Ethically Compliant Sequential Decision Making

Justin Svegliato **Samer B. Nashed** **Shlomo Zilberstein**

College of Information and Computer Sciences
University of Massachusetts Amherst
{jsvegliato,snashed,shlomo}@cs.umass.edu

Abstract

Enabling autonomous systems to comply with an ethical theory is critical given their accelerating deployment in domains that impact society. While many ethical theories have been studied extensively in moral philosophy, they are still challenging to implement by developers who build autonomous systems. This paper proposes a novel approach for building *ethically compliant autonomous systems* that optimize completing a task while following an ethical framework. First, we introduce a definition of an ethically compliant autonomous system and its properties. Next, we offer a range of ethical frameworks for divine command theory, prima facie duties, and virtue ethics. Finally, we demonstrate the accuracy and usability of our approach in a set of autonomous driving simulations and a user study of planning and robotics experts.

Introduction

Enabling autonomous systems to comply with an ethical theory is critical given their accelerating deployment in domains that impact society (Charisi et al. 2017). For example, a self-driving car that drives a route ought to slow near a school zone, crosswalk, or park to avoid endangering pedestrians (Svegliato et al. 2019; Basich et al. 2020). Similarly, an elder care robot that helps caregivers perform medical diagnostics ought to tailor its support based on the physical and mental state of the patient to reduce the risk of injury and the loss of dignity (Shim, Arkin, and Pettinatti 2017). While many ethical theories have been studied extensively in moral philosophy, they are still challenging to implement by developers who build autonomous systems. Hence, there is a growing need to simplify and standardize the process of implementing an ethical theory within autonomous systems.

A simple approach to enabling an autonomous system to comply with an ethical theory is to modify its objective function directly. Modifying this objective function, however, poses two problems. First, adjusting the objective function can lead to unpredictable effects on the behavior of the autonomous system due to the complexity of its decision-making model. In fact, small changes to the objective function can generate large changes to the behavior of the autonomous system (Bostrom 2016). Second, using the objective function to represent both a task and an ethical theory can result in incommensurable conversions as it blends them within the decision-making model implicitly (Taylor et al. 2016). These problems cause the behavior of an autonomous system to fail to reflect the intentions of developers or the values of stakeholders (Hadfield-Menell and Hadfield 2019).

Ideally, any developer who builds an autonomous system with the ability to comply with an ethical theory could instead use an approach that exhibits several desirable properties. First, it should be *general-purpose* by supporting any task or ethical theory as long as they can be represented appropriately. Next, it should be *modular* by encapsulating the task and ethical theory as separate modules that avoid an objective function that blends them implicitly. Finally, it should be *interpretable* by describing the ethical theory in terms of the behavior and environment of the autonomous system.

We propose a novel approach with these properties for building *ethically compliant autonomous systems* that optimize completing a task subject to following an ethical framework. As expected, the task defines the goal that the system must achieve using a *decision-making model*. More importantly, the ethical framework approximates a well-known ethical theory that the system must comply with using a *moral principle* and an *ethical context*. The moral principle evaluates whether or not the system violates the ethical framework and the ethical context includes the contextual information needed to evaluate the system. Formally, this is expressed as an optimization problem with a set of constraints for the task and a constraint for the ethical framework. While our approach supports different decision-making models and ethical frameworks, we consider a Markov decision process as the decision-making model and divine command theory, prima facie duties, and virtue ethics as the ethical frameworks in this paper.

We evaluate our approach in two ways. In a set of autonomous driving simulations, we observe that our approach produces optimal behavior that meets a set of moral requirements. In a user study, we find that planning and robotics experts who use our approach to produce optimal behavior that meets a set of moral requirements make fewer development errors and need less development time than a simple approach that modifies the decision-making model directly.

Our contributions are: (1) a definition of an ethically compliant autonomous system and its properties, (2) a range of ethical frameworks for divine command theory, prima facie duties, and virtue ethics, and (3) a set of autonomous driving simulations and a user study of planning and robotics experts that show the accuracy and usability of our approach.

Related Work

Autonomous systems perform an array of tasks in diverse social contexts. Their potential harms can be mitigated via many strategies: (1) abandonment of technologies that are likely to be abused from a historical context (Browne 2015), such as facial recognition (Brey 2004; Introna and Wood 2004) and online surveillance (Zimmer 2008; Burgers and Robinson 2017), (2) legal intervention that enforces oversight to discourage or prevent malevolent or negligent use (Raymond and Shackelford 2013; Scherer 2015; Goodman and Flaxman 2017; Desai and Kroll 2017), including metaregulation (Pasquale 2017), and (3) technical advances that improve the accuracy and interpretability of algorithms. While these strategies are important, our approach focuses on a different strategy that reduces the likelihood for error during the design and development of autonomous systems.

Similarly, there are various principles (Friedman, Kahn, and Borning 2008; Boden et al. 2017), guidelines (Fallman 2003; Robertson et al. 2019), and standards (Read et al. 2015; Adamson, Havens, and Chatila 2019) that have recently been proposed to lower the chance for error during the design and development of autonomous systems. However, though critical to promoting the intentions of developers or the values of stakeholders, they do not address implementing an ethical theory within autonomous systems. In fact, autonomous systems that try to satisfy a set of moral requirements only through careful construction, called *implicit ethical agents*, may not produce ethical behavior (Moor 2006). Hence, many autonomous systems must be *explicit ethical agents* capable of some notion of moral reasoning (Bench-Capon and Modgil 2017; Dignum et al. 2018).

Efforts to build autonomous systems that are explicit ethical agents take two approaches (Allen, Smit, and Wallach 2005). *Bottom-up approaches* produce ethical behavior by gradually evolving or learning in an environment that rewards and penalizes behavior (Anderson, Anderson, and Berenz 2017; Shaw et al. 2018). Although this is compelling given the natural development of ethical ideas in society, they can lack stability or interpretability. Hence, *top-down approaches* produce ethical behavior by directly following prescriptive rules provided by a human or an ethical theory. These methods often use different logics, such as deontic logic (van der Torre 2003; Bringsjord, Arkoudas, and Bello 2006), temporal logic (Wooldridge and Van Der Hoek 2005; Atkinson and Bench-Capon 2006; Dennis et al. 2016), answer set programming (Berreby, Bourgne, and Ganascia 2015), or planning formalisms (Dennis et al. 2016). Some methods even use metareasoning over many logics (Bringsjord et al. 2011). While we offer a top-down approach in this paper, we do not employ logics since they are challenging to use given the growing complexity of autonomous systems (Abel, MacGlashan, and Littman 2016).

A common top-down approach that addresses the complexity of autonomous systems uses an *ethical governor* to determine online whether an action is required, permitted, or prohibited (Arkin 2008). Applications include eldercare (Shim, Arkin, and Pettinatti 2017) and physical safety (Winfield, Blum, and Liu 2014; Vanderelst and Winfield 2018). However, an ethical governor is *myopic* because it only considers the immediate reaction that must be made by the system at each time step. In contrast, our approach is *nonmyopic* since it reasons about the sequence of actions that must be performed by the system over every time step.

