An Integrated Approach to Moral Autonomous Systems

Justin Svegliato[1] and **Samer Nashed**[1] and **Shlomo Zilberstein**[1]

Abstract. The prevailing methodology for integrating decision making and ethics is to modify autonomous systems in an ad hoc way to incorporate moral sensibility. However, these provisional modifications often lead to behavior that jeopardizes the intentions of developers or the values of stakeholders. We propose a novel approach for building moral autonomous systems that optimally completes a task and follows an ethical framework by decoupling ethical compliance from task completion. This paper offers a formal definition of our approach along with its key properties, an example based on prima facie duties, and a demonstration that uses our open source library.

1 INTRODUCTION

Autonomous systems have traditionally operated without any moral sensibility in the domain of operation. Although there have been attempts to build autonomous systems with moral sensibility, the prevailing methodology relies on making ad hoc adjustments until the desired behavior is produced. For example, a self-driving vehicle with a reward function that encourages completing a route efficiently can be tweaked to discourage inconsiderate or even hazardous driving. However, because these provisional modifications blend task completion and ethical compliance incommensurably, the resulting behavior is often unpredictable. As a result, these systems eventually encounter unanticipated scenarios that lead to behavior that fails to reflect the intentions of developers or the values of stakeholders [8].

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*. However, instead of making ad hoc adjustments to this decision making model, the system follows an ethical framework by adhering to a *moral principle* in an *ethical context*. Such a system can formally be expressed as an optimization problem with an objective function that is constrained by a set of constraints that represents the task and an extra constraint that operationalizes the ethical framework [1]. An optimal solution to the optimization problem is a policy that optimizes completing the task while following the ethical framework.

Our approach offers several benefits. First, it is *general-purpose* because it supports any ethical framework as long as it can be represented appropriately. Second, it is *modular* in that it enables the ethical context and moral principle of any ethical framework to be interchanged with the decision making model of any task. Third, it is *interpretable* since it defines an ethical framework in terms of the behavior and environment of an autonomous system. Fourth, it is *explicit* given that it avoids a common objective that conflates completing a task and following an ethical framework, which reduces implicit value judgments that can harm the design of a moral autonomous system. Our approach is therefore a practical tool that helps engineers and ethicists implement moral autonomous systems.

[1] University of Massachusetts Amherst, USA, email: {jsvegliato, snashed, shlomo}@cs.umass.edu

2 MORAL AUTONOMOUS SYSTEMS

In our approach, a moral autonomous system is an agent who completes a task by using a *decision making model* and follows an ethical framework by adhering to a *moral principle* in an *ethical context*.

The *decision making model* is an *amoral*, *descriptive* 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 for navigating a city [7]. An engineer must select a representation for the decision making model that reflects the properties of the task. While a Markov decision process (MDP) is used for tasks with full observability in this paper, a stochastic shortest path (SSP) problem or a partially observable MDP (POMDP) can be used for tasks with start and goal states or partial observability instead.

The *ethical context* is a *moral*, *prescriptive* model 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 discourteous and reckless driving that permit speeding on a highway in some scenarios but never in a school zone or near a crosswalk [9]. 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 specified as a tuple of values, sets, and functions, the specification of the tuple depends on the ethical framework.

The *moral principle* evaluates the morality of a policy of the decision making model in the ethical context based on the information for 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* [5] or universalize to the moral community without contradiction in *Kantianism* [3]. Given a decision making model and an ethical context, a moral principle can be expressed as a function that maps a policy of the decision making model to its moral status in the ethical context.

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 in an **ethical context** $\mathcal{E}$.*

We now offer a formal description of a moral autonomous system and its goal of finding an optimal policy that not only completes its task but also follows its ethical framework below.

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 ρ in an ethical context $\mathcal{E}$.*

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^π subject to a moral principle $\rho(\pi)$ in the following optimization problem.*

$$\underset{\pi \in \Pi}{\text{maximize}} \; V^\pi \quad \text{subject to } \rho(\pi)$$

Note that standard autonomous systems use an *optimal amoral policy*, $\pi^* \in \Pi$, that considers the task without any ethical framework.

Following an ethical framework may adversely impact completing the task of a moral autonomous system. Engineers and ethicists can measure the cost of this impact on the system in the following way.

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

Following an ethical framework may even prevent completing the task of a moral autonomous system. 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. We formalize a notion of feasibility below.

