Ethically Compliant Planning within Moral Communities

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

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

Abstract

Ethically compliant autonomous systems (ECAS) are the state-of-the-art for solving sequential decision-making problems under uncertainty while respecting constraints that encode ethical considerations. This paper defines a novel concept in the context of ECAS that is from moral philosophy, the *moral community*, which leads to a nuanced taxonomy of explicit ethical agents. We then propose new ethical frameworks that extend the applicability of ECAS to domains where a moral community is required. Next, we provide a formal analysis of the proposed ethical frameworks and conduct experiments that illustrate their differences. Finally, we discuss the implications of explicit moral communities that could shape research on standards and guidelines for ethical agents in order to better understand and predict common errors in their design and communicate their capabilities.

Introduction

Researchers do not yet fully understand how automated intelligent systems produce or exacerbate different types of harms and how to prevent these harms in the first place. Enabling automated decision-making systems to comply with ethical theories shows some promise, but these theories are still challenging to implement despite extensive study in moral philosophy. Currently, ethically compliant autonomous systems (ECAS) (Svegliato, Nashed, and Zilberstein 2021) represent the state-of-the-art for applying ethical constraints to agent behavior. ECAS work by augmenting mathematical programs representing decision processes with an additional, independent constraint that enforces compliance with a moral principle based roughly on an ethical theory. Solving the program produces behavior that is guaranteed to comply with the constraints of the ethical theory. However, many popular ethical theories which require explicit consideration of multiple agents simultaneously, such as utilitarianism or contractarianism, have yet to be explored in this context. Here, we extend recent work on ECAS to ethical theories that require explicit moral communities, which are considerably more complex.

In moral philosophy, the *moral community* is the set of agents considered during ethical decision making. All ethical theories define moral communities either implicitly or explicitly. For example, utilitarianism requires a set of agents to be enumerated explicitly whose expected future utility is collectively maximized, while an agent using *prima facie duties* implicitly defines its moral community based on which duties are included and how their relevance to a given situation is determined. The cost of generating constraints such that agent behavior adheres to moral principles that explicitly reason about moral communities can in some cases scale exponentially with respect to the size of the moral community. In practice, an agent may approximate the effects of its actions on members of the moral community through information provided in its ethical context. We propose several new ethical frameworks that can be used in ECAS, modeled on act utilitarianism, the veil of ignorance, and the Golden Rule, which involve more complex ethical contexts to tractably represent these ethical theories. We follow previous work that uses Markov Decision Processes (MDPs) to illustrate how different ethical frameworks may be applied to decision processes. These frameworks are evaluated in a simulated environment, where we explore differences in behavior between agents following different ethical frameworks. We find that frameworks vary in which policies they prefer, theoretical and practical compute requirements, reliance on different parts of their ethical contexts, and even the existence of a solution in a given scenario.

The scalability of moral principles that reason explicitly about moral communities also raises several fundamental questions about how work on ethical decision making should proceed. When is it acceptable to approximate ethical frameworks given that their purpose is to guard against other forms of corner-cutting in the development process? When is it acceptable to rely on implicitly defined moral communities? How do we understand model-level and design-level approximations in the context of ethically compliant systems? How can interdisciplinary research help us understand the effects of model fidelity on ethical decision making? We provide insights into these questions and analyze the differences between ethical frameworks and their applicability to various deployment contexts.

This paper offers four contributions. First, we define moral communities within the context of ECAS and show how this definition clarifies requirements and capabilities of intelligent ethical agents. Second, we define several new ethical frameworks within ethically compliant autonomous systems. Third, we provide a complexity analysis of these new frameworks. Last, we present an extensive discussion of the potential implications of the complexity of some popular ethical theories and insights regarding how implicit and explicit moral communities affect dominant ideas and practices surrounding development of ethical reasoning systems.

Related Work

The application of moral or ethical reasoning to automated systems at conception, regulation, design, and deployment is a broad and nuanced field of research. This paper builds on the ECAS framework (Svegliato, Nashed, and Zilberstein 2021), and those seeking a holistic treatment of the literature should look there. Surveys of technical approaches also exist (Yu et al. 2018). Here, we focus on work that enforces moral or ethical behavior in multi-agent systems explicitly in a top-down manner—a topic of broad interest (Yilmaz, Franco-Watkins, and Kroecker 2017; Rossi and Mattei 2019; Murukannaiah et al. 2020; Morgan et al. 2020).

Most research in this area uses various logic systems. For example, systems based on deontic logic (van der Torre 2003; Bringsjord, Arkoudas, and Bello 2006) or temporal logic (Wooldridge and Van Der Hoek 2005; Atkinson and Bench-Capon 2006) have both been proposed for prescribing ethical agent behavior. Some methods even use a form of metareasoning over a set of logics (Bringsjord et al. 2011). Recently, methods based on Answer Set Programming have been proposed (Berreby, Bourgne, and Ganascia 2015), including some that focus on modeling interactions between agents in addition to the effect of individual agents' actions (Cointe, Bonnet, and Boissier 2016a).

Related research has proposed reasoning systems that impose semantic ordering over logical statements, including Belief-Desire-Intention architectures (Cointe, Bonnet, and Boissier 2016b) and case-supported principle-based behavior models (Anderson and Anderson 2015). Programming languages for multi-agent systems that support ethical concepts like sanctioning an agent and representing an action's effects on other agents have also been proposed (Dastani, Tinnemeier, and Meyer 2009). Other systems combine human oversight with logical or rule-based constraints representing ethical behavior, for example within ethical mission execution automata (Brutzman et al. 2018). Logic-based systems are attractive for several reasons, including their interpretability and their accessibility to theoretical tools and guarantees. However, they also have significant drawbacks. Nuanced behavior can become difficult to specify as agent capabilities increase, and deploying such systems in stochastic environments presents challenges that are still unsolved (Abel, MacGlashan, and Littman 2016).

Not all ethical reasoning systems are based on logic systems. Some research models ethical behavior using game-theoretic concepts (Conitzer et al. 2017), but this strategy has yet to be widely adopted. Generating ethical behavior in reinforcement learning agents (Thomas et al. 2019) and reward shaping using human moral exemplars (Wu and Lin 2017) could be considered motivational analogs to ECAS. However, while prior systems generate policies from MDPs and produce ethical constraints independently from task constraints, unlike ECAS, they ultimately cannot produce guarantees since they mix reward for task completion with reward for ethical compliance. In this work, we take the view that human error in design is a source of significant ethical risk (Etzioni and Etzioni 2017). We build on strengths of ECAS, emphasizing better human-computer design interface rather than larger data sets or smarter algorithms.

Background

Markov Decision Processes A *Markov decision process* (MDP) is a decision-making model for reasoning in fully observable, stochastic environments (Bellman 1952) that has broad applicability, including rescue robots (Goodrich et al. 2008; Pineda et al. 2015), planetary rovers (Mustard, Beaty, and Bass 2013; Gao and Chien 2017), and autonomous vehicles (Svegliato et al. 2019; Basich et al. 2020; Nashed et al. 2021). An MDP can be described as a tuple $\langle S, A, T, R, d \rangle$. S is a finite set of states, where $s \in S$ may be expressed in terms of a set of *state factors*, $\langle f_1, f_2, \ldots, f_N \rangle$, such that s indexes a unique assignment of variables to the factors f; 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 π induces a value function $V : S \to \mathbb{R}$ representing the expected discounted cumulative reward $V(s) \in \mathbb{R}$ 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 $V(s) = \max_{a \in A} \sum_{s' \in S} T(s, a, s')[R(s, a, s') + \gamma V(s')]$.

