# Ethically Compliant Planning in Moral Autonomous Systems

**Justin Svegliato** and **Samer B. Nashed** and **Shlomo Zilberstein**
College of Information and Computer Sciences, University of Massachusetts Amherst
{jsvegliato,snashed,shlomo}@cs.umass.edu

## Abstract

In many sequential decision-making problems, ethical compliance is enforced by either myopic rule sets or provisional modifications to the objective function. The effect of these strategies is exceedingly difficult to predict, often leading to inadvertent behavior that can jeopardize the values of stakeholders. We propose a novel approach for ethically compliant planning, based on decoupling ethical compliance from task completion within the objective function, that produces optimal policies subject to the constraints of an ethical framework. This paper introduces a formal definition of a moral autonomous system and its key properties. It also offers a range of ethical framework examples for divine command theory, prima facie duties, and virtue ethics. Finally, it demonstrates the effectiveness of our approach in a set of autonomous driving simulations and a user study of MDP experts.

## 1 Introduction

Integrating decision making and ethics in autonomous systems is challenging due to the diversity and complexity of deployment domains and stakeholder value systems. For decision making in the real world, Markov decision processes (MDPs) are a common, general-purpose model because of their support for long-term, nonmyopic reasoning in fully observable, stochastic environments. However, MDPs pose two additional challenges when generating ethically compliant behavior. First, the complexity of these models often obfuscates the effect of the reward function on the behavior of the agent. Seemingly innocuous adjustments may drastically change resulting behavior, leading to unpredictability [Bostrom, 2016]. Second, and more fundamentally, using the reward function to model both desirable and undesirable behavior often involves incommensurable unit conversions. For example, an autonomous vehicle with a reward function that encourages completing a route efficiently and discourages driving recklessly blends task completion and ethical compliance implicitly. The resulting policy may drive too recklessly if offered enough time savings. Thus, in complex environments, autonomous systems may encounter unanticipated scenarios that lead to behavior that fails to reflect the intentions of developers or the values of stakeholders [Taylor *et al.*, 2016; Hadfield-Menell and Hadfield, 2019].

Ideally, researchers and practitioners integrating ethical theories and decision processes could access methods that offer several desirable features. These features include support for interpretability and control over behavior with formal guarantees, nonmyopic reasoning, the acquisition of rules from a non-technical person, and the application of one or more ethical theories simultaneously. Ethicists often describe an ethical theory as a set of moral principles for evaluating if an action is required, permitted, or prohibited in a given scenario [Shafer-Landau, 2009]. Given this interpretation, an ethical theory can be operationalized in a decision process as constraints on the actions of the agent in specific states.

In this paper, we propose a novel approach for building *moral autonomous systems* that produce an optimal policy to a decision-making problem subject to the constraints of an ethical framework. The system models a task with a *decision-making model* and models an ethical framework as a *moral principle* and an *ethical context*. While we use MDPs for the decision-making models in our experiments, our approach supports any decision process expressible as a mathematical program. The moral principle is an *approximation* of an interpretation of an ethical theory that can be represented as a Boolean function that evaluates whether or not a policy violates a particular ethical theory. The ethical context contains all of the information necessary to evaluate the moral principle. Formally, this system is expressed as an optimization problem with a set of constraints representing the task and a constraint that operationalizes the ethical framework. The solution to the optimization problem is a policy that optimizes completing the task while following the ethical framework.

We evaluate our approach in two experiments. First, in an autonomous driving simulation, we confirm that our approach produces optimal behavior while complying with moral requirements. Second, in a user study, we find that MDP experts who use our approach require less development time to produce policies that have higher rates of ethical compliance compared to modifying the reward function directly.

Our main contributions in this paper are: (1) a formal definition of a moral autonomous system and its key properties, (2) a range of ethical framework examples for divine command theory, prima facie duties, and virtue ethics, and (3) a set of autonomous driving simulations and a user study of MDP experts that shows the effectiveness of our approach.

