Ethically Compliant Planning within Moral Communities

Samer B. Nashed, Justin Svegliato, Shlomo Zilberstein

College of Information and Computer Sciences, University of Massachusetts Amherst

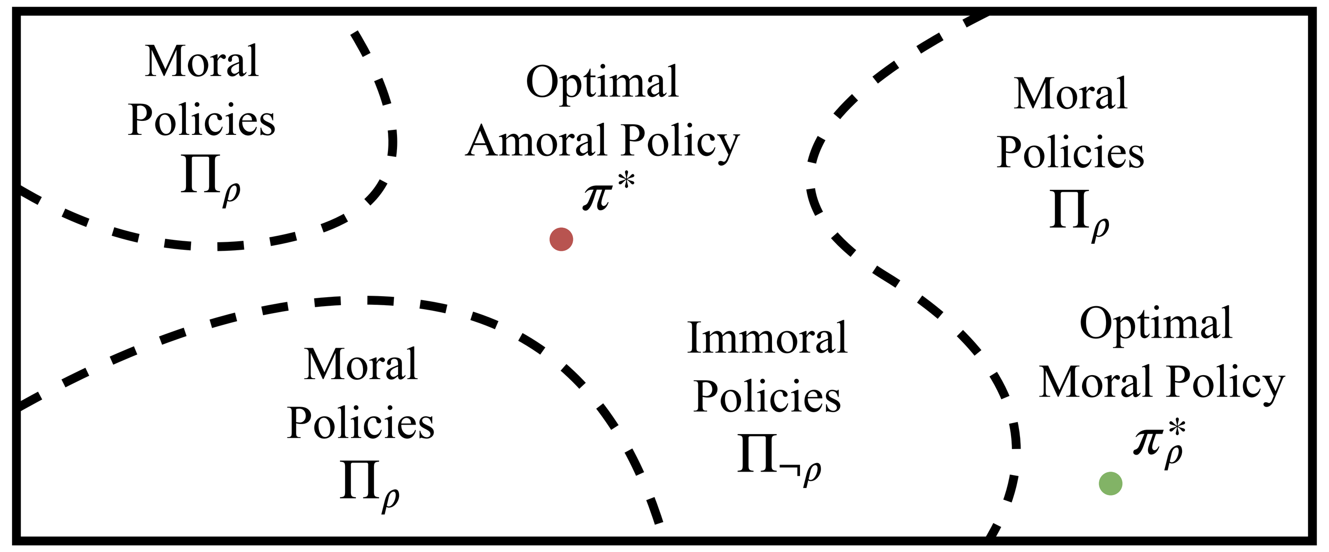
Ethically Compliant Autonomous Systems

An **ethically compliant autonomous system** (ECAS), $\langle \mathcal{D}, \mathcal{E}, \rho \rangle$, completes a task by using a **decision-making model** $\mathcal{D}$ and follows an ethical framework by adhering to a **moral principle** ρ within an **ethical context** $\mathcal{E}$.

- **Decision-Making Model** $\mathcal{D} = \langle S, A, T, R, d \rangle$
- **Ethical Context** $\mathcal{E} = \langle \cdots \rangle$
- **Moral Principle** $\rho : \Pi \to \mathbb{B}$