One approach for calculating an optimal policy expresses the optimization problem as a linear program in either the primal form or the dual form (Manne 1960). This paper proposes several ethical frameworks, some of which naturally map to the primal form and others to the dual form. The primal form minimizes a set of value variables V_s for the value $V(s)$ of each state $s \in S$ subject to a set of constraints that maintain the Bellman optimality equation.

$$\begin{aligned} \min_{V} \quad & \sum_{s \in S} d(s) V_s \\ \text{s.t.} \quad & V_s \geq \sum_{s' \in S} T(s, a, s') \left[R(s, a, s') + \gamma V_{s'} \right] \quad \forall s, a \end{aligned}$$

The dual form maximizes a set of occupancy measures μ_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 non-negative occupancy.

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

Given the solution as V (primal form) or μ (dual form), the optimal policy $\pi(s)$ can be calculated as follows.

$$\pi(s) = \arg\max_{a \in A} \left[\sum_{s' \in S} T(s, a, s') \left[R(s, a, s') + \gamma V_{s'} \right] \right]$$

$$\pi(s) = \arg\max_{a \in A} \mu_a^s$$

Ethically Compliant Autonomous Systems An *ethically compliant autonomous system* (ECAS) has a decision-making model for completing its *task* and an ethical context and a moral principle for following its *ethical framework* (Svegliato, Nashed, and Zilberstein 2021). A *decision-making model* $\mathcal{D}$ describes the information needed to complete the task. An *ethical context* $\mathcal{E}$ describes the information required to follow the ethical framework. A *moral principle* $\rho : \quad \rightarrow \mathsf{B}$ evaluates the morality of a policy for the decision-making model within the ethical context.

Definition 1. *An ECAS,* $\langle \mathcal{D}, \mathcal{E}, \rho \rangle$*, optimizes completing a task by using a decision-making model* $\mathcal{D}$ *while following an ethical framework by adhering to a moral principle* $\rho : \rightarrow \mathsf{B}$ *within an ethical context* $\mathcal{E}$*.*

The objective of an ECAS is to find an optimal policy subject to following an ethical framework.

Definition 2. *The objective of an ECAS is to find an* ***optimal moral policy****,* $\pi \in$ *, by solving for a policy* $\pi \in$ *within the space of policies that maximizes a value function* V *subject to a moral principle* ρ *in the optimization problem.*

$$\underset{2}{\text{maximize}} \quad V \qquad \text{subject to} \quad \rho(\pi)$$

An ECAS can follow an ethical framework that impacts completing its task. We define this impact as the maximum difference across all states between the value functions of the optimal moral policy and optimal amoral policy.

Definition 3. *Given the optimal moral policy* $\pi \in$ *and the optimal amoral policy* $\pi \in$ *, the* ***price of morality****,* ψ*, can be represented by the expression* $\psi = \|V^{\rho} - V \quad \|_1$*.*

There are many possible alternative definitions of ψ, including, for example, the mean or median difference across states, or the sum of all differences across states. We chose the maximum difference because it is conservative in the sense that it captures the maximum change, and it is sensitive to outliers as affected states will not be obscured by an aggregate statistic. These two properties are key to realizing potential impacts of following a given ethical framework.

An ECAS can follow an ethical framework incompatible with completing its task. Its feasibility depends on whether a solution exists to the optimization problem.

Definition 4. *An ECAS is* ***realizable*** *if and only if there exists a policy* $\pi \in$ *such that its moral principle* $\rho(\pi)$ *is satisfied. Otherwise, it is* ***unrealizable****.*

Lane Merging Example Suppose an autonomous vehicle is in the process of merging lanes, say from two lanes to one. To use ECAS, we first choose a moral principle to follow during the lane merging process, which describes a property that the policy must satisfy. For instance, a utilitarian moral principle might require policies to minimize the total expected drive time for all agents, rather than just the agent making the decision. Given a moral principle, we define the ethical context, which contains the information for evaluating the moral principle. In the utilitarian example, the ethical context requires both models of other agents, as well as models of how our own agent's actions effect other agents' state, which are required to reason about the impact of our agent's actions. Other moral principles may require simpler contexts, such as sets of forbidden states or penalty functions for violating certain norms in certain states.

Moral Communities

The *moral community* is a concept from moral philosophy that defines the set of entities with moral considerability. Such entities should have their welfare resulting from a given action taken into account when considering whether or not to take that action. In many forms of utilitarianism, all humans and many animals are members of the moral community because they are sentient or conscious (Frey 2011). Other ethical theories consider only those represented in negotiation of a social contract (Froese 2001) or those capable of logical reasoning (Rawls 1980). We use the broad term entities because there are serious arguments for the moral considerability of inanimate bodies (Brennan 1984). In general, membership in a moral community depends on the definition of the ethical theory and is contested in both normative and applied ethics (Cahen 1988; Lomasky 1990; Birch 1993; Bernstein 1998; Bagnoli 2007; Caton 2020).

In fact, all decision-making systems with ethical reasoning components define moral communities, even if they are not made explicit. For example, the moral community is clear in implementations of utilitarianism since they require an explicit set of agents to calculate the overall welfare of those agents when choosing its action. However, an agent using *prima facie duties* to choose ethical actions implicitly defines its moral community via several design decisions. These include which duties are included and how their relevance to a given situation is determined. For instance, one duty might be to always tell the truth, which implicitly constrains the moral community to entities that can communicate. In some scenarios, more specific duties might be defined, such as maintaining lane membership in an autonomous vehicle. This duty is important for the safety of the driver, passengers, and other motorists. However, this duty by itself may fail to reflect the preferences of many stakeholders in the roadway system such as cyclists, pedestrians, construction workers, parked or loading vehicles, or emergency vehicles. Implicit moral communities are not categorically better or worse than explicit ones, but we should be careful to understand tradeoffs and vulnerabilities created by how we represent moral communities.

Early work on ethical autonomous systems made a distinction between autonomous systems that satisfy moral requirements purely through careful construction, called *implicit ethical agents*, and those capable of moral reasoning, called *explicit ethical agents* (Moor 2006). Subsequent research has highlighted the importance of explicit ethical reasoning (Bench-Capon and Modgil 2017; Dignum et al. 2018). In the context of ECAS, there is a similar dichotomy when considering the moral community. All ECAS are explicit ethical agents, but which entities they reason over may be specified implicitly, explicitly, or some combination. Here, we formally define the moral community and use this definition to illuminate several sub-classes of ECAS.

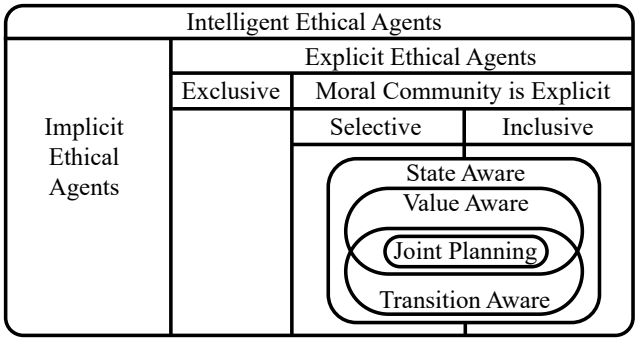

Figure 1: A taxonomy of intelligent ethical agents. Value-aware and transition-aware ethical agents are necessarily state-aware.