## 2 Related Work

Autonomous systems attempt to address a range of problems, are deployed in diverse social contexts, and draw upon a heterogeneous collection of algorithms. The potential harms of these systems can be mitigated through many strategies: (1) abandonment of technologies that are likely to be abused when analyzed in a historical context [Browne, 2015], such as facial recognition [Brey, 2004; Introna and Wood, 2004] and surveillance of online activity [Burgers and Robinson, 2017; Zimmer, 2008], (2) legal or legislative intervention that provides oversight and regulation in enough detail to prevent or discourage malevolent or negligent use [Goodman and Flaxman, 2017; Desai and Kroll, 2017; Raymond and Shackelford, 2013; Scherer, 2015], including meta-regulation [Pasquale, 2017], and (3) algorithmic advances that improve accuracy and interpretability. Though these strategies will continue to play important roles in the future, our approach focuses on a fourth strategy that reduces the opportunity for error during design and development.

Recently, various principles [Boden *et al.*, 2017], guidelines [Robertson *et al.*, 2019], and standards [Adamson *et al.*, 2019] have been proposed for the design and development of autonomous systems. Although these are essential for promoting the values of stakeholders throughout the design process, these initiatives do not offer developers enough detail to operationalize ethical frameworks in autonomous systems. In fact, implicit ethical systems, which satisfy moral requirements through careful design, may not always produce desirable behavior [Moor, 2006]. Many autonomous systems must therefore be capable of explicit moral reasoning [Dignum *et al.*, 2018; Bench-Capon and Modgil, 2017].

Engineering efforts to develop explicit autonomous moral agents take two forms [Allen *et al.*, 2005]. *Bottom-up* approaches generate ethical behavior naturally through learning or evolution [Anderson *et al.*, 2017; Shaw *et al.*, 2018]. While this is theoretically compelling given the natural evolution of ethical ideas in human society, the instability and lack of interpretability are major drawbacks. Instead, we choose a *top-down* approach in which prescriptive rules describing moral behavior are provided to the agent. Many top-down approaches use logics, such as deontic logic [van der Torre, 2003; Bringsjord *et al.*, 2006], temporal logic [Wooldridge and Van Der Hoek, 2005; Atkinson and Bench-Capon, 2006], or Answer Set Programming [Berreby *et al.*, 2015]. Some even propose a form of metareasoning over logics [Bringsjord *et al.*, 2011]. However, as systems become more complex and operate in stochastic, partially observable environments, norm specification represented by logics will become increasingly challenging [Abel *et al.*, 2016].

A common approach for handling these environments employs an *ethical governor* that reasons online about whether an action is required, permitted, or prohibited [Arkin, 2008]. Applications include eldercare [Shim *et al.*, 2017] and physical safety [Vanderelst and Winfield, 2018; Winfield *et al.*, 2014]. These methods use reactive ethical governors that only consider a single action at a time as the situation presents the agent with the opportunity to act. In contrast, our approach is nonmyopic because it considers entire sequences of actions.

We are aware of only one other approach that focuses on proactive ethical governors of policies [Kasenberg and Scheutz, 2018]. However, since it is specific to norms, it is unclear how it could support other forms of ethical reasoning, such as adherence to a moral principle like utilitarianism or deontology. Moreover, both task completion and ethical behavior are defined in terms of real-valued norm weights, the coupling of which elides guarantees of ethical behavior. In contrast, our approach can generate policies that follow arbitrary ethical theories and avoid unpredictable trade-offs between task completion and ethical behavior.

## 3 Background

A *Markov decision process* (MDP) is a decision-making model for reasoning in fully observable, stochastic environments [Bellman, 1966]. 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, $T : S \times A \times S \to [0, 1]$ represents the probability of reaching a state $s' \in S$ after performing an action $a \in A$ in a state $s \in S$, $R : S \times A \times S \to \mathbb{R}$ 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 $d : S \to [0, 1]$ represents the probability of starting in a state $s \in S$. A solution to an MDP is a policy $\pi : S \to A$ indicating that an action $\pi(s) \in A$ should be performed in a state $s \in S$. A policy $\pi$ induces a value function $V^\pi : S \to \mathbb{R}$ 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 $\pi^*$ 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 $\mu_a^s$ 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.