We know of only one other nonmyopic top-down approach to explicit ethical agents (Kasenberg and Scheutz 2018). However, the approach cannot represent different ethical theories, such as utilitarianism or Kantianism, because it is specific to norms. Moreover, the approach cannot guarantee ethical behavior since both task completion and ethical compliance are defined by real-valued weights. Our approach instead produces desirable behavior that complies with different ethical theories and avoids unpredictable trade-offs between task completion and ethical compliance.

Background

A *Markov decision process* (MDP) is a decision-making model for reasoning in fully observable, stochastic environments (Bellman 1952). An MDP can be described as a tuple $\langle S, A, T, R, d \rangle$, where S is a finite set of states, A is a finite set of actions, $T : S \times A \times S \rightarrow [0, 1]$ represents the probability of reaching a state $s' \in S$ after performing an action $a \in A$ in a state $s \in S$, $R : S \times A \times S \rightarrow \mathbb{R}$ represents the expected immediate reward of reaching a state $s' \in S$ after performing an action $a \in A$ in a state $s \in S$, and $d : S \rightarrow [0, 1]$ represents the probability of starting in a state $s \in S$. A solution to an MDP is a policy $\pi : S \rightarrow A$ indicating that an action $\pi(s) \in A$ should be performed in a state $s \in S$. A policy π induces a value function $V^\pi : S \rightarrow \mathbb{R}$ representing the expected discounted cumulative reward $V^\pi(s) \in \mathbb{R}$ for each state $s \in S$ given a discount factor $0 \leq \gamma < 1$. An optimal policy π^* maximizes the expected discounted cumulative reward for every state $s \in S$ by satisfying the Bellman optimality equation $V^*(s) = \max_{a \in A} \sum_{s' \in S} T(s, a, s')[R(s, a, s') + \gamma V^*(s')]$.

A common approach for finding an optimal policy expresses the optimization problem as a linear program in either the primal form or the dual form (Manne 1960). In this paper, we propose ethical frameworks that naturally map to the dual form. The dual form maximizes a set of occupancy measures μ^s_a for the discounted number of times an action $a \in A$ is performed in a state $s \in S$ subject to a set of constraints that maintain consistent and nonnegative occupancy.

$$\begin{aligned}
\max_{\mu} \quad & \sum_{s \in S} \sum_{a \in A} \mu^s_a \sum_{s' \in S} R(s, a, s') \\
\text{s.t.} \quad & \sum_{a' \in A} \mu^{s'}_{a'} = d(s') + \gamma \sum_{s \in S} \sum_{a \in A} T(s, a, s') \mu^s_a \quad && \forall s' \\
& \mu^s_a \geq 0 && \forall s, a
\end{aligned}$$

Ethically Compliant Autonomous Systems

We propose a novel approach for building *ethically compliant autonomous systems* that decouples ethical compliance from task completion. The system optimizes completing a task by using a *decision-making model* subject to following an ethical framework by adhering to a *moral principle* within an *ethical context*. We describe these attributes of an ethically compliant autonomous system below.

First, the system has a *decision-making model* that describes the information needed to complete the task. For example, a self-driving vehicle could have a decision-making model that includes a map of a city (Basich et al. 2021). A developer must select a representation for the decision-making model that reflects the properties of the task. For many tasks, an MDP, a decision process that assumes full observability, can be used easily. However, for more complex tasks with partial observability, start and goal states, or multiple agents, it is possible to use a decision process like a partially observable MDP, a stochastic shortest path problem, or a decentralized MDP instead. In short, the decision-making model is an *amoral*, descriptive model for completing the task but not following the ethical framework.

Next, the system has an *ethical context* that describes the information required to follow the ethical framework. For instance, an autonomous vehicle could have an ethical context that includes any details related to inconsiderate or hazardous driving that permit speeding on a highway in some scenarios but never around a school zone or near a crosswalk (Vanderelst and Winfield 2018). Similar to the decision-making model, a developer must select a representation for the ethical context that informs the fundamental principles of the ethical framework. Although the ethical context can be represented as a tuple of different values, sets, and functions, the particular specification of the tuple depends on the ethical framework. In brief, the ethical context is a *moral*, prescriptive model for following the ethical framework but not completing the task.

Finally, the system has a *moral principle* that evaluates the morality of a policy for the decision-making model within the ethical context. This considers the information that describes how to both complete the task and follow the ethical framework. As an illustration, a moral principle could require a policy to maximize the well-being of the moral community in *utilitarianism* (Bentham 1789; Mill 1895) or universalize to the moral community without contradiction in *Kantianism* (Kant and Schneewind 2002). Given a decision-making model and an ethical context, a developer must express the moral principle as a general function that maps a policy to its moral status in the following way.

Definition 1. *A **moral principle**, $\rho : \Pi \to \mathbb{B}$, represents whether a policy $\pi \in \Pi$ of a **decision-making model** $\mathcal{D}$ is moral or immoral within an **ethical context** $\mathcal{E}$.*

Now, putting these attributes together, we offer a description of an ethically compliant autonomous system below.

Definition 2. *An **ethically compliant autonomous system**, $\langle \mathcal{D}, \mathcal{E}, \rho \rangle$, optimizes completing a task by using a decision-making model $\mathcal{D}$ while following an ethical framework by adhering to a moral principle ρ within an ethical context $\mathcal{E}$.*

An ethically compliant autonomous system has the objective of finding an optimal policy that completes its task and follows its ethical framework. This can naturally be expressed as an optimization problem that solves for a policy within the space of policies that *maximizes* the value of the policy *subject to* the constraint that the policy satisfies the moral principle. We now turn to a description of the objective of an ethically compliant autonomous system below.

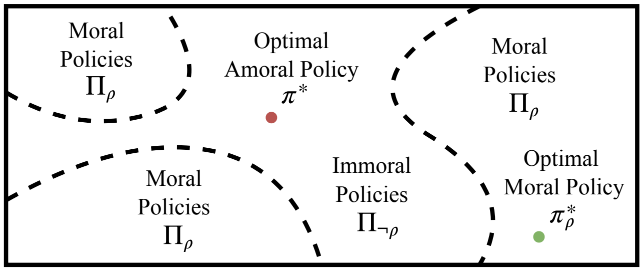

Figure 1: A simple view of the goal of an ethically compliant autonomous system (*green*) and the goal of a standard autonomous system (*red*) in terms of the space of policies.

Definition 3. *The objective of an ethically compliant autonomous system is to find an **optimal moral policy**, $\pi^*_\rho \in \Pi$, by solving for a policy $\pi \in \Pi$ within the space of policies Π that maximizes a value function V^π subject to a moral principle ρ in the following optimization problem.*

$$\begin{aligned} &\underset{\pi \in \Pi}{\text{maximize}} && V^\pi \\ &\text{subject to} && \rho(\pi) \end{aligned}$$

The objective of a standard autonomous system has typically been to find an *optimal amoral policy*, $\pi^* \in \Pi$, that only completes its task without following any ethical framework.