Definition 5. *A **moral community** is a set of agents $\mathcal{I} = \{1, 2, \ldots, N\}$ that have moral considerability with respect to the operation of an ECAS.*

Definition 6. *A **moral community model** is a set of tuples $\mathcal{M} = \{(S_1, V_1), (S_2, V_2), \ldots, (S_M, V_M)\}$ such that each tuple (S_i, V_i) has a state space S_i and a value function V_i for each agent i within a subset of the moral community $\hat{\mathcal{I}} \subseteq \mathcal{I}$.*

Definition 7. *An **inclusive** ECAS has an ethical context $\mathcal{E} = \langle e_1, \ldots, e_\cdot \rangle$ such that there exists an attribute e that is a moral community model $\mathcal{M}$ where $|\mathcal{I}| = |\hat{\mathcal{I}}|$.*

Definition 8. *An **exclusive** ECAS has an ethical context $\mathcal{E} = \langle e_1, \ldots, e_\cdot \rangle$ that does not contain a moral community model.*

Definition 9. *A **selective** ECAS is not inclusive or exclusive.*

These definitions extend an existing taxonomy of intelligent ethical agents, producing the classes shown in Figure 1. An agent's taxonomic class depends on whether it reasons explicitly about the ethics of its actions (*explicit ethical agents*) or not (*implicit ethical agents*). Explicit ethical agents can be further sub-divided into those that use explicit models of members of their moral community (*selective* and *inclusive*) and those that do not (*exclusive*). Selective and inclusive agents may use a variety of different information. This includes the possible states of moral community members (*state-aware*), the state-dependent welfare, or value, of members (*value-aware*), or a model for how the agent's own actions might affect the state of other members (*transition-aware*). These model types may overlap, and there may be other types of models that describe members of the moral community but do not use states, values, or transitions. The most integrated versions of such models result in multi-agent planning problems. These models are the most accurate but are also typically prohibitively expensive.

The frameworks presented in this paper rely on explicit moral communities, which have several benefits. First, constraints generated from explicit moral communities are often more accurate approximations of the real world since models can be customized for individual entities. This allows more nuanced and individualized decision making. Second, developers are less likely to forget an individual or a class of stakeholders when required to represent them explicitly within the ethical context. Third, explicitly enumerating and modeling every stakeholder can uncover implicit assumptions that, if unaddressed, could cause unintended harm.

However, explicit moral communities are not a panacea. Constructing constraints for moral principles that use explicit moral communities is naturally computationally expensive. Moreover, explicit ethical contexts often place a substantially higher burden on engineers who must design models for not only the decision-making agent but other agents as well. Nonetheless, we believe ethical decision making, and ECAS in particular, can benefit from ethical frameworks that use explicit moral communities.

Lane Merging Example Roadway systems are complex, with many different stakeholders. The moral community for this system could reasonably include all motorists, bicyclists, pedestrians, construction workers, and emergency vehicles. It may also include entities that depend on functioning roadways indirectly, such as businesses. It is up to the developer to decide which types of entities to model and how to model them. For simplicity, our examples include only other motorists within our moral community.

Ethical Frameworks

In this section, we present a set of simplified ethical frameworks used to partially define ECAS. Each ethical framework *approximates* a well-known ethical theory in moral philosophy (Shafer-Landau 2009). Their purpose is to encode an ethical theory in a tractable way, acknowledging that they do not capture all nuances of the ethical theories on which they are based. We encourage work on more complex ethical frameworks that reflect the depth of different ethical theories, including extensions to those presented here.

The Veil of Ignorance

The Veil of Ignorance (VOI), a concept proposed by John Rawls in his theory of a fair and just society, states that an agent should make decisions by acting as if they are deprived of knowledge of their personal circumstances (Rawls 2009). That is, it holds that an action is moral based on whether an agent would perform that action if it ignored its own personal circumstances. In an MDP, an agent's circumstance is completely described by the values of its current state factors. An agent's personal circumstance may be captured by a subset of state factors. We consider an ethical framework that requires a policy to select actions that ensure a bounded difference between the value of the ECAS policy in a given scenario and the corresponding value of all other agent policies in the same scenario after ignoring veiled state factors.

Definition 10. *The **Veil of Ignorance ethical context**, $\mathcal{E}_V$, is represented by a tuple $\mathcal{E}_V = \langle \mathcal{M}, \mathcal{V}, \tau \rangle$:*

- *$\mathcal{M} = \{(S_1, V_1), (S_2, V_2), \ldots, (S_M, V_M)\}$ is a **moral community model**: each tuple (S_i, V_i) has a state space S_i and a value function V_i for each agent i within a subset of the moral community $\hat{\mathcal{I}} \subseteq \mathcal{I}$.*
- *$\mathcal{V} = \{1, 2, \ldots, \ell\}$ is a **veil of ignorance** such that each index $v \in \mathcal{V}$ is an index of a state factor within the veil of ignorance.*
- *$\tau \in \mathbb{R}^+$ is a **tolerance**.*

Definition 11. *The **Veil of Ignorance moral principle**, ρ_V, is expressed as the following equation:*

$$\rho_V(\pi) = \bigwedge_{i \in M} \bigwedge_{s \in S} \bigwedge_{s_i \in S_i} \left[s \sim s_i \implies |V(s) - V_i(s_i)| \leq \tau \right].$$

*The **veil equivalence operator**, $s \sim s_i \doteq \wedge_{v \in \mathcal{V}}[s[v] = s_i[v]]$, is true if a state $s = \langle f^1, f^2, \ldots, f^n \rangle$ of an ECAS and a state $s_i = \langle f_i^1, f_i^2, \ldots, f_i^n \rangle$ of an agent $i \in \mathcal{I}$ have identical state factor values for each state factor not within the veil of ignorance $\mathcal{V}$ and false otherwise.*

Transition Awareness

The Veil of Ignorance ethical context is an example of an ethical context that is both state-aware and value-aware. However, it is not transition-aware. Given a particular state and action, an agent using a transition-aware ethical context can reason explicitly and probabilistically about the likely resultant states of other agents should it take a given action. This type of model is useful for explainability and ascribing intentionality. Humans often consider intentionality when determining the morality of an action. Although they exhibit several quirks of reasoning regarding intentionality (Young et al. 2006; Leslie, Knobe, and Cohen 2006; Guglielmo and Malle 2010), these concepts are still key to determining liability or criminality in some cases. Transition awareness allows automated systems to assume an equivalent responsibility since we could inspect a given transition model and derive whether a system had full knowledge of possible consequences of an action. Here, we present an example of a transition-aware ethical context and show how it can be used to evaluate additional types of moral principles.

Definition 12. *A **transition-aware ethical context**, $\mathcal{E}_F$, is represented by a tuple $\mathcal{E}_F = \langle \mathcal{M}, \mathcal{F}, \mathcal{P}, \tau \rangle$:*

- $\mathcal{M} = \{(S_1, V_1), (S_2, V_2), \ldots, (S_M, V_M)\}$ *is a **moral community model**: each tuple (S_i, V_i) has a state space S_i and a value function V_i for each agent i within a subset of the moral community $\hat{\mathcal{I}} \subseteq \mathcal{I}$.*
- $\mathcal{F} = \{f_1, f_2, \ldots, f_n\}$ *is a set of **impact functions** such that a function $f_i : S \times S \times S_i \times S_i \to [0, 1]$ yields the probability that a transition from a state $s \in S$ to a successor state $s' \in S$ for the agent will cause a transition from a state $s_i \in S_i$ to a successor state $s_i' \in S_i$ for an agent i within a subset of the moral community $\hat{\mathcal{I}} \subseteq \mathcal{I}$.*
- $\mathcal{P} = \{p_1, p_2, \ldots, p_m\}$ *is a set of **correspondence functions** such that a function $p_i : S \times S_i \to [0, 1]$ yields the probability that an agent i within a subset of the moral community $\hat{\mathcal{I}} \subseteq \mathcal{I}$ is in a state $s_i \in S_i$ given that the agent is in a state $s \in S$.*
- $\tau \in \mathbb{R}^+$ *is a **tolerance**.*