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

## 4 Moral Autonomous Systems

We propose a novel approach for building *moral autonomous systems* that decouples ethical compliance from task completion. The system completes a task by using a *decision-making model* and follows an ethical framework by adhering to a *moral principle* within an *ethical context*. We describe these three components of a moral 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 [Nashed *et al.*, 2018;

Svegliato *et al.*, 2019]. An engineer 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 and hazardous driving that permit speeding on a highway in some scenarios but never in a school zone or near a crosswalk [Vanderelst and Winfield, 2018]. Similar to the decision-making model, an ethicist must select a representation for the ethical context that informs the fundamental principles of the ethical framework. While the ethical context can be represented as a tuple of different values, sets, and functions, the specification of the tuple depends on the ethical framework. In summary, the ethical context is a *moral*, descriptive 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 of the decision-making model within the ethical context by considering the information that describes how to complete the task and follow the ethical framework. As an illustration, a moral principle could require a policy to maximize the overall 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 moral principle can be expressed as a function that maps a policy to its moral status.

**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** $\mathcal{E}$.*

By putting all of these attributes together, we provide a formal description of a moral autonomous system as follows.

**Definition 2.** *A **moral autonomous system**, $\langle \mathcal{D}, \mathcal{E}, \rho \rangle$, completes a task by using a decision-making model $\mathcal{D}$ and follows an ethical framework by adhering to a moral principle $\rho$ within an ethical context $\mathcal{E}$.*

A moral autonomous system has the goal of finding an optimal policy that completes its task and follows its ethical framework. This can be expressed as an optimization problem solving for a policy in the space of policies that maximizes the value of the policy subject to the constraint that the policy satisfies the moral principle. We define the goal of a moral autonomous system as follows.

**Definition 3.** *The goal of a moral autonomous system is to find an **optimal moral policy**, $\pi^*_\rho \in \Pi$, by solving for a policy $\pi \in \Pi$ that maximizes a value function $V^\pi$ subject to a moral principle $\rho(\pi)$ in the following optimization problem.*

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

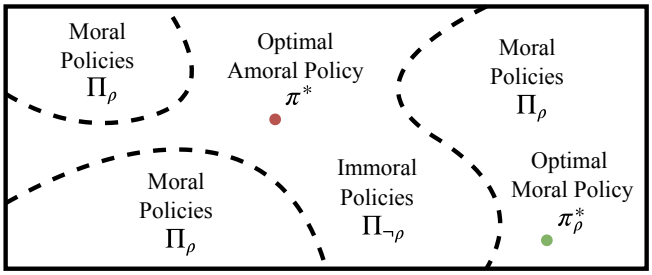


Figure 1: A simple view of the goal of a moral autonomous system and a standard autonomous system in terms of the space of policies

However, the goal 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 depicts the goal of a moral autonomous system and a standard autonomous system. For a moral principle $\rho$, the space of policies $\Pi$ is partitioned into a moral region $\Pi_\rho$ and an immoral region $\Pi_{\neg\rho}$. The moral region contains the optimal moral policy $\pi^*_\rho \in \Pi$ of the moral autonomous system while the immoral region contains the optimal amoral policy $\pi^* \in \Pi$ of the standard autonomous system. In general, the optimal amoral policy $\pi^* \in \Pi$ can be contained by either the moral region $\Pi_\rho$ or the immoral region $\Pi_{\neg\rho}$.

A moral autonomous system may follow an ethical framework that adversely impacts completing its task. Engineers and ethicists can assess 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 define this cost below.

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

A moral autonomous system may even follow an ethical framework that is mutually exclusive with completing its task. In this situation, engineers and ethicists should reconsider the moral implications of the system and could augment the decision-making model or adjust the ethical context if deemed safe. Naturally, depending on whether or not there is a solution to the optimization problem, the system can be considered either feasible or infeasible as follows.

**Definition 5.** *A moral autonomous system is **realizable** if and only if there exists a policy $\pi \in \Pi$ such that its moral principle $\rho(\pi)$ is satisfied. Otherwise, the system is **unrealizable**.*

It is natural to find the optimal moral policy by solving the optimization problem of a moral autonomous system using mathematical programming. This process involves four steps. First, the moral principle can be mapped to a moral constraint in terms of the occupancy measures of a policy. We show that this mapping can always be performed as follows.