Figure 1 illustrates the objective of both an ethically compliant autonomous system and a standard autonomous system. For a moral principle ρ, the space of policies Π has been partitioned into a moral region Π_ρ and an immoral region $\Pi_{\neg\rho}$. The moral region Π_ρ contains the optimal moral policy $\pi^*_\rho \in \Pi_\rho$ of the ethically compliant autonomous system while the immoral region $\Pi_{\neg\rho}$ contains the optimal amoral policy $\pi^* \in \Pi_{\neg\rho}$ of the standard autonomous system. In general, the optimal amoral policy $\pi^* \in \Pi$ can be contained by either the moral region Π_ρ or the immoral region $\Pi_{\neg\rho}$.

An ethically compliant autonomous system may follow an ethical framework that negatively impacts completing its task. In this situation, a developer can evaluate the cost of this impact by calculating the maximum difference across all states between the value function of the optimal moral policy and the value function of the optimal amoral policy. We describe this idea more formally below.

Definition 4. *Given the optimal moral policy $\pi^*_\rho \in \Pi$ and the optimal amoral policy $\pi^* \in \Pi$, the **price of morality**, ψ, can be represented by the expression $\psi = \|V^{\pi^*_\rho} - V^{\pi^*}\|_\infty$.*

In fact, an ethically compliant autonomous system may even follow an ethical framework that is mutually exclusive with completing its task. In this situation, a developer should reconsider the moral implications of the system and could augment the decision-making model or adjust the ethical context if deemed safe. Intuitively, the system can be called either feasible or infeasible depending on whether or not there is a solution to the optimization problem. We express this notion more formally below.

Definition 5. *An ethically compliant autonomous system is **realizable** if there exists a policy $\pi \in \Pi$ such that its moral principle $\rho(\pi)$ is satisfied. Otherwise, it is **unrealizable**.*

Moral Constraint	Type	Conjunctions	Operations	Computations
$c_{\rho_\mathcal{F}}(\mu) = \wedge_{s \in S, a \in A, f \in F} \big(T(s,a,f)\mu^s_a = 0\big)$	Linear	$\lvert S\rvert\lvert A\rvert\lvert F\rvert$	2	$2\lvert S\rvert\lvert A\rvert\lvert F\rvert$
$c_{\rho_\Delta}(\mu) = \sum_{s \in S, a \in A} \mu^s_a \sum_{s' \in S} T(s,a,s') \sum_{\delta \in \Delta_{s'}} \phi(\delta, s') \leq \tau$	Linear	1	$3\lvert S\rvert\lvert A\rvert\lvert S\rvert\lvert \Delta\rvert + 1$	$3\lvert S\rvert\lvert A\rvert\lvert S\rvert\lvert \Delta\rvert + 1$
$c_{\rho_\mathcal{M}}(\mu) = \wedge_{s \in S, a \in A} \big(\mu^s_a \leq [\alpha(s,a)]\big)$	Linear	$\lvert S\rvert\lvert A\rvert$	$1 + 3L\lvert \mathcal{M}\rvert$	$\lvert S\rvert\lvert A\rvert\big(1 + 3L\lvert \mathcal{M}\rvert\big)$

Table 1: The moral constraints that have been derived from the moral principle of each ethical framework.

Naturally, to find the optimal moral policy by solving the optimization problem of an ethically compliant autonomous system, we use mathematical programming. This process involves four steps. First, the moral principle is mapped to a moral constraint in terms of the occupancy measures of a policy. We show that this mapping can be performed below.

Theorem 1. *A moral principle, $\rho : \Pi \rightarrow \mathbb{B}$, can be expressed as a moral constraint $c_\rho(\mu)$ in terms of the matrix of occupancy measures μ for a given policy $\pi \in \Pi$.*

Proof (Sketch) 1. *We start with a moral principle $\rho(\pi)$ using a deterministic or stochastic policy $\pi(s)$ or $\pi(a|s)$. First, recall that the discounted number of times that an action $a \in A$ is performed in a state $s \in S$ is an occupancy measure μ^s_a. Next, observe that the discounted number of times that a state $s \in S$ is visited is the expression $\sum_{a \in A} \mu^s_a$. Finally, a policy $\pi(s)$ or $\pi(a|s)$ is thus $\arg\max_{a \in A} \left[\mu^s_a / \sum_{a \in A} \mu^s_a\right]$ or $\mu^s_a / \sum_{a \in A} \mu^s_a$. Therefore, by substitution, we end with a moral constraint $c_\rho(\mu)$ that maps to a moral principle $\rho(\pi)$.*

Second, the moral principle is considered either linear or nonlinear depending on the form of its moral constraint. If the moral constraint is linear in the occupancy measures of a policy, the moral principle is linear. Otherwise, it is nonlinear. Although we use linear moral principles for the ethical theories considered in this paper, it is possible to use nonlinear moral principles for ethical theories like utilitarianism and Kantianism. We formalize this property below.

Definition 6. *A moral principle, $\rho : \Pi \rightarrow \mathbb{B}$, is* ***linear*** *if it can be expressed as a moral constraint $c_\rho(\mu)$ that is linear with respect to the matrix of occupancy measures μ for a given policy $\pi \in \Pi$. Otherwise, it is* ***nonlinear****.*

Third, the optimization problem is described as mathematical program. As expected, to represent task completion, following the linear program of an MDP in the dual form, the program maximizes a set of occupancy measures μ^s_a for the discounted number of times an action $a \in A$ is performed in a state $s \in S$ subject to a set of constraints that maintain consistent and nonnegative occupancy. More importantly, to represent ethical compliance, the program uses a moral constraint $c_\rho(\mu)$ derived from the moral principle $\rho(\pi)$ given a matrix of occupancy measures μ for a policy π.

Fourth, the mathematical program is solved to find the optimal moral policy. Given a linear moral principle, it can be solved with linear programming techniques, such as the simplex method or the criss-cross algorithm (Bertsimas and Tsitsiklis 1997). However, given a nonlinear moral principle, it can be solved with nonlinear programming techniques (Bertsekas 1997). Note that this four-step process can also be used with the primal form of the linear program.

Ethical Frameworks

In this section, we offer a range of ethical frameworks that can be used to build an ethically compliant autonomous system. Each ethical framework approximates a well-known ethical theory in moral philosophy (Shafer-Landau 2009). During the design of an ethical framework, a developer must select a representation for the ethical context and the moral principle. This involves choosing the contextual details of the ethical context and the logical structure of the moral principle that best describe the moral implications of the system.

Table 1 offers the moral constraints that have been derived from the moral principle of each ethical framework. For each moral constraint, there are several columns that describe its computational tractability. The *Type* column lists whether the moral constraint is linear or nonlinear with respect to the occupancy measures of a policy. The *Conjunctions* column states the number of logical conjunctions that compose the moral constraint. The *Operations* column indicates an upper bound on the number of arithmetic, comparison, and logical operations that must be performed for each logical conjunction. The *Computations* column contains an upper bound on the number of computations that must be executed for the moral constraint to evaluate the moral status of a policy.