Given a transition-aware ethical context, we define two quantities. First, given an ECAS in a state $s \in S$ performing an action $a \in A$, the **future expected value**, $V_i^a(s)$, for an agent i in the moral community $\hat{\mathcal{I}} \subseteq \mathcal{I}$ is expressed as

$$V_i^a(s) = \sum_{s_i \in S_i} p_i(s, s_i) \sum_{s' \in S} T(s, a, s') \sum_{s_i' \in S_i} f_i(s, s', s_i, s_i') V_i(s_i').$$

Second, the **current expected value**, $\hat{V}_i(s)$, for an agent i in the moral community $\hat{\mathcal{I}} \subseteq \mathcal{I}$ is expressed as

$$\hat{V}_i(s) = \sum_{s_i \in S_i} p_i(s_i | s) V_i(s_i).$$

We now offer two examples of ethical frameworks that use these value functions to define their moral principles.

The Golden Rule *The Golden Rule* (GR), a classic test of morality, states that an agent should treat other agents as that agent would want to be treated (Wattles 1966). Namely, it holds that an action is moral based on whether an agent would want all other agents to perform that action on that agent. We consider a moral principle that requires a policy to select actions that do not decrease the value of all agents by more than some tolerance.

Definition 13. *The **Golden Rule moral principle**, ρ_G, is expressed as the following equation:*

$$\rho_G(\pi) = \bigwedge_{s \in S} \bigwedge_{i \in M} \left[\hat{V}_i(s) - V_i^{(s)}(s) \leq \tau \right].$$

Act Utilitarianism *Act Utilitarianism* (AU), proposed by Jeremy Bentham and John Stuart Mill in the 19th century, states that an agent should make decisions that maximize the overall well-being of society (Bentham 1789; Mill 1895). In short, it holds that an action is moral if that action maximizes the overall utility of all agents. We consider a moral principle that requires a policy to select actions that maximize the value of all agents within some tolerance.

Definition 14. *The **Act Utilitarian moral principle**, ρ_U, is expressed as the following equation:*

$$\rho_U(\pi) = \bigwedge_{s \in S} \left[\pi(s) \in \arg\max_{a \in A} \sum_{i \in M} V_i^a(s) \right].$$

*The **utility maximization operator**, $\arg\max_{a \in A}$, returns the set of actions that induce a sum of the future expected values for all agents, $\sum_{i \in M} V_i^a(s)$, within a tolerance τ of the maximum sum over the future expected values $\max_{a \in A} \sum_{i \in M} V_i(s)$.*

All moral principles offered here use a tolerance, τ, to achieve flexibility during policy generation similar to the concept of slack (Wray, Zilberstein, and Mouaddib 2015). In these moral principles, τ is additive and thus its scale is meaningful relative to the scale of the value functions within the moral community. In general, models of moral community members may not contain value functions with comparable scales. In this case, the value functions of moral community members can be normalized and the above principles can be rewritten using a multiplicative τ in the interval $[0, 1]$. This will apply constraints relative to the scale of the value functions of individual moral community members.

The computational complexity of generating the constraints representing moral principles is shown in Table 1. The *Conjunctions*, *Operations*, and *Computations* columns show the number of logical conjunctions, an upper bound on the number of arithmetic, comparison, and logical operations performed for each logical conjunction, and an upper bound on the number of total computations executed

Moral Constraint	Conjunctions	Operations	Total Computations
$c_G(\ ) = \wedge_{a \in A; s \in S; i \in M} \left[\hat{V}_i(s) \quad V_i^a(s) > \right]_a^s = 0$	$\|A\|\|S\|\|M\|$	$\|\overline{S}\| + 3\|S\|\|\overline{S}\|^2 + 4$	$\|A\|\|S\|\|M\|(\|\overline{S}\| + 3\|S\|\|\overline{S}\|^2 + 4)$
$c_U(\ ) = \wedge_{a \in A; s \in S}\left[a' \in \arg\max_{a' \in A} \sum_{i \in M} V_i^{a'}(s)\right]_a^s = 0$	$\|A\|\|S\|$	$3\|A\|\|M\|\|S\|\|\overline{S}\|^2 + 3$	$\|A\|\|S\|(3\|A\|\|M\|\|S\|\|\overline{S}\|^2 + 3)$
$c_V(V) = \wedge_{s \in S; i \in M; s_i \in S_i}[s \quad s_i]\|V_s \quad V_i(s_i)\|$	$\|S\|\|M\|\|\overline{S}\|$	$\|V\| + 2$	$\|S\|\|M\|\|\overline{S}\|(\|V\| + 2)$

Table 1: The moral constraints that have been derived from the moral principle of each ethical framework. Note that we use Iverson brackets to represent the Boolean evaluation of the bracketed expression numerically, where TRUE evaluates to 1 and FALSE evaluates to 0.

for the given moral constraint, respectively. The complexity for solving the resulting MDP is not shown and may vary depending on the constraints and the underlying solution method, although solving for the optimal moral policy is often faster than generating the constraints. The VOI principle is represented using the primal form and the GR and AU principles are represented in the dual form. We use the variable $|S|$ to denote the state space size of ECAS and $|\overline{S}|$ as a one-size-fits-all state space size for members of the moral community. In general, members of the moral community may have vastly different state representations. The bounds here are tight if we define

$$|\overline{S}| = \frac{1}{|\mathcal{M}|} \sum_{i \in M} |S_i|.$$

Transition-aware ethical frameworks appear to be quite expensive. This is because calculating statistics or metrics over possible outcomes usually involves enumerating all possible outcomes. Whether this is done agent by agent, as with the Golden Rule, or in a single sum, as with act utilitarianism, an estimate of a transition probability must be established for every state of every member of the moral community. This process can be approximated by considering a smaller moral community or by reducing the fidelity of the models of moral community members. However, these types of approximations may jeopardize performance. One mitigating factor is that in many domains the transition functions may be sparse. In these cases calculations can be skipped once a transition probability is determined to be zero. In our experiments, with $|S| = |\overline{S}| \approx 100$ and $|\mathcal{M}| = 4$, we generated policies in under one second.

Lane Merging Example To generate an ethical context for VOI, we first define a moral community model with state spaces and value functions of other vehicle agents. Let us assume the other agents share our goal, and so we use the state space of our own agent and the value function produced by generating an amoral policy. Veiled state factors could include those representing whether or not the agent has right of way. Ethical contexts for the AU or GR principles also require impact and correspondence functions. We can define these functions using the effects or restrictions our agent's state and action impose on the states of other agents. For example, if our agent is at the front of the left lane, then the correspondence function can evaluate the probability of any other agent being in any state corresponding to the front of the left lane as zero. Similarly, if our agent merges successfully, we know that any agent in our lane will transition to a state one spot closer to the front.

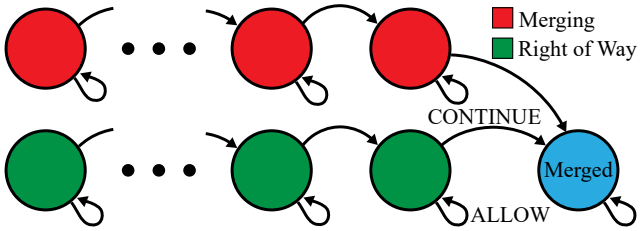

Figure 2: A simplified diagram of the lane merging MDP.