**Theorem 1.** *A moral principle, $\rho : \Pi \to \mathbb{B}$, can be expressed as a moral constraint $c_\rho(\mu)$ in terms of the matrix of occupancy measures $\mu$ 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 $\pi(a|s)$. Recall*

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

*that the discounted number of times that an action $a \in A$ is performed in a state $s \in S$ is an occupancy measure $\mu_a^s$. Observe that the discounted number of times that a state $s \in S$ is visited is the expression $\sum_{a\in A}\mu_a^s$. A policy $\pi(s)$ or $\pi(a|s)$ is thus* $\arg\max_{a\in A}\left[\mu_a^s / \sum_{a\in A}\mu_a^s\right]$ *or* $\mu_a^s / \sum_{a\in A}\mu_a^s$*. Therefore, by substitution, we end with a moral constraint $c_\rho(\mu)$.*

Second, the moral principle can be classified as 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, the moral principle is nonlinear. We formalize this property below.

**Definition 6.** *A moral principle, $\rho : \Pi \to \mathbb{B}$, 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 $\mu$ for a given policy $\pi \in \Pi$. Otherwise, the moral principle is **nonlinear**.*

Third, the optimization problem can be described as a mathematical program. For task completion, following the linear program of an MDP in the dual form, the program maximizes a set of occupancy measures $\mu_a^s$ 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. However, for ethical compliance, the program has a moral constraint $c_\rho(\mu)$ derived from the moral principle $\rho(\mu)$ given a matrix of occupancy measures $\mu$.

Fourth, the mathematical program can be solved to find the optimal moral policy. Given a linear moral principle, it can be solved using techniques designed for linear programming, such as the simplex method or the criss-cross algorithm [Bertsimas and Tsitsiklis, 1997]. However, given a nonlinear moral principle, it can be solved using techniques designed for nonlinear programming instead [Bertsekas, 1997]. Note that, while we use the dual form of the linear program of an MDP, this process can also be used with the primal form.

# 5 Ethical Frameworks

In this section, we offer a range of ethical framework examples that can be used to build a moral autonomous system. Each ethical framework is influenced by an interpretation of an ethical theory in moral philosophy [Shafer-Landau, 2009]. During the design of an ethical framework, ethicists and engineers 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 most accurately describe the capabilities of the agent, the effect of its actions on its environment, and the moral implications of its behavior. In short, an ethical framework, composed of an ethical context and a moral principle, is an *approximation* of an interpretation of an ethical theory.

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 now present a set of simplified ethical frameworks. They are not definitive and do not capture all nuances of ethical theories. Their purpose is to tractably operationalize an ethical theory within a decision process. We encourage the development of more complex ethical frameworks that reflect the depth of different ethical theories, including those below.

## 5.1 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]. We consider a simplified ethical framework in which a moral autonomous system uses a policy that selects actions that have a nil probability of transitioning to any forbidden state [Mouaddib *et al.*, 2015]

**Definition 7.** *A **DCT ethical context**, $\mathcal{E}_\mathcal{F}$, can 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**, $\rho_\mathcal{F}$, can be expressed as the following equation:*

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

## 5.2 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]. We consider a simplified ethical framework in which a moral autonomous system uses a policy that selects actions that do not neglect duties of different penalties within some tolerance.

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

- *$\Delta$ is a set of **duties**,*

- $\phi : \Delta \times S \to \mathbb{R}^+$ *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****,* $\rho_\Delta$*, can be expressed as the following equation:*

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

*The* ***expected cumulative penalty****,* $J^\pi : S \to \mathbb{R}$*, 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* $s' \in S$*.*

### 5.3 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]. We consider a simplified ethical framework in which a moral autonomous system uses a policy that selects actions that align with any moral trajectory performed by a moral exemplar.

**Definition 11.** *A* ***VE ethical context****,* $\mathcal{E}_\mathcal{M}$*, can 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****,* $\rho_\mathcal{M}$*, can be expressed as the following equation:*

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

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

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

*where* $m(s_i)$ *and* $m(a_i)$ *are the* $i$*th state and the* $i$*th 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* $L$*.*

## 6 Autonomous Driving

We turn to an application of moral autonomy to autonomous driving. A moral self-driving vehicle must complete a navigation task by driving from an origin to a destination in a city. However, to follow an ethical framework, the moral self-driving vehicle must adjust its route and speed depending on the type and pedestrian traffic of each road. We describe how to separate task completion and ethical compliance below.