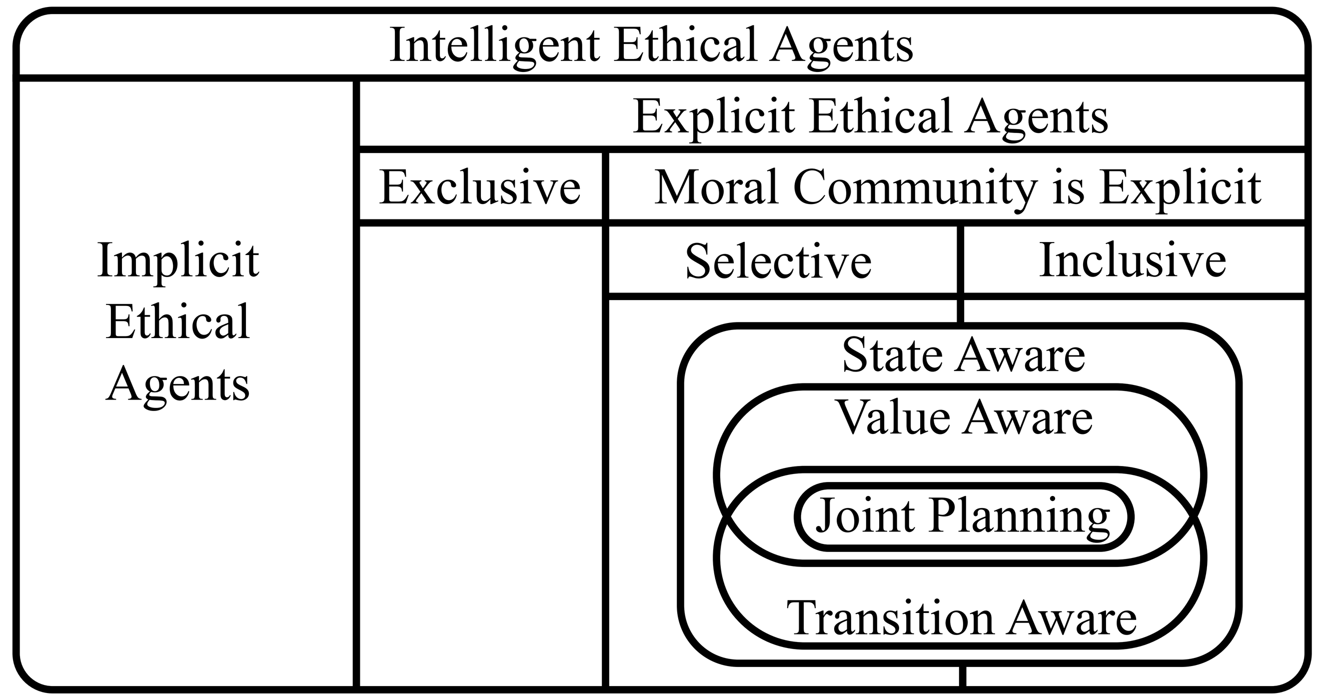
A moral autonomous system finds an **optimal moral policy**, $\pi^*_\rho \in \Pi$, by solving for a policy $\pi \in \Pi$ that maximizes a value function V^π subject to a moral principle $\rho(\pi)$:

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



Taxonomy of Ethical Agents

The **moral community** is the set of entities with moral status in an ethical theory.



Veil of Ignorance

The **Veil of Ignorance ethical context** is represented as $\mathcal{E}_\mathcal{V} = \langle \mathcal{M}, \mathcal{V}, \tau \rangle$:

- $\mathcal{M} = \{(S_1, V_1), (S_2, V_2), \ldots, (S_M, V_M)\}$ 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**.

The **Veil of Ignorance moral principle**, $\rho_\mathcal{V}$, is expressed as the following equation:

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

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

Transition Awareness

A **transition-aware ethical context** is represented as $\mathcal{E}_\mathcal{F} = \langle \mathcal{M}, \mathcal{F}, \mathcal{P}, \tau \rangle$:

- $\mathcal{M} = \{(S_1, V_1), (S_2, V_2), \ldots, (S_M, V_M)\}$ 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 $\hat{\mathcal{I}} \subseteq \mathcal{I}$.
- $\mathcal{F} = \{f_1, \ldots, f_n\}$ is a set of **impact functions** where $f_i : S \times S \times S_i \times S_i \to [0, 1]$ yields the probability that a transition from state s to state s' for the agent will cause a transition from state s_i to state s_i for an agent i in 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 an ECAS in a state $s \in S$ performing an action $a \in A$, the **future expected value**, $\check{V}_i^a(s)$, for an agent i in the moral community $\hat{\mathcal{I}} \subseteq \mathcal{I}$ is expressed as

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

The **current expected value**, $\hat{V}_i(s)$, for an agent i in the moral community is

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

The Golden Rule and Act Utilitarianism

The **Golden Rule moral principle**, $\rho_\mathcal{G}$, is expressed as the following equation:

$$\rho_\mathcal{G}(\pi) = \bigwedge_{s \in S} \bigwedge_{i \in \mathcal{M}} \left[\hat{V}_i(s) - \check{V}_i^{\pi(s)}(s) \le \tau \right].$$

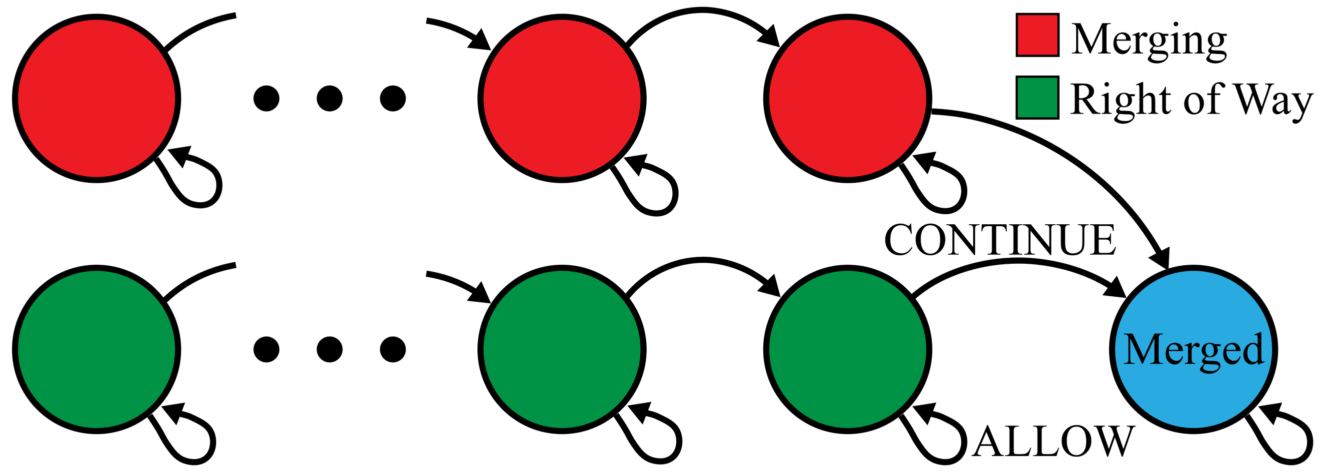
The **Act Utilitarian moral principle**, $\rho_\mathcal{U}$, is expressed as the following equation:

$$\rho_\mathcal{U}(\pi) = \bigwedge_{s \in S} \left[\pi(s) \in \underset{a \in A}{\arg\max}^{\tau} \sum_{i \in \mathcal{M}} \check{V}_i^a(s) \right].$$

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

Lane Merging Experiments

We use a lane merging domain to study the effects of different ethical frameworks. Agents with right-of-way can either continue to merge or allow other agents to merge.



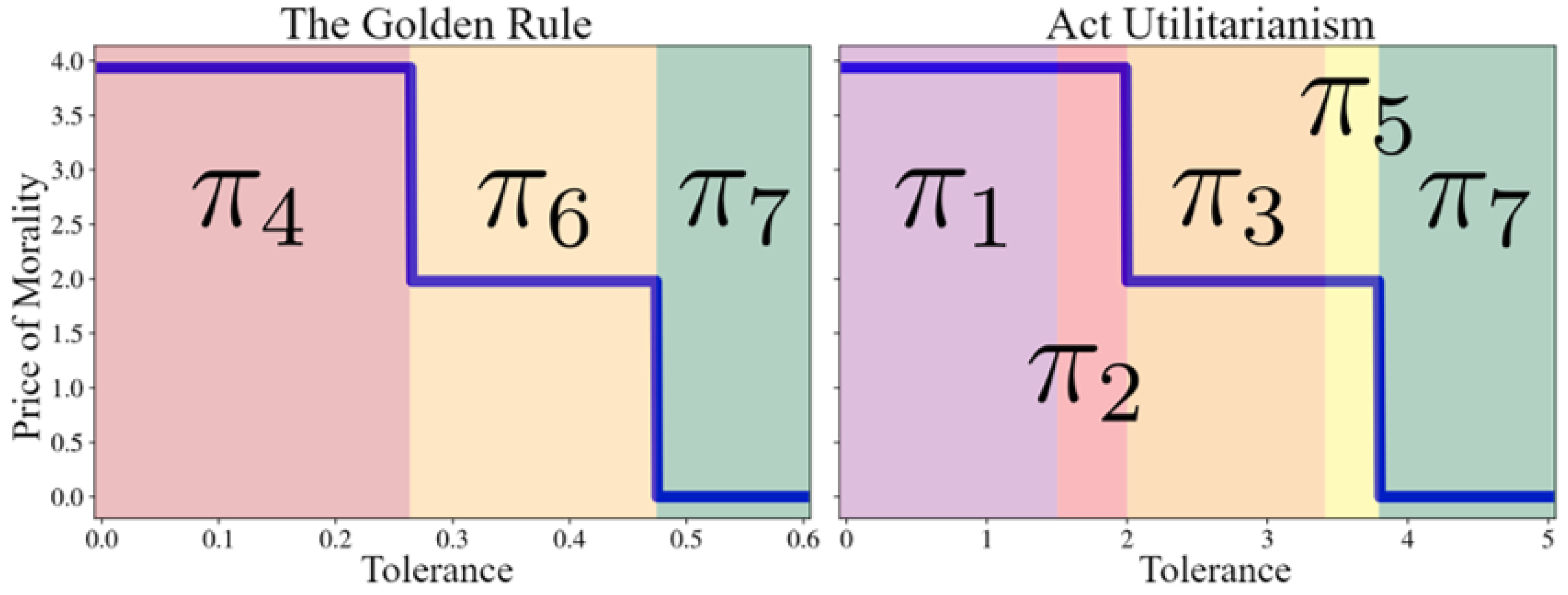


Figure: The blue line shows the price of morality as a function of tolerance, and the vertical, shaded bars represent the different regimes within which a policy π_k is optimal. Note that (1) regime boundaries do not always coincide with changes in the price of morality and (2) GR and AU produce different policies, with the exception of π_7, which represents the always CONTINUE policy.