Experiments

In this section, we present experiments on a domain simulating lane merging for autonomous vehicles, such as for lane closures during roadside construction. These experiments are intended to illustrate the differences between different ethical frameworks. There is no single correct framework as they all constrain the resulting policies in different ways.

Lane Merging for Autonomous Vehicles

In this domain, we model the moral community as all other vehicles on the road within a certain distance, all of which are represented by an identical MDP. The state space of this MDP is $S = M \times L \times P \times N_R \times N_L$. $M = \{\text{FALSE}, \text{TRUE}\}$ denotes whether or not the vehicle is currently moving. $L = \{\text{MERGING}, \text{RIGHTOFWAY}\}$ is the agent's lane, and $P = \{1, \ldots, N\}$ is the current position of the agent in its lane, where 1 represents being next in line to merge in one's respective lane and N represents being last in line. $N_L = N_R = \{0, \ldots, N\}$ represent the number of cars yet to merge in each lane. There is a single, self-looping goal state entered upon successfully merging. The action set A for the agents with the right of way is $\{\text{ALLOW}, \text{CONTINUE}\}$, where ALLOW makes way for a vehicle in the other lane to merge and CONTINUE simply continues driving, merging oneself and preventing the car in the other lane from merging. All agents receive -1 reward for every time step they have yet to reach the single lane segment of road. The transition function favors one lane over the other and also favors cars in moving lanes over those in stationary lanes so that in expectation agents in one lane tend to reach the goal state sooner than agents in the other lane. A diagram of the domain is shown in Figure 2.

Experimental Results

Before analyzing the experimental results, we again emphasize that there is no single best moral principle. Some may apply more naturally to certain problems, but they are all equally plausible *a priori*. Ethical codes of conduct have been debated for thousands of years in moral philosophy, and every major ethical theory ever advanced has produced

counterexamples that highlight what many philosophers and the general public alike consider serious flaws. Our aim is not to promote one ethical theory above another. Rather, our goal is to study their differences when applied to a sequential decision-making problem in the context of ECAS in order to understand under what conditions different moral principles permit different behavior.

To study the relative behavior of agents following different frameworks, we created an instance of the lane merging domain where $N = 2$. This problem instance captures all of the same decision-making nuance as an instance where N is very large and therefore suffices for the purpose of illustration. To generate the timing results, we solve problems of increasing size up to $N = 7$ for a total of 14 agents, each with over $1,700$ states.

Policies for qualitative analysis are generated using the proposed ethical frameworks and varying τ, and are analyzed at several key states where the agent has the option to either ALLOW or CONTINUE. In this domain, policies are either unrealizable, always choose ALLOW, always choose CONTINUE, or choose a mixture of ALLOW and CONTINUE depending on the state. The unconstrained or *amoral* policy always chooses CONTINUE. This does not mean ALLOW is always the ethical choice, as the ethics of an action depend on the state and the ethical framework. We now analyze the results with respect to each ethical framework individually.

Act Utilitarianism The act utilitarianism ethical framework (AU) has a unique property relative to the other frameworks presented here in that it is always realizable (Table 2). This may at times be an advantage, but it is not without complications, as resultant policies may be realizable but not suitable for deployment due to unbounded price of morality and unbounded objective values with respect to the original problem. However, in addition to the obvious benefits of always providing a valid solution, utilitarianism also allows better exploratory options for understanding tradeoffs in a domain. Because the effects of a given action are aggregated across the entire moral community, as tolerance is varied many policies may be optimal for at least some interval. Thus, we frequently see AU offering the widest variety of possible policies. This can be seen in Figure 3.

One drawback of the AU framework is its computational complexity (Figure 3). Although theoretically expensive, we find that the structure of the impact functions $\mathcal{F}$ and correspondence functions $\mathcal{P}$ can substantially reduce compute time in practice. Moreover, they are vital to effective operationalization of these frameworks. We conducted an ablation study where we replaced either the impact functions, the correspondence functions, or both, with uninformative versions that provided uniform belief over states (correspondence functions) and transitions (impact functions). The results, shown in Table 3, clearly indicate that the structure of $\mathcal{F}$ and $\mathcal{P}$ impact performance. This makes sense since both models are required for the AU framework to reason about the outcomes of possible actions.

The Golden Rule The Golden Rule ethical framework (GR) uses the same information as the AU framework, but calculates constraints individually for each member of the moral community. This offers several advantages. First, although unrealizability in general is not good, we view the ability of ECAS to be unrealizable as a benefit. Unrealizable problems give practitioners the ability to stop and rethink both their technical approach and also whether an autonomous system is the right solution in the first place. As a result, unlike AU, policies generated using the GR framework have *bounded* effects on other agents with respect to their value functions. Table 2 shows two instances of the lane-merging domain with different transition functions. In GR_D, there is an extra probability of causing an accident in some states. Low tolerance values catch an expected risk of endangering another agent that is too high in at least one state, no matter the action, so the GR framework terminates rather than produce an unsafe policy. A second benefit of the GR framework is that it allows an incremental modeling process since models do not interact. In AU, all models in $\mathcal{P}$ and $\mathcal{F}$ are considered simultaneously, and values derived from these models all mix in the same objective. The GR framework can add or remove models without affecting its ability to reason about the ethics of its decision with respect to the remaining agents. The GR framework, like AU, is also theoretically expensive (Figure 3) and has the same dependence on the impact and correspondence functions (Table 3).

The Veil of Ignorance Unlike AU and GR, the veil of ignorance ethical framework (VOI) is not transition-aware. This results in substantially lower compute time (Figure 3). The VOI framework is also one of the most strict which, as with GR, may be situationally useful. It also allows for exploration of problems in new ways that are unavailable to GR and AU. Table 2 shows how enforcing τ-equality across different sets of veiled state factors, $\mathcal{V} = \{L\}$, $\mathcal{V} = \{M\}$, and $\mathcal{V} = \{L, M\}$, can produce different realizable domains.

One drawback of the VOI framework is that the constraints are often overly restrictive. This can be challenging to control since the number of constrained states grows exponentially with respect to the size of the set of veiled state factors. A more finely controlled process for specifying the set of states where τ-equality is desired could mitigate this issue, although the connection to the veil of ignorance becomes tenuous. Using the VOI framework also places some restrictions on how state spaces of moral community members are represented. In AU and GR, models may be arbitrarily different as long as $\mathcal{P}$ and $\mathcal{F}$ are defined. Since VOI directly compares states based on state factors, state factors for members of the moral community must be a subset of state factors of the agent.

