designing virtuous sex robots. *International Journal of Social Robotics* 13, 1 (2021), 55–66.

[199] Marek Petrik and Ronny Luss. 2016. Interpretable policies for dynamic product recommendations. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*.

[200] Marek Petrik and Shlomo Zilberstein. 2009. Constraint relaxation in approximate linear programs. In *Proceedings of the International Conference on Machine Learning*. 809–816.

[201] Luis Pineda, Takeshi Takahashi, Hee-Tae Jung, Shlomo Zilberstein, and Roderic Grupen. 2015. Continual planning for search and rescue robots. In *Proceedings of the IEEE/RAS International Conference on Humanoid Robots*. IEEE, 243–248.

[202] Luis Pineda, Kyle Hollins Wray, and Shlomo Zilberstein. 2017. Fast SSP solvers using short-sighted labeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

[203] Pascal Poupart, Aarti Malhotra, Pei Pei, Kee-Eung Kim, Bongseok Goh, and Michael Bowling. 2015. Approximate linear programming for constrained partially observable Markov decision processes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 1. 3342–3348.

[204] Warren B Powell. 2016. Perspectives of approximate dynamic programming. *Annals of Operations Research* 241, 1-2 (2016), 319–356.

[205] T Powers. 2005. Deontological machine ethics. In *2005 AAAI Fall Symposium on Machine Ethics*. 79–86.

[206] Thomas M Powers. 2006. Prospects for a Kantian machine. *IEEE Intelligent Systems* 21, 4 (2006), 46–51.

[207] David Pujol, Ryan McKenna, Satya Kuppam, Michael Hay, Ashwin Machanavajjhala, and Gerome Miklau. 2020. Fair decision making using privacy-protected data. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 189–199.

[208] Aida Rahmattalabi, Phebe Vayanos, Kathryn Dullerud, and Eric Rice. 2022. Learning Resource Allocation Policies from Observational Data with an Application to Homeless Services Delivery. *arXiv preprint arXiv:2201.10053* (2022).

[209] Inioluwa Deborah Raji, I. Elizabeth Kumar, Aaron Horowitz, and Andrew Selbst. 2022. The Fallacy of AI Functionality. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 959–972.

[210] Miriam Rateike, Ayan Majumdar, Olga Mineeva, Krishna P Gummadi, and Isabel Valera. 2022. Don't Throw it Away! The Utility of Unlabeled Data in Fair Decision Making. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 1421–1433.

[211] Balaraman Ravindran and Andrew G Barto. 2002. Model minimization in hierarchical reinforcement learning. In *Proceedings of the International Symposium on Abstraction, Reformulation, and Approximation*. Springer, 196–211.

[212] Lydia Reader, Pegah Nokhiz, Cathleen Power, Neal Patwari, Suresh Venkatasubramanian, and Sorelle Friedler. 2022. Models for understanding and quantifying feedback in societal systems. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 1765–1775.

[213] Kit T Rodolfa, Erika Salomon, Lauren Haynes, Iván Higuera Mendieta, Jamie Larson, and Rayid Ghani. 2020. Case study: Predictive fairness to reduce misdemeanor recidivism through social service interventions. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 142–153.

[214] Andrzej Ruszczyński. 2010. Risk-averse dynamic programming for Markov decision processes. *Mathematical programming* 125, 2 (2010), 235–261.

[215] Sandhya Saisubramanian, Ece Kamar, and Shlomo Zilberstein. 2021. A multi-objective approach to mitigate negative side effects. In *Proceedings of the 29th International Conference on International Joint Conferences on Artificial Intelligence*. 354–361.

[216] Sandhya Saisubramanian and Shlomo Zilberstein. 2019. Adaptive outcome selection for planning with reduced models. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 1655–1660.

[217] Javier Sánchez-Monedero, Lina Dencik, and Lilian Edwards. 2020. What does it mean to 'solve' the problem of discrimination in hiring? Social, technical and legal perspectives from the UK on automated hiring systems. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 458–468.