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**.*

We find an optimal moral policy of a moral autonomous system by solving the optimization problem 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 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 [4]. For ethical compliance, the program has an extra moral constraint $c_\rho(\mu)$ that represents the moral principle $\rho(\mu)$ given a matrix of occupancy measures μ. Formally, an optimal moral policy, $\pi^*_\rho \in \Pi$, is found by solving the following program.

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

Note that an MDP can be represented by a tuple $\langle S, A, T, R, d\rangle$ with a set of states S, a set of actions A, a transition function T, a reward function R, and a start state function d given a discount factor γ [2].

3 PRIMA FACIE DUTIES

We offer an example of a moral autonomous system that can complete any task while following an ethical framework influenced by *prima facie duties* [6]. Prima facie duties, a pluralistic, nonabsolutist ethical theory, holds that the morality of an action depends on whether that action fulfills fundamental moral duties that can contradict each other. We consider an 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 below.

Definition 6. *A **prima facie duties ethical context**, $\mathcal{E}_\Delta$, can be represented by a tuple, $\mathcal{E}_\Delta = \langle \Delta, \phi, \tau\rangle$, where*

- *Δ 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 7. *A **prima facie duties moral principle**, ρ_Δ, 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:*

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

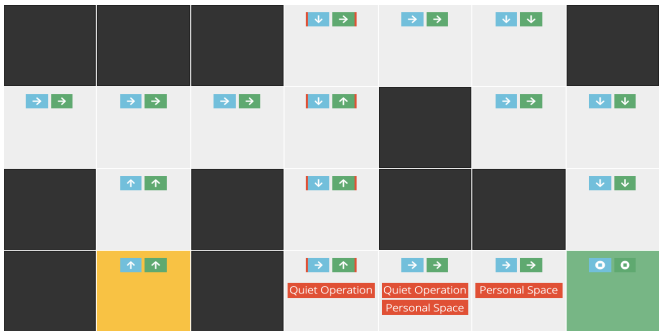

Figure 1. A demonstration of a moral autonomous system that navigates a library with the prima facie duties *Quiet Operation* and *Personal Space*

4 DEMONSTRATION

We demonstrate a moral autonomous system that navigates a library with prima facie duties. The decision making model represents the library as a grid world in which an agent must go from a start square to a goal square by moving in a cardinal direction with a small chance of slipping in an orthogonal direction. The ethical context and moral principle represent the prima facie duties of *Quiet Operation* and *Personal Space* that prevent being too load or too close to a human.

Figure 1 is an illustration of the moral autonomous system taken from the customizable dashboard of our open source *Morality.js* library.[2] Here, the system takes the longer but moral path instead of the immoral but shorter path from the *orange* start square to the *green* goal square, with a price of morality of 12.7 (15.2%), in order to avoid being too load or too close to a human. Note that the optimal amoral policy is in *blue* while the optimal moral policy is in *green* with *red* tags that highlight any difference between each policy.

5 CONCLUSION

We propose an integrated approach for building moral autonomous systems that optimize completing a task while following an ethical framework based on decoupling ethical compliance from task completion. This enables engineers and ethicists to work together to build moral autonomous systems that are general-purpose, modular, interpretable, and explicit. Future work will develop ethical frameworks for moral autonomous systems influenced by traditional ethical theories, including utilitarianism, natural law theory, and Kantianism.

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

REFERENCES

[1] Ronald C Arkin, 'Governing lethal behavior: Embedding ethics in a hybrid deliberative/reactive robot architecture', in *HRI*. ACM, (2008).
[2] Richard Bellman, 'Dynamic programming', *Science*, (1966).
[3] Immanuel Kant and Jerome B Schneewind, *Groundwork for the metaphysics of morals*, Yale University Press, 2002.
[4] Alan S Manne, 'Linear programming and sequential decisions', *Management Science*, (1960).
[5] John Stuart Mill, *Utilitarianism*, Longmans, Green and Company, 1895.
[6] David Ross and William David Ross, *The right and the good*, Oxford University Press, 2002.
[7] Justin Svegliato, Kyle Hollins Wray, Stefan J Witwicki, Joydeep Biswas, and Shlomo Zilberstein, 'Belief space metareasoning for exception recovery', in *IROS*, (2019).
[8] Jessica Taylor, Eliezer Yudkowsky, Patrick LaVictoire, and Andrew Critch, 'Alignment for advanced ML systems', *MIRI*, (2016).
[9] Dieter Vanderelst and Alan Winfield, 'An architecture for ethical robots inspired by the simulation theory of cognition', *CSR*, (2018).

[2] https://www.moralityjs.com