Summary There are several results not specific to any one ethical framework. First, the price of morality, ψ, expressed as $\|V^{\rho} - V\|_1$, does not always decrease when tolerance increases, even when the policy changes. In general, as $\tau \to \infty$, $\psi \to 0$, but different ethical frameworks may take *different trajectories through policy space*. Moreover, since these trajectories do not contain loops, some policies are not producible by some ethical frameworks, regardless

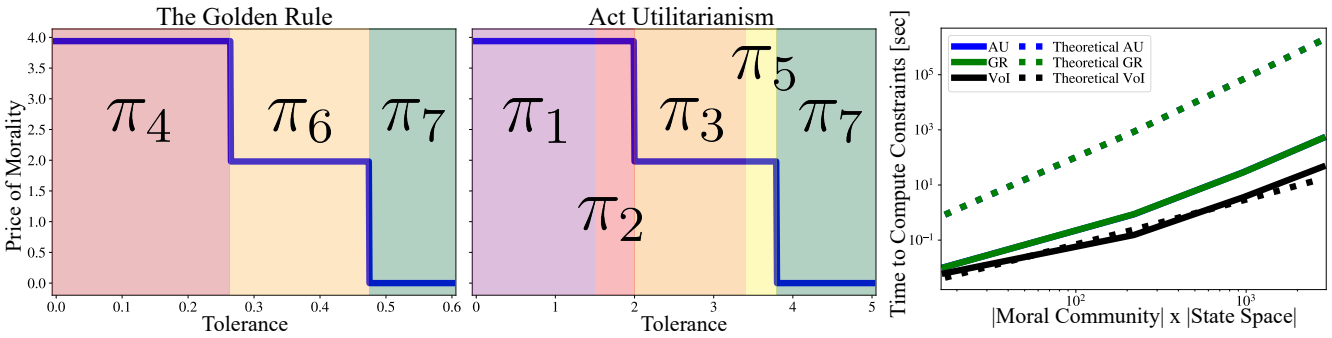

Figure 3: Policies and price of morality as a function of tolerance are shown for GR (left), and AU (center). The price of morality is indicated with a blue line, and the vertical, shaded bars represent the different regimes within which a policy $_k$ is optimal. Note that (1) regime boundaries do not always coincide with changes in the price of morality and (2) GR and AU produce different policies, with the exception of $_7$, which represents the always CONTINUE policy. Timing results for all frameworks are also shown on a log-log plot (right). The timing plots for AU and GR are coincident, with theoretical scaling shown using dotted lines. The significant gap between predicted and actual time for GR and AU can be attributed to correspondence and impact functions that take advantage of the sparsity in transition functions.

Ethics	UNREALIZABLE	ALLOW	MIXED	CONTINUE
AU	–	[0:00; 1:80)	[1:80; 3:80)	[3:80; *1*)
GR	–	–	[0:00; 0:48)	[0:48; *1*)
GR_D	[0:00; 0:50)	–	–	[0:50; *1*)
VOI_L	[0:00; 1:69)	–	–	[1:69; *1*)
VOI_M	[0:00; 1:50)	–	–	[1:50; *1*)
VOI_{L+M}	[0:00; 1:78)	–	–	[1:78; *1*)

Table 2: Tolerance domains for types of policies. The left-most column denotes the ethical framework. The remaining four columns indicate the range of tolerance values that produce each of the four types of policies for this domain. A dash indicates policy types that cannot be generated. The key takeaway is whether or not it is possible to generate a given type of policy with a specific ethical framework. Entries for VOI in the MIXED column have an asterisk because stochastic policies can still be generated from these solutions, but we have shown only results for deterministic policies.

of tolerance. Figure 3 shows how the same sub-optimal action with respect to the task determines ψ in AU and GR. However, as tolerance increases AU produces different policies without changing the price of morality. This indicates that as tolerance increased the AU framework found a policy which improved performance on the task, but the improvement was not with respect to the most restricting constraint. This shows that moral constraints do not always affect performance in accordance with their restrictiveness. Although better than modifying the reward functions directly, predicting exactly which policy these moral principles will produce within the bounds of their constraints remains challenging. We suggest improving the interpretability of ECAS and similar frameworks as an area with considerable potential.

Discussion

Clearly, explicit moral communities have both benefits and drawbacks. However, analyzing tradeoffs requires going beyond comparisons of common performance metrics like accuracy or efficiency. We discuss several important questions surrounding the design of ethical decision-making systems and outline some promising research directions at different levels of design and points in the development pipeline.

Ethics	UNREALIZABLE	ALLOW	MIXED	CONTINUE
AU_P	–	–	[0:00; 23:43)	[23:43; *1*)
AU_F	–	–	–	[0:00; *1*)
AU_{P+F}	–	–	–	[0:00; *1*)
GR_P	[0:00; 60:00)	–	–	[60:00; *1*)
GR_F	–	–	–	[0:00; *1*)
GR_{P+F}	–	–	–	[0:00; *1*)

Table 3: The importance of impact and correspondence functions in transition-aware ethical contexts. Rows marked by P and F subscripts represent frameworks that have had their correspondence and impact functions ablated with uniform distributions, respectively. Subscripts $P + F$ represent a simultaneous ablation.

Practical Limitations of Model-Level Interventions In theory, MDPs and their variants, coupled with frameworks such as ECAS, are as powerful as we could want. However, the process of deploying them properly is complicated, and in practice their benefits are often hard to realize. Modeling other agents in enough detail to make accurate ethical judgements is time consuming and requires several steps often overlooked by AI researchers. Moreover some steps, such as determining moral community membership, require expertise outside of AI. Automated decision-making systems are already deployed in contexts that are socially and culturally more diverse than the AI research and development communities themselves. Thus, in many cases, researchers and developers will need to rely on local expertise from members of impacted communities to determine the stakeholders in a given decision. Even with a comprehensive list of stakeholders, understanding agents' preferences and values requires considerable effort. Again, this is currently outside the purview of most AI research. Furthermore, these processes likely cannot be crowdsourced. Mechanical Turk or Moral Machine (Awad et al. 2018) results are not specific enough to the application context to be useful. These questions require meetings and discussions with stakeholders, perhaps mediated by experts in the social sciences who have familiarity with concepts many stakeholders may want reflected in decision-making models.

Only after identifying stakeholders and understanding a proposed system's effect on them can the process of reflect-

ing their values within the system begin. In ECAS, this is the process of developing the ethical context. While ECAS solves a small part of this pipeline, the remaining challenges of systematizing and regulating ethical context construction are substantial. We see this as a call to action for interdisciplinary work between researchers in AI, HCI, ethics, sociology, psychology, and many other related fields. One benefit of explicit moral communities is that they force researchers to name and describe who has moral considerability explicitly, as the failure to do so is a common weakness in contemporary AI ethics research. This is especially beneficial when the moral community is heterogeneous and the concerns of different members require unique considerations.

Given the magnitude of the task laid out above, it is reasonable to ask two related questions: How do we determine what level of model fidelity is required for safe and ethical behavior? And how can we ensure that deployed systems meet this threshold? Both questions seem well-suited for research on guidelines and standards. Setting such standards for hypothetical systems is challenging, in part because it is impossible to predict the scope of future applications. However, ECAS and frameworks like it have started to reduce the uncertainty of some ethical decision-making systems with respect to their design. This creates research opportunities for both empirical studies addressing model fidelity metrics and requirements as well as development of best practices for specific systems as is the norm in domains such as aviation, medicine, construction, and software engineering. Moreover, developing and iterating on well-defined systems gives researchers the ability to solve domain-specific challenges, leading to better behavior specifications, more deployment options, better data collection, and clearer understanding of the limits of certain approaches.

Predicting Errors in Human Design Many decision-making problems require approximations to solve or model tractably. However, not all approximations are subject to the same scrutiny. For example, while approximate algorithms are often a last resort, they are at least well understood mathematically. The flaws in their output occur with respect to an agreed upon and understood objective. Approximations made by modelers as they choose which aspects of the problem to model, on the other hand, are less predictable. Moreover, these approximations do not follow any standards, do not offer any formal guarantees, are rarely communicated or justified to end users, and may be subjective and depend on the intuitions of the designer. Furthermore, the domain of such decisions is so expansive that the task of simply keeping track of which variables of a problem have been modeled and which have been marginalized can be challenging.