We present a set of simplified ethical frameworks examples below. Their purpose is to encode an ethical theory in a tractable way that may not capture all nuances of an ethical theory. We encourage work on more complex ethical frameworks that reflect the nuances of different ethical theories.

Divine Command Theory

Divine command theory (DCT), a monistic, absolutist ethical theory, holds that the morality of an action is based on whether a divine entity commands or forbids that action (Idziak 1979; Quinn 2013). Similar to earlier work on dead ends (Kolobov, Mausam, and Weld 2012), we consider an ethical framework that requires a policy that selects actions that have a nil probability of transitioning to any forbidden state (Mouaddib, Jeanpierre, and Zilberstein 2015).

Definition 7. *A* ***DCT ethical context****, $\mathcal{E}_\mathcal{F}$, is represented by a tuple, $\mathcal{E}_\mathcal{F} = \langle \mathcal{F} \rangle$, where $\mathcal{F}$ is a set of* ***forbidden states****.*

Definition 8. *A* ***DCT moral principle****, $\rho_\mathcal{F}$, is expressed as the following equation:*

$$\rho_\mathcal{F}(\pi) = \bigwedge_{s \in S, f \in \mathcal{F}} \big(T(s, \pi(s), f) = 0\big).$$

Note that the *DCT moral constraint* $c_{\rho_\mathcal{F}}$ in Table 1 needs $2|S||A||F|$ computations in the worst case because 2 operations are performed in $|S||A||F|$ conjunctions.

Prima Facie Duties

Prima facie duties (PFD), a pluralistic, nonabsolutist ethical theory, holds that the morality of an action is based on whether that action fulfills fundamental moral duties that can contradict each other (Ross 1930; Morreau 1996). Related to recent work on norm conflict resolution (Kasenberg and Scheutz 2018), we consider an ethical framework that requires a policy that selects actions that do not neglect duties of different penalties within some tolerance.

Definition 9. *A **PFD ethical context**, $\mathcal{E}_\Delta$, is represented by a tuple, $\mathcal{E}_\Delta = \langle \Delta, \phi, \tau \rangle$, where*

- *Δ is a set of **duties**,*
- *$\phi : \Delta \times S \rightarrow \mathbb{R}^+$ is a **penalty function** that represents the expected immediate penalty for neglecting a duty $\delta \in \Delta$ in a state $s \in S$, and*
- *$\tau \in \mathbb{R}^+$ is a **tolerance**.*

Definition 10. *A **PFD moral principle**, ρ_Δ, is expressed as the following equation:*

$$\rho_\Delta(\pi) = \sum_{s \in S} d(s) J^\pi(s) \leq \tau.$$

*The **expected cumulative penalty**, $J^\pi : S \rightarrow \mathbb{R}$, is below:*

$$J^\pi(s) = \sum_{s' \in S} T(s, \pi(s), s') \big[\sum_{\delta \in \Delta_{s'}} \phi(\delta, s') + J^\pi(s') \big],$$

where $\Delta_{s'}$ is the set of duties neglected in a state $s' \in S$.

Note that the *PFD moral constraint* c_{ρ_Δ} in Table 1 requires $3|S||A||S||\Delta| + 1$ computations in the worst case as $3|S||A||S||\Delta| + 1$ operations are performed in 1 conjunction.

Virtue Ethics

Virtue ethics (VE), a monistic, absolutist ethical theory, holds that the morality of an action is based on whether a virtuous person who acts in character performs that action in a similar situation (Anscombe 1958; Hursthouse 1999). Drawing on its natural connection to learning by demonstration from a human operator with domain expertise (Atkeson and Schaal 1997), we consider an ethical framework that requires a policy that selects actions that align with any moral trajectory performed by a moral exemplar.

Definition 11. *A **VE ethical context**, $\mathcal{E}_\mathcal{M}$, is represented by a tuple, $\mathcal{E}_\mathcal{M} = \langle \mathcal{M} \rangle$, where $\mathcal{M}$ is a set of moral trajectories.*

Definition 12. *A **VE moral principle**, $\rho_\mathcal{M}$, is expressed as the following equation:*

$$\rho_\mathcal{M}(\pi) = \bigwedge_{s \in S} \alpha(s, \pi(s)).$$

*The **alignment function**, $\alpha : S \times A \rightarrow \mathbb{B}$, is below:*

$$\alpha(s, a) = \exists_{m \in \mathcal{M}, 0 \leq i \leq \ell} \big(s = m(s_i) \land a = m(a_i) \big),$$

where $m(s_i)$ and $m(a_i)$ are the ith state and the ith action of a moral trajectory $m = \langle s_0, a_0, s_1, a_1, \ldots, s_{\ell-1}, a_{\ell-1}, s_\ell \rangle$ of length $\ell \leq L$ bounded by a maximum length L.

Note that the *VE moral constraint* $c_{\rho_\mathcal{M}}$ in Table 1 involves $|S||A|(1 + 3L|\mathcal{M}|)$ computations in the worst case since $1 + 3L|\mathcal{M}|$ operations are performed in $|S||A|$ conjunctions.

Autonomous Driving

We turn to an application of ethically compliant autonomous systems to autonomous driving. An ethically compliant self-driving vehicle must complete a navigation task by driving from an origin to a destination within a city. However, to follow a given ethical framework, the ethically compliant self-driving vehicle must adjust its route and speed depending on the type and pedestrian traffic of each road to avoid harming people and damaging property. Note that our approach can be used in many other applications, such as a security robot that patrols a college campus or a robot assistant that navigates a grocery store to help customers or prevent theft. We describe how to separate task completion and ethical compliance in an ethically compliant self-driving vehicle below.

Task Completion

The vehicle must complete a navigation task by driving from a start location $\lambda_0 \in \Lambda$ to a goal location $\lambda_g \in \Lambda$ along a set of roads Ω in a city with a set of locations Λ. At each location $\lambda \in \Lambda$, the vehicle must turn onto a road $\omega \in \Omega$. Each road $\omega \in \Omega$ is a type $\upsilon \in \Upsilon$ that indicates either a *city street*, *county road*, or *highway* with a *low*, *medium*, or *high* speed limit. Once the vehicle turns onto a road $\omega \in \Omega$, the vehicle observes the pedestrian traffic $\theta \in \Theta$ as either *light* or *heavy* with a probability $\Pr(\Theta = \theta)$. After the vehicle observes the pedestrian traffic $\theta \in \Theta$, the vehicle accelerates to a speed $\sigma \in \Sigma$ that reflects either a *low*, *normal*, or *high* speed *under*, *at*, or *above* the speed limit. To drive along the road $\omega \in \Omega$ from the current location $\lambda \in \Lambda$ to the next location $\lambda' \in \Lambda$, the vehicle cruises at the speed $\sigma \in \Sigma$. This is repeated until the vehicle arrives at the goal location $\lambda_g \in \Lambda$.