### 6.1 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 $\Omega$ in a city with a set of locations $\Lambda$. 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 $\sigma \in \Sigma$. Note that this is repeated until arriving at the goal location $\lambda_g \in \Lambda$.

We represent the decision-making model of a 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 $S_\Lambda$ for being at a location $\lambda \in \Lambda$ and a set of road states $S_\Omega$ 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 $A_\Omega$ for turning onto a road $\omega \in \Omega$, a set of accelerate actions $A_\Sigma$ for accelerating to a speed $\sigma \in \Sigma$, a stay action $\otimes$, and a cruise action $\odot$. The transition function $T : S \times A \times S \to [0, 1]$ 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 $R : S \times A \times S \to \mathbb{R}$ reflects the duration of a turn action $a \in A_\Omega$ from a location state $S_\Lambda$ 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 $S_\Lambda$ (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 $d : S \to [0, 1]$ 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 $s \in S$.

### 6.2 Ethical Compliance

The vehicle must follow one of the ethical frameworks. First, the vehicle can follow DCT with forbidden states comprised of *hazardous* states $\mathcal{H}$ and *inconsiderate* states $\mathcal{I}$. Hazardous states $\mathcal{H}$ contain any road state at high speed while 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* $\delta_1$ and *careful operation* $\delta_2$. Smooth operation $\delta_1$ is neglected in any road state at low speed with light pedestrian traffic while careful operation $\delta_2$ is neglected in any road state at high speed or at normal speed with heavy pedestrian traffic. When smooth operation $\delta_1$ and careful operation $\delta_2$ are neglected, they incur a low and high penalty that changes with any pedestrian traffic. Neglecting duties is permitted until a limit $\epsilon$. With the PFD moral principle $\rho_\Delta$, 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, $\phi : \Delta \times S \to \mathbb{R}^+$ 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

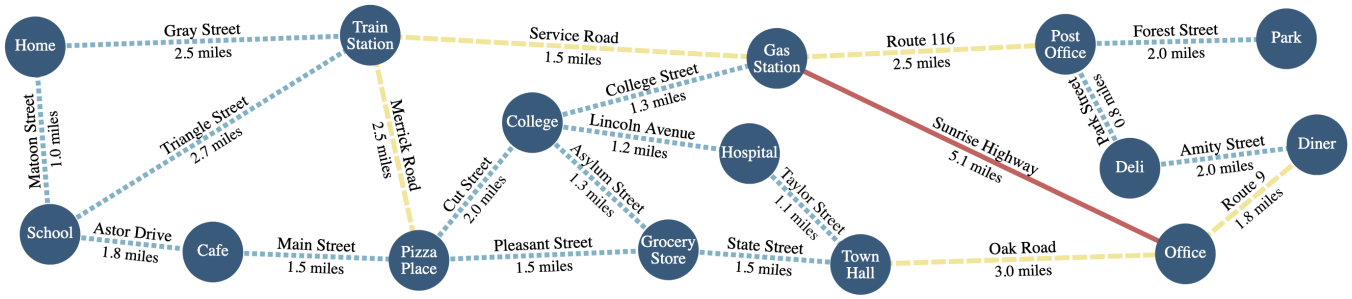


Figure 2: A city with different places connected by city streets, county roads, and highways

with heavy pedestrian traffic while 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.

## 7 Experiments

We now demonstrate that the application of moral autonomy to autonomous driving is effective in a set of simulations and a user study. In the set of simulations, an amoral self-driving vehicle and a moral self-driving vehicle that follows different ethical frameworks both 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 sec. Accelerating 10 mph requires 2 sec. 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 sec.

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.

Table 2 highlights that the price of morality incurred by the behavior of the agent is appropriate given each ethical framework. Naturally, the amoral self-driving vehicle does not incur a price of morality. The moral self-driving vehicle, however, 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 4 indicates that the behavior of the agent is correct given each ethical framework. The amoral self-driving vehicle drives the shortest route at high speed. The moral self-driving vehicle, however, 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 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 the highway road states and the set of populated location states.