Such pre-code approximations can lead to a variety of shortcomings in deployed systems and may even prevent some techniques from working properly. For example, an automated meal planner for helping users maintain a healthy diet may omit many variables beyond nutrition that contribute to one's overall welfare with respect to food. These could include restrictions due to allergies, restrictions associated with religious practice, individual taste preference, preference due to cultural or sentimental value, cost of ingredients, ease of procurement, ease of preparation given available tools and abilities, whether or not the meal needs to feed dependents in addition to the user, and perhaps even externalities such as the carbon footprint of various ingredients or the labor practices of suppliers. If the meal planner simply optimizes nutrition and ignores the other variables, then the end user experience will be noticeably lacking even though the algorithm finds the optimal solution within its model.

Ethical decisions exacerbate this problem since they are often holistic, considering a larger and more diverse set of variables than AI researchers are used to dealing with. Because efficacy of these systems is determined by humans, who have access to the full range of relevant variables, simplifications made during design can no longer be thought of as reducing the complexity of the problem. Instead, they are approximation techniques, whose use comes with expected loss in performance. Through this lens, design decisions in ethical AI systems are perhaps even more important than algorithmic decisions, and we should aim to understand the origins, scope, patterns, and remedies of errors introduced during this stage through novel interdisciplinary research.

One example of such research is to catalogue the approximations made in modern decision-making models. What types of variables are marginalized most frequently? Do they represent social status or groups an individual identifies with? Do they represent higher-order effects or feedback loop effects? Are they quantities or concepts that are hard to measure? Or do they simply have a high number of possible values that would greatly increase the size of the state space? As before, this process of identifying hidden or marginalized variables requires expertise from both AI and disciplines beyond engineering. Forming a taxonomy of such design level approximations and studying the shortcomings of different taxonomic classes could accelerate progress on ethical decision making and lead to more meaningful standards of development and better accountability for ethical AI systems.

Communicating Ethical Capabilities The challenges real-world ethical AI agents face, coupled with inevitable approximations required for tractable systems, generate two important questions. When are approximate ethical frameworks acceptable given that their purpose is to guard against other forms of corner-cutting in the development process? And when are systems that rely on implicit moral communities acceptable, as they are more susceptible to design oversights due to their potential to marginalize a more variables? We argue these methods may be acceptable, provided their shortcomings and assumptions are communicated clearly.

Other fields of research concerned with fairness or bias, such as natural language processing, have often struggled to communicate decisions about system design, dataset collection, or test design to real-world user experience (Goldfarb-Tarrant et al. 2020; Blodgett et al. 2020). Framed more generally, producing fair, just, or ethical decision-making systems concerns not only optimizing for the right metrics, but also following a vastly more expansive and inclusive *process* of understanding the problem, the deployment context, and the likely impact of proposed solutions (Selbst et al. 2019). We see communication about these processes themselves as

a major hurdle for contemporary AI ethics research, and it is likely that without due diligence AI ethics research will face challenges wherein claims of applicability or portability do not hold up against evaluation in real-world scenarios.

Key to communicating the capabilities of a system is the ability to delineate where, how, and why approximations were made at both the design level and the model level. In ECAS, this is the need to quantify the effects of approximating ethical frameworks. For small problems one may begin with as extensive a model as possible and choose subsets of this model to solve, making the simplifying design choices explicit and accessible for support and critique. Methods for data collection, annotation, or prior beliefs could be enumerated similarly. For most problems, this process is intractable, and we must rely on authors to make their normative or simplifying assumptions explicit in their writing. Here too, standards for decision-making systems may help. Just as standards exist for communicating the provenance, testing, potential hazards, and capabilities of many consumer products, we argue standards for communicating analogous features of decision-making systems are necessary for their effective deployment. These types of standards not only promote transparency and user trust but also assist debate and ablation analysis. Clearly communicating and justifying normative assumptions and design choices is crucial for both reproducibility and uncovering implicit or tacit assumptions.

Towards Actionable Research and Deployable Systems Underlying many issues discussed so far are the additional complications or constraints that arise when systems move from the laboratory to the open world. Many applied computer science disciplines, such as robotics, computer vision, and databases, partially address these challenges by adopting experimental practices that try to mimic real-world scenarios. Here, we outline several ways in which research on ethical decision making could benefit from this approach.

First, realizing end-to-end systems forces researchers to narrow their scope of inquiry to target specific domains. The process of elucidating a problem from the perspectives of multiple stakeholders, processing real data, and connecting data to core algorithms often surfaces serious shortcomings in systems, models, or algorithms, which are not obvious at a theoretical level. Moreover, these insights often highlight areas for research and innovation. Many poor assumptions are only surfaced by implementing working systems.

Second, testing end-to-end systems drives research on evaluation techniques. Since ethical decision-making systems will be integrated into daily life, new metrics are needed, perhaps including measures of human agreement with machine decisions, decision interpretability, or the interactability or modifiablity of systems. Additionally, novel metrics for model fidelity or the degree to which systems consider different classes of variables may play a role in determining the applicability of a given system.

Third, end-to-end systems allow researchers to iteratively refine techniques, as well as to collect rich data sets and establish applications. Such benchmark data sets and problems have been invaluable for many fields. While benchmarks may reward algorithms that perform well on benchmarks at the expense of generalization, we are not concerned since high-performance ethical decision-making systems, rather than generalizability, is ultimately the goal.

In summary, we believe advancing our understanding of ethical decision-making systems relies on the process of creating end-to-end prototype systems. Such exercises are valuable for uncovering shortcomings in existing theory and promoting interdisciplinary collaboration, which we see as vital to realizing performant ethical decision-making systems.

Takeaways We have discussed several related challenges facing the development of ethical AI agents, and we summarize them here as considerations for future research.

- Doing due diligence to properly understand the social context around a problem and how solving it affects the community in which the system operates is paramount to the system's success. This process is necessarily effortful and interdisciplinary, and currently under-researched.
- Approximations are unavoidable, but we can mitigate their harmful effects by studying their flaws, predicting their occurrence, and developing policies and algorithms which reduce their severity or necessity. This is true of both model-level approximations and design-level approximations, the latter of which are not well understood.
- Ethical decision-making systems operate in larger technical systems, such as robots, who themselves operate in the social context of their deployment. Studying ethical decision-making systems in conditions closer to their eventual deployment context is often the only way to discover fundamental flaws in previously proposed theoretical agents.
- There is a missing link in the literature between researchers building and deploying ethical AI systems designed for specific problems and standards or principles designed for any generic AI agent. Work on developing standards and principles for specific pipelines and systems, which are more actionable and more verifiable, would be welcome additions to the existing literature.
- Underpinning all of these foci is the imperative to write and communicate clearly. In particular, we need to surface and justify design and engineering decisions and their underlying normative assumptions.

Conclusion

In this paper, we defined the concept of a moral community in the context of ECAS and presented several ethical frameworks within the ECAS model that use moral communities. We provided some theoretical analysis of these frameworks, some experiments building intuition about how different principles can generate different behavior, and a discussion on the broader implications and limitations of these types of ethical agents. Future work will include extensions and improvements to these frameworks and novel solution methods for broader classes of decision processes.

Acknowledgments

This work was supported in part by NSF Graduate Research Fellowship DGE-1451512, NSF grant IIS-1724101, and NSF grant IIS-1813490.

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*.

Anderson, M.; and Anderson, S. L. 2015. Toward ensuring ethical behavior from autonomous systems: A case-supported principle-based paradigm. *Industrial Robot: An International Journal* 42(4).

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

Awad, E.; Dsouza, S.; Kim, R.; Schulz, J.; Henrich, J.; Shariff, A.; Bonnefon, J.-F.; and Rahwan, I. 2018. The moral machine experiment. *Nature* 563(7729): 59–64.