More formally, we represent the decision-making model of the navigation task by an MDP $\mathcal{D} = \langle S, A, T, R, d \rangle$. The set of states $S = S_\Lambda \cup S_\Omega$ has a set of location states S_Λ for being at a location $\lambda \in \Lambda$ and a set of road states S_Ω for being on a road $\omega \in \Omega$ of a type $\upsilon \in \Upsilon$ with a pedestrian traffic $\theta \in \Theta$ at a speed $\sigma \in \Sigma$. The set of actions $A = A_\Omega \cup A_\Sigma \cup \{\otimes, \odot\}$ has a set of turn actions A_Ω for turning onto a road $\omega \in \Omega$, a set of accelerate actions A_Σ for accelerating to a speed $\sigma \in \Sigma$, a stay action $\otimes$, and a cruise action $\odot$. The transition function $T : S \times A \times S \rightarrow [0, 1]$ reflects the dynamics of a turn action $a \in A_\Omega$ and a stay action $\otimes$ in a location state $\lambda \in S_\Lambda$ or an accelerate action $a \in A_\Sigma$ and a cruise action $\odot$ in a road state $s \in S_\Omega$ (with a self-loop for any invalid action $a \in A$). The reward function $R : S \times A \times S \rightarrow \mathbb{R}$ reflects the duration of a turn action $a \in A_\Omega$ from a location state S_Λ to a road state $s \in S_\Omega$, a stay action $\otimes$ at a location state $\lambda \in S_\Lambda$, an accelerate action $a \in A_\Sigma$ at a road state $s \in S_\Omega$, and a cruise action $\odot$ from a road state $s \in S_\Omega$ to a location state S_Λ (with an infinite duration for any invalid action $a \in A$ and a nil duration for a stay action $\otimes$ at a state $s \in S$ that represents the goal location $\lambda_g \in \Lambda$). The start state function $d : S \rightarrow [0, 1]$ has unit probability at a state $s \in S$ that represents the start location $\lambda_0 \in \Lambda$ and nil probability at every other state $s' \in S$.

Ethical Compliance

The vehicle must follow one of the ethical frameworks. First, the vehicle can follow DCT with forbidden states comprised

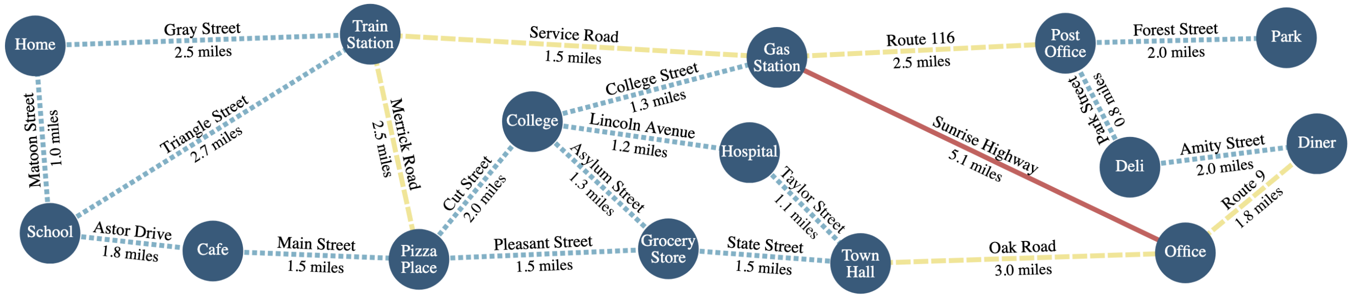

Figure 2: A city with different places that are connected by city streets, county roads, and highways.

of *hazardous* states $\mathcal{H}$ and *inconsiderate* states $\mathcal{I}$. Hazardous states $\mathcal{H}$ contain any road state at high speed and inconsiderate states $\mathcal{I}$ contain any road state at normal speed with heavy pedestrian traffic. With the DCT moral principle $\rho_{\mathcal{F}}$, we represent the DCT ethical context by a tuple, $\mathcal{E}_{\mathcal{F}} = \langle \mathcal{F} \rangle$, where $\mathcal{F} = \mathcal{H} \cup \mathcal{I}$ is the set of forbidden states.

Next, the vehicle can follow PFD with duties comprised of *smooth operation* δ_1 and *careful operation* δ_2. Smooth operation δ_1 is neglected in any road state at low speed with light pedestrian traffic while careful operation δ_2 is neglected in any road state at high speed or at normal speed with heavy pedestrian traffic. When smooth operation δ_1 and careful operation δ_2 are neglected, they incur a low and high penalty that changes with any pedestrian traffic. Neglecting duties is permitted until a limit ϵ. With the PFD moral principle ρ_{Δ}, we represent the PFD ethical context by a tuple, $\mathcal{E}_{\Delta} = \langle \Delta, \phi, \tau \rangle$, where $\Delta = \{\delta_1, \delta_2\}$ is the set of duties, $\phi : \Delta \times S \rightarrow \mathbb{R}^+$ is the penalty function that represents the expected immediate penalty for neglecting smooth operation $\delta_1 \in \Delta$ and careful operation $\delta_2 \in \Delta$ in a state $s \in S$ with a pedestrian traffic $\theta \in \Theta$, and $\tau = \epsilon$ is the tolerance.

Finally, the vehicle can follow VE with moral trajectories comprised of *cautious* trajectories $\mathcal{C}$ and *proactive* trajectories $\mathcal{P}$. Cautious trajectories $\mathcal{C}$ exemplify driving on any road state at normal speed with light pedestrian traffic or at low speed with heavy pedestrian traffic and proactive trajectories $\mathcal{P}$ exemplify avoiding any highway road states and a set of populated location states. With the VE moral principle $\rho_{\mathcal{M}}$, we represent the VE ethical context by a tuple, $\mathcal{E}_{\mathcal{M}} = \langle \mathcal{M} \rangle$, where $\mathcal{M} = \mathcal{C} \cup \mathcal{P}$ is the set of moral trajectories.

Experiments

We now demonstrate that the application of ethically compliant autonomous systems to autonomous driving is effective in a set of simulations and a user study.

In the set of simulations, a standard self-driving vehicle that cannot follow any ethical framework and an ethically compliant self-driving vehicle that can follow different ethical frameworks must complete a set of navigation tasks.

Each navigation task can use a different start location $\lambda_0 \in \Lambda$ and goal location $\lambda_g \in \Lambda$ based on the city in Figure 2. The speed limits of city streets, county roads, and highways are 25, 45, and 75 MPH. The probability $\Pr(\Theta = \theta)$ of observing light or heavy pedestrian traffic $\theta \in \Theta$ is 0.8 and 0.2. A low, normal, and high speed is 10 MPH under, at, and 10 MPH above the speed limit. Turning onto a road $\omega \in \Omega$ from a location $\lambda \in \Lambda$ requires 5 seconds. Accelerating 10 MPH requires 2 seconds. Cruising requires a time equal to the distance of the road $\omega \in \Omega$ divided by the speed $\sigma \in \Sigma$. Staying at a location $\lambda \in \Lambda$ other than the goal location $\lambda_g \in \Lambda$ requires 120 seconds.