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

In the user study, planning and robotics experts had to complete two tasks in a randomized order. In both tasks, developers were given a complete decision-making model for navigating efficiently around the example city and had to enforce the following moral requirements. The agent 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 achieve the desired behavior by modifying the existing decision-making model, an MDP, by changing its reward function or transition function. In the other task, developers were asked to achieve the same desired behavior but by defining the ethical context for the prima facie duties ethical framework.

Figure 5 illustrates that our method led to better policies than the other method. In our method, all policies satisfy the requirements and optimize the navigation task with exactly one violation. However, in the other method, the majority of policies fail to optimize the navigation task or even satisfy the requirements: aggressive policies in the upper right corner are faster but immoral while conservative policies in the lower left corner are slower but moral. It is also encouraging that our method (24 min) had a significantly lower mean development time than the other method (45 min).

| 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 as a percentage of the value of the optimal amoral policy for all vehicle options on each navigation task

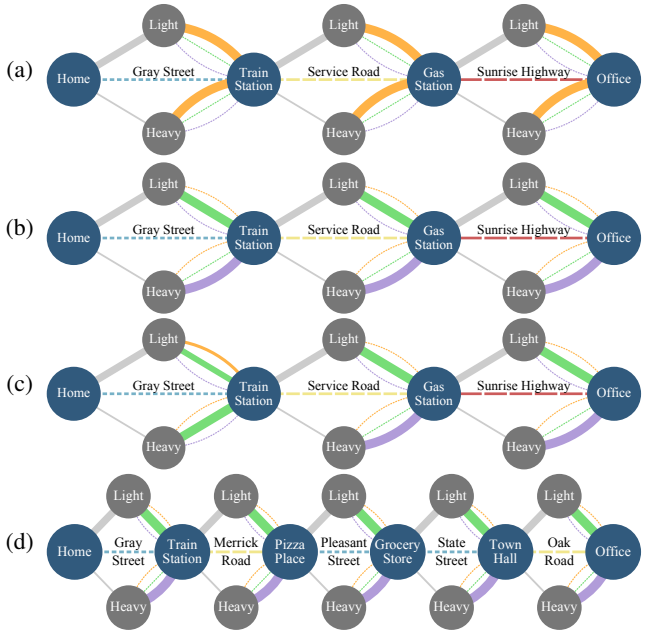


Figure 4: The optimal policies for select vehicle options on a navigation task with (a) no ethical framework, (b) DCT with $\mathcal{H} \cup \mathcal{I}$, (c) PFD with $\epsilon = 9$, and (d) VE with $\mathcal{C} \cup \mathcal{P}$. A *blue* node denotes a location and a *gray* node denotes pedestrian traffic. With a thickness for likelihood, a *gray* line denotes turning onto a road and an *orange*, *green*, or *purple* line denotes cruising at high, normal, or low speed.

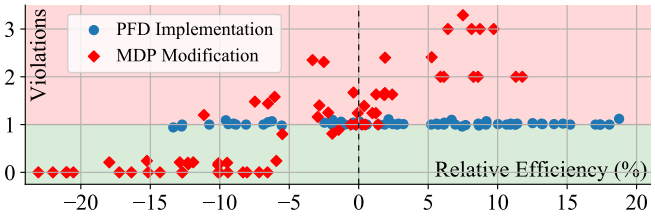


Figure 5: The results of the user study. For each exercise and location in the city, there is a point that denotes the resulting policy. For that policy, the horizontal axis is its time savings relative to the policy from the opposing exercise while the vertical axis is its number of violations. The moral and immoral regions are in *green* and *red*.

Our open source library, *Morality.js*, which is available on the website `https://www.moralityjs.com` with the customizable grid world environment dashboard seen in Figure 3, was used for all experiments [Svegliato *et al.*, 2020].

## Acknowledgments

This work was supported in part by an NSF Graduate Research Fellowship DGE-1451512 and the NSF grants IIS-1724101 and IIS-1813490.

## References

[Abel *et al.*, 2016] David Abel, James MacGlashan, and Michael L Littman. Reinforcement learning as a framework for ethical decisions. In *AAAI Workshop on AI, Ethics, and Society*, 2016.

[Adamson *et al.*, 2019] Greg Adamson, John C Havens, and Raja Chatila. Designing a value-driven future for ethical autonomous and intelligent systems. *IEEE*, 2019.