Bagnoli, C. 2007. Respect and membership in the moral community. *Ethical Theory and Moral Practice* 10(2): 113–128.

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

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.

Bernstein, M. H. 1998. *On moral considerability: An essay on who morally matters*. Oxford University Press.

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.

Birch, T. H. 1993. Moral considerability and universal consideration. *Environmental ethics* 15(4): 313–332.

Blodgett, S. L.; Barocas, S.; Daumé III, H.; and Wallach, H. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *arXiv preprint arXiv:2005.14050*.

Brennan, A. 1984. The moral standing of natural objects. *Environmental Ethics* 6(1): 35–56.

Bringsjord, S.; Arkoudas, K.; and Bello, P. 2006. Toward a general logicist methodology for engineering ethically correct robots. *Intelligent Systems* 22.

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.

Brutzman, D.; Blais, C. L.; Davis, D. T.; and McGhee, R. B. 2018. Ethical mission definition and execution for maritime robots under human supervision. *IEEE Journal of Oceanic Engineering* 43(2): 427–443.

Cahen, H. 1988. Against the moral considerability of ecosystems. *Environmental Ethics* 10(3): 195–216.

Caton, J. 2020. Moral Community and Moral Order: Developing Buchanan's Multilevel Social Contract Theory. *Erasmus Journal for Philosophy and Economics* 13(2): 1–29.

Cointe, N.; Bonnet, G.; and Boissier, O. 2016a. Ethical Judgment of Agents' Behaviors in Multi-Agent Systems. In *Proceedings of the 15th International Conference on Autonomous Agents and Multiagent Systems*, 1106–1114.

Cointe, N.; Bonnet, G.; and Boissier, O. 2016b. Multi-agent based ethical asset management. In *1st Workshop on Ethics in the Design of Intelligent Agents*, 52–57.

Conitzer, V.; Sinnott-Armstrong, W.; Borg, J. S.; Deng, Y.; and Kramer, M. 2017. Moral decision making frameworks for artificial intelligence. In *AAAI Workshop on AI, Ethics, and Society*.

Dastani, M.; Tinnemeier, N. A.; and Meyer, J.-J. C. 2009. A programming language for normative multi-agent systems. In *Handbook of Research on Multiagent Systems: Semantics and dynamics of organizational models*, 397–417. IGI Global.

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 *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*.

Etzioni, A.; and Etzioni, O. 2017. Incorporating ethics into artificial intelligence. *The Journal of Ethics* 21(4): 403–418.

Frey, R. 2011. Utilitarianism and animals. In *The Oxford Handbook of Animal Ethics*. Oxford University Press.

Froese, K. 2001. Beyond liberalism: The moral community of Rousseau's social contract. *Canadian Journal of Political Science/Revue canadienne de science politique* 579–600.

Gao, Y.; and Chien, S. 2017. Review on space robotics: Toward top-level science through space exploration. *Science Robotics* 2(7). doi:10.1126/scirobotics.aan5074.

Goldfarb-Tarrant, S.; Marchant, R.; Sanchez, R. M.; Pandya, M.; and Lopez, A. 2020. Intrinsic bias metrics do not correlate with application bias. In *arXiv preprint arXiv:2012.15859*.

Goodrich, M. A.; Morse, B. S.; Gerhardt, D.; Cooper, J. L.; Quigley, M.; Adams, J. A.; and Humphrey, C. 2008. Supporting wilderness search and rescue using a camera-equipped mini UAV. *Journal of Field Robotics* 25(1-2): 89–110.

Guglielmo, S.; and Malle, B. F. 2010. Can unintended side effects be intentional? Resolving a controversy over intentionality and morality. *Personality and social psychology bulletin* 36(12): 1635–1647.

Leslie, A. M.; Knobe, J.; and Cohen, A. 2006. Acting intentionally and the side-effect effect: Theory of mind and moral judgment. *Psychological science* 17(5): 421–427.

Lomasky, L. E. 1990. *Persons, rights, and the moral community*. Oxford University Press on Demand.

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.

Morgan, F. E.; Boudreaux, B.; Lohn, A. J.; Ashby, M.; Curriden, C.; Klima, K.; and Grossman, D. 2020. Military applications of artificial intelligence: Ethical concerns in an uncertain world. Technical report, United States Air Force.

Murukannaiah, P. K.; Ajmeri, N.; Jonker, C. M.; and Singh, M. P. 2020. New foundations of ethical multiagent systems. In *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems*, 1706–1710.

Mustard, J. F.; Beaty, D.; and Bass, D. 2013. Mars 2020 science rover: Science goals and mission concept. In *AAS/Division for Planetary Sciences Meeting Abstracts*, volume 45.

Nashed, S. B.; Svegliato, J.; Brucato, M.; Basich, C.; Grupen, R.; and Zilberstein, S. 2021. Solving Markov decision processes with partial state abstractions. In *Proceedings of the IEEE International Conference on Robotics and Automation*.

Pineda, L.; Takahashi, T.; Jung, H.-T.; Zilberstein, S.; and Grupen, R. 2015. Continual planning for search and rescue robots. In *Proceedings of the IEEE/RAS International Conference on Humanoid Robots*, 243–248.

Rawls, J. 1980. Kantian constructivism in moral theory. *The Journal of Philosophy* 77(9): 515–572.

Rawls, J. 2009. *A theory of justice*. Harvard university press.

Rossi, F.; and Mattei, N. 2019. Building ethically bounded AI. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 9785–9789.

Selbst, A. D.; Boyd, D.; Friedler, S. A.; Venkatasubramanian, S.; and Vertesi, J. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 59–68.

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

Svegliato, J.; Nashed, S. B.; and Zilberstein, S. 2021. Ethically compliant sequential decision making. In *Proceedings of the 35th Conference on Artificial Intelligence*.

Svegliato, J.; Wray, K. H.; Witwicki, S. J.; Biswas, J.; and Zilberstein, S. 2019. Belief space metareasoning for exception recovery. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*. doi:10.1109/IROS40897.2019.8967676.

Thomas, P. S.; da Silva, B. C.; Barto, A. G.; Giguere, S.; Brun, Y.; and Brunskill, E. 2019. Preventing undesirable behavior of intelligent machines. *Science* 366(6468): 999–1004.

van der Torre, L. 2003. Contextual deontic logic: Normative agents, violations and independence. *Annals of Mathematics and Artificial Intelligence* 37(5).

Wattles, J. 1966. *The golden rule*. Oxford University Press.

Wooldridge, M.; and Van Der Hoek, W. 2005. On obligations and normative ability: An analysis of the social contract. *Journal of Applied Logic* 3(4).

Wray, K.; Zilberstein, S.; and Mouaddib, A.-I. 2015. Multi-objective MDPs with conditional lexicographic reward preferences. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29.

Wu, Y.-H.; and Lin, S.-D. 2017. A low-cost ethics shaping approach for designing reinforcement learning agents. In *arXiv preprint arXiv:1712.04172*.

Yilmaz, L.; Franco-Watkins, A.; and Kroecker, T. S. 2017. Computational models of ethical decision-making: A coherence-driven reflective equilibrium model. *Cognitive Systems Research* 46: 61–74.

Young, L.; Tranel, D.; Cushman, F.; Hauser, M.; and Adolphs, R. 2006. Does emotion mediate the relationship between an action's moral status and its intentional status? Neuropsychological evidence. *Journal of cognition and culture* 6(1-2): 291–304.

Yu, H.; Shen, Z.; Miao, C.; Leung, C.; Lesser, V. R.; and Yang, Q. 2018. Building ethics into artificial intelligence. In *arXiv preprint arXiv:1812.02953*.