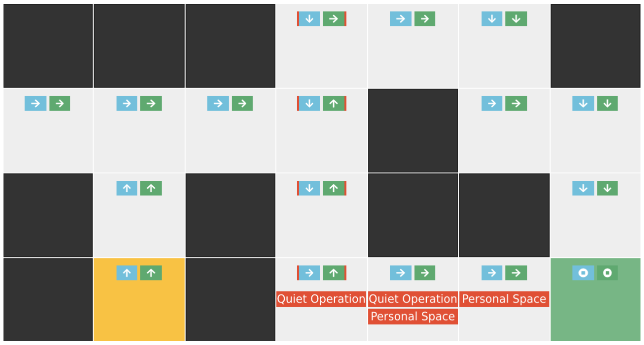

Figure 3: An agent completes a task and follows an ethical framework with a *blue* amoral path and a *green* moral path in the *Morality.js* customizable grid world environment.

Each ethical framework can use different settings. For DCT, the forbidden states $\mathcal{F}$ can be just hazardous states $\mathcal{H}$ or both hazardous states $\mathcal{H}$ and inconsiderate states $\mathcal{I}$. For PFD, the tolerance $\tau = \epsilon$ can be the limit $\epsilon = 3$, $\epsilon = 6$, or $\epsilon = 9$. For VE, the moral trajectories can be just cautious trajectories $\mathcal{C}$ or both cautious trajectories $\mathcal{C}$ and proactive trajectories $\mathcal{P}$ that avoid any highway road states and a set of populated location states that contains the *School* and *College* locations with many students on campus.

Table 2 shows that the price of morality incurred by the agent is appropriate for each ethical framework. The standard self-driving vehicle does not incur a price of morality. However, the ethically compliant self-driving vehicle incurs a price of morality that increases with more forbidden states for DCT, decreases with more tolerance for PFD, and increases with more moral trajectories for VE.

Figure 5 shows that the behavior performed by the agent is appropriate for each ethical framework. The standard self-driving vehicle drives the shortest route at high speed. However, the ethically compliant self-driving vehicle differs for each ethical framework. For DCT, the vehicle drives the shortest route at low or normal speed based on pedestrian traffic. For PFD, the vehicle drives the shortest route at low

Ethics	Setting	Task 1 (%)	Task 2 (%)	Task 3 (%)
None	—	0	0	0
DCT	$\mathcal{H}$	14.55	15.33	20.12
	$\mathcal{H} \cup \mathcal{I}$	21.13	22.35	27.92
PFD	$\epsilon = 3$	16.07	16.52	24.30
	$\epsilon = 6$	11.96	11.80	21.37
	$\epsilon = 9$	7.91	7.15	18.87
VE	$\mathcal{C}$	21.13	22.35	27.92
	$\mathcal{C} \cup \mathcal{P}$	40.89	94.43	30.28

Table 2: The price of morality relative to the value of the optimal amoral policy for each vehicle on all navigation tasks.

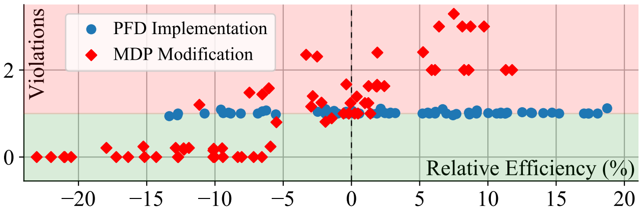

Figure 4: The results of the user study. For each task and location in the city, a point denotes the resulting policy. For each policy, the horizontal axis is its time savings relative to the policy from the opposing task while the vertical axis is its number of violations, averaged over 10 simulations. The moral and immoral regions are highlighted in *green* and *red*.

or normal speed based on pedestrian traffic aside from driving on the first road at normal or high speed with some probability for light pedestrian traffic and at normal speed for heavy pedestrian traffic due to the tolerance. For VE, the vehicle drives at low or normal speed based on pedestrian traffic but drives a different route to avoid any highway road states and the set of populated location states.

In the user study, 7 planning and robotics experts with experience in MDPs but not ethics had to complete two tasks that implemented an ethically compliant self-driving vehicle in a random order. In both tasks, developers were given the decision-making model for navigating the city from any start location to the Office location. They then had to enforce the following moral requirements: the self-driving vehicle should drive at *high* speed with *light* pedestrian traffic or at *normal* speed with *heavy* pedestrian traffic at most once in expectation but should never drive at *high* speed with *heavy* pedestrian traffic. In one task, developers were asked to enforce these requirements by modifying the reward function of the decision-making model, specifically an MDP. In the other task, developers were asked to enforce these requirements by defining the ethical context of an ethical framework, specifically PFD. The user study therefore evaluates the accuracy and usability of modifying a decision-making model versus defining an ethical framework.

Figure 4 shows that defining the ethical context of the ethical framework led to better policies than modifying the reward function of the decision-making model. In our approach, all policies optimize the navigation task and satisfy the requirements with exactly one violation. However, in the

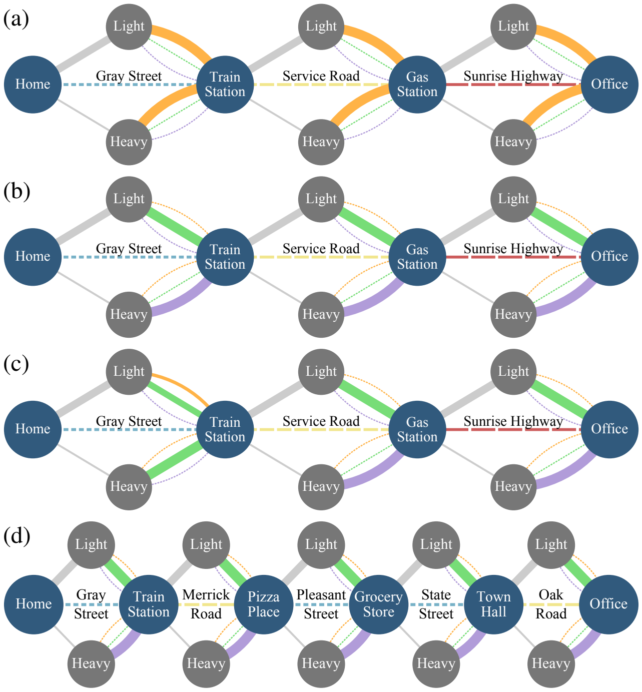

Figure 5: The optimal policies for select vehicles with (a) no ethical framework, (b) DCT with $\mathcal{H} \cup \mathcal{I}$, (c) PFD with $\epsilon = 9$, and (d) VE with $\mathcal{C} \cup \mathcal{P}$ on a navigation task. A *blue* node denotes a location and a *gray* node denotes pedestrian traffic. With a thickness that represents probability, a *gray* line denotes turning onto a road and an *orange*, *green*, or *purple* line denotes cruising at high, normal, or low speed.

other approach, most policies fail to optimize the navigation task or satisfy the requirements: aggressive policies in the upper right corner with more than one violation are faster but immoral while conservative policies with less than one violation in the lower left corner are slower but moral. It is also encouraging that our method (24 minutes) had a lower mean development time than the other method (45 minutes).