[218] Andrew J Schaefer, Matthew D Bailey, Steven M Shechter, and Mark S Roberts. 2005. Modeling medical treatment using Markov decision processes. In *Operations research and health care*. Springer, 593–612.

[219] Philipp Schillinger, Sergio García, Alexandros Makris, Konstantinos Roditakis, Michalis Logothetis, Konstantinos Alevizos, Wei Ren, Pouria Tajvar, Patrizio Pelliccione, Antonis Argyros, et al. 2021. Adaptive heterogeneous multi-robot collaboration from formal task specifications. *Robotics and Autonomous Systems* 145 (2021), 103866.

[220] Candice Schumann, Zhi Lang, Nicholas Mattei, and John P Dickerson. 2019. Group fairness in bandit arm selection. *arXiv preprint arXiv:1912.03802* (2019).

[221] Pola Schwöbel and Peter Remmers. 2022. The Long Arc of Fairness: Formalisations and Ethical Discourse. *arXiv preprint arXiv:2203.06038* (2022).

[222] Russ Shafer-Landau. 2009. *The fundamentals of ethics*. Oxford University Press.

[223] Shubham Sharma, Yunfeng Zhang, Jesús M Ríos Aliaga, Djallel Bouneffouf, Vinod Muthusamy, and Kush R Varshney. 2020. Data augmentation for discrimination prevention and bias disambiguation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 358–364.

[224] Xinyue Shen, Steven Diamond, Yuantao Gu, and Stephen Boyd. 2016. Disciplined convex-concave programming. In *2016 IEEE 55th Conference on Decision and Control (CDC)*. IEEE, 1009–1014.

[225] Jaeeun Shim and Ronald Arkin. 2017. An intervening ethical governor for a robot mediator in patient-caregiver relationships. In *A World with Robots*. Springer, 77–91.

[226] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. 2018. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science* 362, 6419 (2018), 1140–1144.

[227] Harvineet Singh, Rina Singh, Vishwali Mhasawade, and Rumi Chunara. 2021. Fairness violations and mitigation under covariate shift. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*. 3–13.

[228] Dylan Slack, Sorelle A Friedler, and Emile Givental. 2020. Fairness warnings and Fair-MAML: Learning fairly with minimal data. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 200–209.

[229] Trevor Smith, Yuhao Chen, Nathan Hewitt, Boyi Hu, and Yu Gu. 2021. Socially aware robot obstacle avoidance considering human intention and preferences. *International journal of social robotics* (2021), 1–18.

[230] Trey Smith and Reid Simmons. 2006. Focused real-time dynamic programming for MDPs: Squeezing more out of a heuristic. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 1227–1232.

[231] Till Speicher, Muhammad Ali, Giridhari Venkatadri, Filipe Nunes Ribeiro, George Arvanitakis, Fabrício Benevenuto, Krishna P Gummadi, Patrick Loiseau, and Alan Mislove. 2018. Potential for discrimination in online targeted advertising. In *Proceedings of the Conference on Fairness, Accountability and Transparency*. 5–19.

[232] Luke Stark and Jevan Hutson. 2022. Physiognomic artificial intelligence. *Fordham Intellectual Property, Media & Entertainment Law Journal XXXII* (2022).

[233] Jack Stilgoe. 2018. We need new rules for self-driving cars. *Issues in Science and Technology* 34, 3 (2018), 52–57.

[234] Alejandro Suárez-Hernández, Carme Torras, and Guillem Alenya. 2019. Practical resolution methods for mdps in robotics exemplified with disassembly planning. *IEEE Robotics and Automation Letters* 4, 3 (2019), 2282–2288.

[235] Richard S. Sutton. 1995. Learning to predict by the methods of temporal differences. *Machine Learning* 3, 1 (1995), 9–44.