[Allen *et al.*, 2005] Colin Allen, Iva Smit, and Wendell Wallach. Artificial morality: Top-down, bottom-up, and hybrid approaches. *Ethics and Information Technology*, 2005.

[Anderson *et al.*, 2017] Michael Anderson, Susan L Anderson, and Vincent Berenz. A value driven agent: An instantiation of a case-supported principle-based behavior paradigm. In *AAAI Workshop on AI, Ethics, and Society*, 2017.

[Anscombe, 1958] Gertrude Elizabeth Margaret Anscombe. Modern moral philosophy. *Philosophy*, 1958.

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

[Atkinson and Bench-Capon, 2006] Katie Atkinson and Trevor Bench-Capon. Addressing moral problems through practical reasoning. In *International Workshop on Deontic Logic and Artificial Normative Systems*. Springer, 2006.

[Bellman, 1966] Richard Bellman. Dynamic programming. *Science*, 1966.

[Bench-Capon and Modgil, 2017] Trevor Bench-Capon and Sanjay Modgil. Norms and value based reasoning: justifying compliance and violation. *Artificial Intelligence and Law*, 2017.

[Bentham, 1789] Jeremy Bentham. An introduction to the principles of morals. *London: Athlone*, 1789.

[Berreby *et al.*, 2015] Fiona Berreby, Gauvain Bourgne, and Jean-Gabriel Ganascia. Modelling moral reasoning and ethical responsibility with logic programming. In *Logic for Programming, Artificial Intelligence, and Reasoning*. Springer, 2015.

[Bertsekas, 1997] Dimitri P Bertsekas. Nonlinear programming. *Journal of the Operational Research Society*, 1997.

[Bertsimas and Tsitsiklis, 1997] Dimitris Bertsimas and John N Tsitsiklis. *Introduction to linear optimization*. Athena Scientific Belmont, MA, 1997.

[Boden *et al.*, 2017] Margaret Boden, Joanna Bryson, Darwin Caldwell, Kerstin Dautenhahn, Lilian Edwards, Sarah Kember, Paul Newman, Vivienne Parry, Geoff Pegman, Tom Rodden, et al. Principles of robotics: Regulating robots in the real world. *Connection Science*, 2017.

[Bostrom, 2016] Nick Bostrom. Superintelligence: Paths, dangers, strategies. *Science Fiction and Philosophy: From Time Travel to Superintelligence*, 2016.

[Brey, 2004] Philip Brey. Ethical aspects of facial recognition systems in public places. *Journal of Information, Communication and Ethics in Society*, 2004.

[Bringsjord *et al.*, 2006] Selmer Bringsjord, Konstantine Arkoudas, and Paul Bello. Toward a general logicist methodology for engineering ethically correct robots. *Intelligent Systems*, 2006.

[Bringsjord *et al.*, 2011] Selmer Bringsjord, Joshua Taylor, Bram Van Heuveln, Konstantine Arkoudas, Micah Clark, and Ralph Wojtowicz. 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, 2011.

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

[Burgers and Robinson, 2017] Tobias Burgers and David RS Robinson. Networked authoritarianism is on the rise. *Sicherheit und Frieden*, 2017.

[Desai and Kroll, 2017] Deven R Desai and Joshua A Kroll. Trust but verify: A guide to algorithms and the law. *Harvard Journal of Law & Technology*, 2017.

[Dignum *et al.*, 2018] Virginia Dignum, Matteo Baldoni, Cristina Baroglio, Maurizio Caon, Raja Chatila, Louise Dennis, Gonzalo Génova, Galit Haim, Malte S Kließ, Maite Lopez-Sanchez, et al. Ethics by design: necessity or curse? In *AAAI/ACM Conference on AI, Ethics, and Society*, 2018.

[Goodman and Flaxman, 2017] Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, 2017.

[Hadfield-Menell and Hadfield, 2019] Dylan Hadfield-Menell and Gillian K Hadfield. Incomplete contracting and AI alignment. In *AAAI/ACM Conference on AI, Ethics, and Society*, 2019.

[Hursthouse, 1999] Rosalind Hursthouse. *On virtue ethics*. Oxford University Press, 1999.

[Idziak, 1979] Janine Marie Idziak. *Divine command morality*. Edwin Mellen Press, 1979.