Our open source library, *Morality.js*, available on the website https://www.moralityjs.com with the customizable grid world environment dashboard in Figure 3, was used in all experiments (Svegliato, Nashed, and Zilberstein 2020a,b).

Conclusion

We propose a novel approach for building ethically compliant autonomous systems that optimize completing a task while following an ethical framework. It simplifies and standardizes the process of implementing an ethical theory within autonomous systems as it is general-purpose, modular, and interpretable. We then offer a range of ethical frameworks for divine command theory, prima facie duties, and virtue ethics. Finally, we demonstrate the accuracy and usability of our approach in a set of autonomous driving simulations and a user study of planning and robotics experts. Future work will develop nuanced ethical frameworks for the ethical theories in this paper and explore new ethical frameworks for ethical theories like utilitarianism and Kantianism.

Acknowledgments

We thank the anonymous reviewers for their valuable comments and Alan Labouseur for his helpful feedback on the *Morality.js* library. This work was supported in part by NSF Graduate Research Fellowship DGE-1451512, NSF grant IIS-1724101, and NSF grant IIS-1813490.

Ethics Statement

Although we discuss important ethical considerations surrounding ethically compliant autonomous systems throughout the paper, we highlight three ethical implications below.

First, we stress that simply defining some ethically compliant autonomous system does not guarantee that it exhibits perfect ethical compliance with respect to its developers or stakeholders in the real world. In fact, similar to any autonomous system, the quality of an ethically compliant autonomous system is limited by the accuracy of its decision-making model, ethical framework, and ethical context. If these attributes have not been specified in a way that reflects the intentions of its developers or the values of its stakeholders, the system may still result in undesirable consequences. Moreover, any conflict between stakeholders can perhaps be resolved using multiple ethical frameworks together by forming a moral principle that is a conjunction over the moral principle for each ethical framework. It is therefore critical that developers seek continual participation and feedback from a range of stakeholders who interact with the system in as many diverse situations as possible.

Next, while our approach gives autonomous systems the ability to satisfy an arbitrary set of moral requirements, we emphasize that developers must still remain transparent about the moral requirements of their autonomous systems. This could be in the form of ethical documentation that specifies the moral requirements of the autonomous system and its limitations. For example, if a self-driving vehicle has an ethical framework that considers the level of pedestrian traffic but not the presence of wildlife along a route, there should be ethical documentation in the user manual that is provided to the owners of the vehicle. Hence, by providing ethical documentation prior to the deployment of the autonomous system, deployers can ensure that any conditions necessary for ethical compliance are satisfied throughout operation.

Finally, even though our approach gives autonomous system the ability to comply with a given ethical theory, we highlight that developers must still think carefully about the design and development of their autonomous systems. This involves selecting the moral requirements of the autonomous system, which can include determining the best *ethical theory*, the best *ethical framework* for that ethical theory, and the best *settings* for that ethical framework. In other words, our approach does not substitute for the deliberate process of determining the best way to build an ethically compliant autonomous system. In fact, developers should avoid intentionally selecting ethically compliant autonomous systems that are easy to implement in practice. However, as the vast discourse surrounding the best ethical theory to use in autonomous systems continues to evolve over time, our approach can be used in a way that reflects this discussion.

References

Abel, D.; MacGlashan, J.; and Littman, M. L. 2016. Reinforcement learning as a framework for ethical decisions. In *AAAI Workshop on AI, Ethics, and Society*.

Adamson, G.; Havens, J. C.; and Chatila, R. 2019. Designing a value-driven future for ethical autonomous and intelligent systems. *Proceedings of the IEEE* 107(3): 518–525.

Allen, C.; Smit, I.; and Wallach, W. 2005. Artificial morality: Top-down, bottom-up, and hybrid approaches. *Ethics and Information Technology* 7(3): 149–155.

Anderson, M.; Anderson, S. L.; and Berenz, V. 2017. A value driven agent: An instantiation of a case-supported principle-based behavior paradigm. In *AAAI Workshop on AI, Ethics, and Society*.

Anscombe, G. E. M. 1958. Modern moral philosophy. *Philosophy* 33(124): 1–19.

Arkin, R. C. 2008. Governing lethal behavior: Embedding ethics in a hybrid deliberative/reactive robot architecture. In *3rd ACM/IEEE International Conference on Human Robot Interaction*. ACM.

Atkeson, C. G.; and Schaal, S. 1997. Robot learning from demonstration. In *International Conference On Machine Learning*, volume 97, 12–20.

Atkinson, K.; and Bench-Capon, T. 2006. Addressing moral problems through practical reasoning. In *International Workshop on Deontic Logic and Artificial Normative Systems*. Springer.

Basich, C.; Svegliato, J.; Wray, K. H.; Witwicki, S.; Biswas, J.; and Zilberstein, S. 2020. Learning to optimize autonomy in competence-aware systems. In *19th International Conference on Autonomous Agents and Multiagent Systems*.

Basich, C.; Svegliato, J.; Zilberstein, S.; Wray, K. H.; and Witwicki, S. J. 2021. Improving competence for reliable autonomy. In *ECAI Workshop on Agents and Robots for Reliable Engineered Autonomy*.

Bellman, R. 1952. On the theory of dynamic programming. *National Academy of Sciences of the United States of America* 38(8): 716.

Bench-Capon, T.; and Modgil, S. 2017. Norms and value based reasoning: Justifying compliance and violation. *Artificial Intelligence and Law* 25(1): 29–64.

Bentham, J. 1789. *An introduction to the principles of morals*. London: Athlone.

Berreby, F.; Bourgne, G.; and Ganascia, J.-G. 2015. Modelling moral reasoning and ethical responsibility with logic programming. In *Logic for Programming, Artificial Intelligence, and Reasoning*. Springer.

Bertsekas, D. P. 1997. Nonlinear programming. *Journal of the Operational Research Society* 48(3): 334–334.

Bertsimas, D.; and Tsitsiklis, J. N. 1997. *Introduction to linear optimization*. Athena Scientific Belmont, MA.

Boden, M.; Bryson, J.; Caldwell, D.; Dautenhahn, K.; Edwards, L.; Kember, S.; Newman, P.; Parry, V.; Pegman, G.; Rodden, T.; Sorrell, T.; Wallis, M.; Whitby, B.; and Winfield, A. 2017. Principles of robotics: regulating robots in the real world. *Connection Science* 29(2): 124–129.

Bostrom, N. 2016. *Superintelligence: Paths, Dangers, Strategies*. Wiley Online Library.

Brey, P. 2004. Ethical aspects of facial recognition systems in public places. *Journal of Information, Communication, and Ethics In Society* 2(2): 97–109.