[236] Justin Svegliato, Samer B Nashed, and Shlomo Zilberstein. 2021. Ethically compliant sequential decision making. In *Proceedings of the 35th Conference on Artificial Intelligence*.

[237] Chris Sweeney and Maryam Najafian. 2020. Reducing sentiment polarity for demographic attributes in word embeddings using adversarial learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 359–368.

[238] Latanya Sweeney. 2013. Discrimination in online ad delivery. *Commun. ACM* 56, 5 (2013), 44–54.

[239] Bruce A Swett, Erin N Hahn, and Ashley J Llorens. 2021. Designing robots for the battlefield: State of the art.

[240] Guanhong Tao, Weisong Sun, Tingxu Han, Chunrong Fang, and Xiangyu Zhang. 2022. RULER: discriminative and iterative adversarial training for deep neural network fairness. In *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 1173–1184.

[241] Bahar Taskesen, Jose Blanchet, Daniel Kuhn, and Viet Anh Nguyen. 2021. A statistical test for probabilistic fairness. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*. 648–665.

[242] Philip Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. 2015. High confidence policy improvement. In *Proceedings of the International Conference on Machine Learning*. 2380–2388.

[243] Nicholay Topin, Stephanie Milani, Fei Fang, and Manuela Veloso. 2021. Iterative bounding MDPs: Learning interpretable policies via non-interpretable methods. *arXiv preprint arXiv:2102.13045* (2021).

[244] Yao-Hung Hubert Tsai, Martin Q Ma, Han Zhao, Kun Zhang, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2021. Conditional contrastive learning: Removing undesirable information in self-supervised representations. *arXiv e-prints* (2021), arXiv–2106.

[245] Berk Ustun, Alexander Spangher, and Yang Liu. 2019. Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 10–19.

[246] Leendert van der Torre. 2003. Contextual deontic logic: Normative agents, violations and independence. *Annals of mathematics and artificial intelligence* 37, 1-2 (2003), 33–63.

[247] Tijs Vandemeulebroucke, Bernadette Dierckx de Casterlé, and Chris Gastmans. 2018. The use of care robots in aged care: A systematic review of argument-based ethics literature. *Archives of gerontology and geriatrics* 74 (2018), 15–25.

[248] Dieter Vanderelst and Alan Winfield. 2018. An architecture for ethical robots inspired by the simulation theory of cognition. *Cognitive Systems Research* 48 (2018), 56–66.

[249] Lingraj S Vannur, Balaji Ganesan, Lokesh Nagalapatti, Hima Patel, and MN Tippeswamy. 2021. Data augmentation for fairness in personal knowledge base population. In *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining*. 143–152.

[250] Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. 2019. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature* 575, 7782 (2019), 350–354.

[251] Karel Vredenburg, Ji-Ye Mao, Paul W Smith, and Tom Carey. 2002. A survey of user-centered design practice. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 471–478.

[252] Akifumi Wachi and Yanan Sui. 2020. Safe reinforcement learning in constrained Markov decision processes. In *Proceedings of the International Conference on Machine Learning*. 9797–9806.

[253] Christina Wadsworth, Francesca Vera, and Chris Piech. 2018. Achieving fairness through adversarial learning: An application to recidivism prediction. *arXiv preprint arXiv:1807.00199* (2018).

[254] Mingyang Wan, Daochen Zha, Ninghao Liu, and Na Zou. 2023. In-processing modeling techniques for machine learning fairness: A survey. *ACM Transactions on Knowledge Discovery from Data* 17, 3 (2023), 1–27.

[255] Jialu Wang, Yang Liu, and Caleb Levy. 2021. Fair classification with group-dependent label noise. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*. 526–536.

[256] Christopher Watkins and Peter Dayan. 1992. Q-learning. *Machine Learning* 8, 3-4 (1992), 279–292.

[257] Ralph Wedgwood and MC Timmons. 2022. The reasons aggregation theorem. *Oxford Studies in Normative Ethics Volume 12* (2022), 127.