[Introna and Wood, 2004] Lucas Introna and David Wood. Picturing algorithmic surveillance: The politics of facial recognition systems. *Surveillance & Society*, 2004.

[Kant and Schneewind, 2002] Immanuel Kant and Jerome B Schneewind. *Groundwork for the metaphysics of morals*. Yale University Press, 2002.

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

[Manne, 1960] Alan S Manne. Linear programming and sequential decisions. *Management Science*, 1960.

[Mill, 1895] John Stuart Mill. *Utilitarianism*. Longmans, Green and Company, 1895.

[Moor, 2006] James H Moor. The nature, importance, and difficulty of machine ethics. *Intelligent Systems*, 2006.

[Morreau, 1996] Michael Morreau. Prima facie and seeming duties. *Studia Logica*, 1996.

[Mouaddib *et al.*, 2015] Abdel-Illah Mouaddib, Laurent Jeanpierre, and Shlomo Zilberstein. Handling advice in mdps for semi-autonomous systems. In *ICAPS Workshop on Planning and Robotics*, Jerusalem, Israel, 2015.

[Nashed *et al.*, 2018] Samer B Nashed, David M Ilstrup, and Joydeep Biswas. Localization under topological uncertainty for lane identification of autonomous vehicles. In *IEEE International Conference on Robotics and Automation*, 2018.

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

[Quinn, 2013] Philip L Quinn. Divine command theory. *The Blackwell Guide to Ethical Theory*, 2013.

[Raymond and Shackelford, 2013] Anjanette H Raymond and Scott J Shackelford. Technology, ethics, and access to justice: should an algorithm be deciding your case. *Michigan Journal International Law*, 2013.

[Robertson *et al.*, 2019] Lindsay J Robertson, Roba Abbas, Gursel Alici, Albert Munoz, and Katina Michael. Engineering-based design methodology for embedding ethics in robots. *IEEE*, 2019.

[Ross, 1930] William D Ross. *The right and the good*. Oxford University Press, 1930.

[Scherer, 2015] Matthew U Scherer. Regulating artificial intelligence systems: Risks, challenges, competencies, and strategies. *Harvard Journal of Law & Technology*, 2015.

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

[Shaw *et al.*, 2018] Nolan P Shaw, Andreas Stöckel, Ryan W Orr, Thomas F Lidbetter, and Robin Cohen. Towards provably moral AI agents in bottom-up learning frameworks. In *AAAI/ACM Conference on AI, Ethics, and Society*, 2018.

[Shim *et al.*, 2017] Jaeeun Shim, Ronald Arkin, and Michael Pettinatti. An intervening ethical governor for a robot mediator in patient-caregiver relationship. In *IEEE International Conference on Robotics and Automation*, 2017.

[Svegliato *et al.*, 2019] Justin Svegliato, Kyle Hollins Wray, Stefan J Witwicki, Joydeep Biswas, and Shlomo Zilberstein. Belief space metareasoning for exception recovery. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2019.

[Svegliato *et al.*, 2020] Justin Svegliato, Samer Nashed, and Shlomo Zilberstein. An integrated approach to moral autonomous systems. In *24th European Conference on Artificial Intelligence*, 2020.

[Taylor *et al.*, 2016] Jessica Taylor, Eliezer Yudkowsky, Patrick LaVictoire, and Andrew Critch. Alignment for advanced machine learning systems. *Machine Intelligence Research Institute*, 2016.

[van der Torre, 2003] Leendert van der Torre. Contextual deontic logic: Normative agents, violations and independence. *Annals of Mathematics and Artificial Intelligence*, 2003.

[Vanderelst and Winfield, 2018] Dieter Vanderelst and Alan Winfield. An architecture for ethical robots inspired by the simulation theory of cognition. *Cognitive Systems Research*, 2018.

[Winfield *et al.*, 2014] Alan FT Winfield, Christian Blum, and Wenguo Liu. Towards an ethical robot: internal models, consequences and ethical action selection. In *Conference Towards Autonomous Robotic Systems*. Springer, 2014.

[Wooldridge and Van Der Hoek, 2005] Michael Wooldridge and Wiebe Van Der Hoek. On obligations and normative ability: An analysis of the social contract. *Journal of Applied Logic*, 2005.

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