Bringsjord, S.; Arkoudas, K.; and Bello, P. 2006. Toward a general logicist methodology for engineering ethically correct robots. *IEEE Intelligent Systems* 21(4): 38–44.

Bringsjord, S.; Taylor, J.; Van Heuveln, B.; Arkoudas, K.; Clark, M.; and Wojtowicz, R. 2011. Piagetian roboethics via category theory: Moving beyond mere formal operations to engineer robots whose decisions are guaranteed to be ethically correct. In *Machine Ethics*. Cambridge University Press.

Browne, S. 2015. *Dark matters: On the surveillance of blackness*. Duke University Press.

Burgers, T.; and Robinson, D. R. 2017. Networked authoritarianism is on the rise. *Sicherheit und Frieden* 248–252.

Charisi, V.; Dennis, L.; Fisher, M.; Lieck, R.; Matthias, A.; Slavkovik, M.; Sombetzki, J.; Winfield, A. F.; and Yampolskiy, R. 2017. Towards moral autonomous systems. In *arXiv preprint arXiv:1703.04741*.

Dennis, L.; Fisher, M.; Slavkovik, M.; and Webster, M. 2016. Formal verification of ethical choices in autonomous systems. *Robotics and Autonomous Systems* 77: 1–14.

Desai, D. R.; and Kroll, J. A. 2017. Trust but verify: A guide to algorithms and the law. *Harvard Journal of Law and Technology* 31: 1.

Dignum, V.; Baldoni, M.; Baroglio, C.; Caon, M.; Chatila, R.; Dennis, L.; Génova, G.; Haim, G.; Kließ, M. S.; Lopez-Sanchez, M.; et al. 2018. Ethics by Design: necessity or curse? In *AAAI/ACM Conference on AI, Ethics, and Society*.

Fallman, D. 2003. Design-oriented human-computer interaction. In *SIGCHI Conference on Human Factors in Computing Systems*.

Friedman, B.; Kahn, P. H.; and Borning, A. 2008. Value sensitive design and information systems. *The handbook of information and computer ethics* 69–101.

Goodman, B.; and Flaxman, S. 2017. European Union regulations on algorithmic decision-making and a "right to explanation". *AI magazine* 38(3): 50–57.

Hadfield-Menell, D.; and Hadfield, G. K. 2019. Incomplete contracting and AI alignment. In *AAAI/ACM Conference on AI, Ethics, and Society*.

Hursthouse, R. 1999. *On virtue ethics*. Oxford University Press.

Idziak, J. M. 1979. *Divine command morality*. Edwin Mellen Press.

Introna, L.; and Wood, D. 2004. Picturing algorithmic surveillance: The politics of facial recognition systems. *Surveillance & Society* 2(2/3): 177–198.

Kant, I.; and Schneewind, J. B. 2002. *Groundwork for the metaphysics of morals*. Yale University Press.

Kasenberg, D.; and Scheutz, M. 2018. Norm conflict resolution in stochastic domains. In *32nd AAAI Conference on Artificial Intelligence*.

Kolobov, A.; Mausam; and Weld, D. S. 2012. A theory of goal-oriented MDPs with dead ends. In *28th Conference on Uncertainty in Artificial Intelligence*.

Manne, A. S. 1960. Linear programming and sequential decisions. *Management Science* 6(3): 259–267.

Mill, J. S. 1895. *Utilitarianism*. Longmans, Green and Company.

Moor, J. H. 2006. The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems* 21(4): 18–21.

Morreau, M. 1996. Prima Facie and seeming duties. *Studia Logica* 57(1): 47–71.

Mouaddib, A.-I.; Jeanpierre, L.; and Zilberstein, S. 2015. Handling advice in MDPs for semi-autonomous systems. In *ICAPS Workshop on Planning and Robotics*. Jerusalem, Israel.

Pasquale, F. 2017. Toward a fourth law of robotics: Preserving attribution, responsibility, and explainability in an algorithmic society. *Ohio State Law Journal* 78: 1243.

Quinn, P. L. 2013. Divine command theory. *The Blackwell guide to ethical theory* 81–102.

Raymond, A. H.; and Shackelford, S. J. 2013. Technology, ethics, and access to justice: Should an algorithm be deciding your case. *Michigan Journal of International Law.* 35: 485.

Read, G. J.; Salmon, P. M.; Lenné, M. G.; and Stanton, N. A. 2015. Designing sociotechnical systems with cognitive work analysis: Putting theory back into practice. *Ergonomics* 58(5): 822–851.

Robertson, L. J.; Abbas, R.; Alici, G.; Munoz, A.; and Michael, K. 2019. Engineering-based design methodology for embedding ethics in autonomous robots. *Proceedings of the IEEE* 107(3): 582–599.

Ross, W. D. 1930. *The right and the good*. Oxford University Press.

Scherer, M. U. 2015. Regulating artificial intelligence systems: Risks, challenges, competencies, and strategies. *Harvard Journal of Law and Technology* 29: 353.

Shafer-Landau, R. 2009. *The fundamentals of ethics*. Oxford University Press.

Shaw, N. P.; Stöckel, A.; Orr, R. W.; Lidbetter, T. F.; and Cohen, R. 2018. Towards provably moral AI agents in bottom-up learning frameworks. In *AAAI/ACM Conference on AI, Ethics, and Society*.

Shim, J.; Arkin, R.; and Pettinatti, M. 2017. An Intervening Ethical Governor for a robot mediator in patient-caregiver relationship. In *IEEE International Conference on Robotics and Automation*.

Svegliato, J.; Nashed, S.; and Zilberstein, S. 2020a. An integrated approach to moral autonomous systems. In *24th European Conference on Artificial Intelligence*.

Svegliato, J.; Nashed, S. B.; and Zilberstein, S. 2020b. Ethically compliant planning in moral autonomous systems. In *IJCAI Workshop in Artificial Intelligence Safety*.

Svegliato, J.; Wray, K. H.; Witwicki, S. J.; Biswas, J.; and Zilberstein, S. 2019. Belief space metareasoning for exception recovery. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*.

Taylor, J.; Yudkowsky, E.; LaVictoire, P.; and Critch, A. 2016. Alignment for advanced machine learning systems. In *Machine Intelligence Research Institute*.

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

Vanderelst, D.; and Winfield, A. 2018. An architecture for ethical robots inspired by the simulation theory of cognition. *Cognitive Systems Research* 48: 56–66.

Winfield, A. F.; Blum, C.; and Liu, W. 2014. Towards an ethical robot: Internal models, consequences and ethical action selection. In *Conference Towards Autonomous Robotic Systems*. Springer.

Wooldridge, M.; and Van Der Hoek, W. 2005. On obligations and normative ability: Towards a logical analysis of the social contract. *Journal of Applied Logic* 3(3-4): 396–420.

Zimmer, M. 2008. The gaze of the perfect search engine: Google as an infrastructure of dataveillance. In *Web Search*. Springer.