[258] Min Wen, Osbert Bastani, and Ufuk Topcu. 2021. Algorithms for fairness in sequential decision making. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 1144–1152.

[259] Vincent Wiegel and Jan van den Berg. 2009. Combining moral theory, modal logic and MAS to create well-behaving artificial agents. *International Journal of Social Robotics* 1, 3 (2009), 233–242.

[260] Christo Wilson, Avijit Ghosh, Shan Jiang, Alan Mislove, Lewis Baker, Janelle Szary, Kelly Trindel, and Frida Polli. 2021. Building and auditing fair algorithms: A case study in candidate screening. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 666–677.

[261] Alan Winfield, Christian Blum, and Wenguo Liu. 2014. Towards an ethical robot: Internal models, consequences and ethical action selection. In *Proceedings of the Conference Towards Autonomous Robotic Systems*. Springer, 85–96.

[262] Kyle Hollins Wray, Luis Pineda, and Shlomo Zilberstein. 2016. Hierarchical approach to transfer of control in semi-autonomous systems. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*.

[263] Kyle Hollins Wray, Stefan J. Witwicki, and Shlomo Zilberstein. 2017. Online decision-making for scalable autonomous systems. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*.

[264] Ziwei Wu and Jingrui He. 2022. Fairness-aware Model-agnostic Positive and Unlabeled Learning. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 1698–1708.

[265] Vahid Yazdanpanah, Enrico H Gerding, Sebastian Stein, Mehdi Dastani, Catholijn M Jonker, Timothy J Norman, and Sarvapali D Ramchurn. 2022. Reasoning about responsibility in autonomous systems: challenges and opportunities. *AI & SOCIETY* (2022), 1–12.

[266] Gary Chan Kok Yew. 2021. Trust in and ethical design of carebots: the case for ethics of care. *International Journal of Social Robotics* 13, 4 (2021), 629–645.

[267] Sung Wook Yoon, Alan Fern, and Robert Givan. 2007. FF-Replan: A baseline for probabilistic planning. In *Proceedings of the International Conference on Automated Planning and Scheduling*, Vol. 7. 352–359.

[268] Huizhen Yu and Dimitri P Bertsekas. 2009. Basis function adaptation methods for cost approximation in MDPs. In *Proceedings of the IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning*. IEEE, 74–81.

[269] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P Gummadi. 2019. Fairness constraints: A flexible approach for fair classification. *The Journal of Machine Learning Research* 20, 1 (2019), 2737–2778.

[270] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. 2017. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*. 962–970.

[271] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 335–340.

[272] Dell Zhang and Jun Wang. 2021. Recommendation Fairness: From Static to Dynamic. *arXiv preprint arXiv:2109.03150* (2021).

[273] Hongjing Zhang and Ian Davidson. 2021. Towards fair deep anomaly detection. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*. 138–148.

[274] Xueru Zhang and Mingyan Liu. 2021. Fairness in learning-based sequential decision algorithms: A survey. In *Handbook of Reinforcement Learning and Control*. Springer, 525–555.

[275] Yi Zhang and Jitao Sang. 2020. Towards accuracy-fairness Paradox: Adversarial example-based data augmentation for visual debiasing. In *Proceedings of the 28th ACM International Conference on Multimedia*. 4346–4354.

[276] Han Zhao, Amanda Coston, Tameem Adel, and Geoffrey J Gordon. 2019. Conditional learning of fair representations. *arXiv preprint arXiv:1910.07162* (2019).

[277] Chang Zhou, Jianxin Ma, Jianwei Zhang, Jingren Zhou, and Hongxia Yang. 2021. Contrastive learning for debiased candidate generation in large-scale recommender systems. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 3985–3995.

[278] Qin Zhu, Tom Williams, and Ruchen Wen. 2019. Confucian robot ethics. *Computer Ethics-Philosophical Enquiry (CEPE) Proceedings* 2019, 1 (2019), 12.