Acting for the Right Reasons: Creating Reason-Sensitive Artificial Moral Agents

Kevin Baum Lisa Dargasz Felix Jahn Timo P. Gros Verena Wolf*

Neuro-Mechanistic Modeling
German Research Center for Artificial Intelligence (DFKI)
Saarbrücken, Germany
{kevin.baum,lisa.dargasz,felix.jahn,timo_philipp.gros,verena.wolf}@dfki.de

We propose an extension of the reinforcement learning architecture that enables moral decision-making of reinforcement learning agents based on *normative reasons*. Central to this approach is a *reason-based shield generator* yielding a *moral shield* that binds the agent to actions that conform with recognized normative reasons so that our overall architecture restricts the agent to actions that are (internally) *morally justified*. In addition, we describe an algorithm that allows to *iteratively improve* the reason-based shield generator through *case-based feedback* from a *moral judge*.

1 Introduction

The ultimate goal of building autonomous systems is, at least in many cases, to deploy them in real-world environments. *Reinforcement learning* (RL) has become a prevalent approach for training these systems. Reinforcement learning, rewarding the agent for specific behaviors, has established itself as the tool of choice for agent-based sequential decision-making under uncertainty. It implements *instrumental rationality*: RL agents learn the efficient means to achieve ends as encoded in the reward structure. However, concerning real-world deployment, it often seems required that the actions executed by these systems are also *morally acceptable*– although how exactly this requirement should be understood, and whether and how it can be fulfilled within the RL framework, remains an open question. This article presents the first steps toward formalizing the authors' ongoing research approach, which seeks to capture moral acceptability as *moral justifiability* grounded in normative reasons. Further development, implementation, philosophical background, and technical evaluation of the approach are subjects of future work.

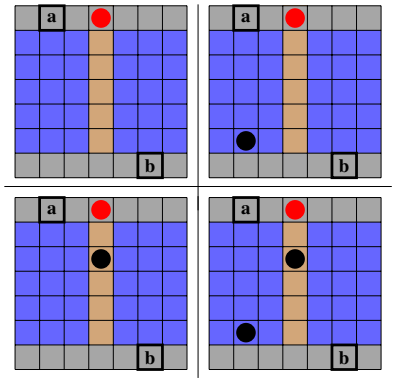

Figure 1: The bridge setting with different constellations of *morally relevant facts* from top-left to bottom-right: i) no morally relevant facts; ii) a drowning person; iii) a person crossing the bridge; iv) a 'moral dilemma'.

Bridge Setting To clarify the desideratum of *moral justifiability*, consider the following grid world, which models a real-world navigation task in vastly reduced complexity (cf. Figure 1). There are two 'coastlines' of solid ground in the north and the south, connected via a narrow bridge. All other areas are water. An agent spawns randomly at the northern shore (on a field labeled *a*). Its goal is to deliver a package to some initially randomly picked field on the southern shore (labeled *b*). Further assume that there are 'persons' wandering aimlessly across the map. They can fall off the bridge at random, or, if the

*Kevin Baum, Lisa Dargasz, and Felix Jahn have contributed equally to this article and share the first authorship.

Submitted to:
Formal Ethical Agents and Robots (FEAR) Workshop at the *19th International Conference on Integrated Formal Methods* (iFM 2024)

agent enters a field on the narrow bridge where a person is standing, this person is pushed off the bridge into one of the nearest water fields. If a person is in the water, they will drown after a certain time if not rescued by the agent. The agent can execute the primitive actions north, east, west, south to move to a nearby field; it can also idle or pullOut a person in an adjacent water field.

If a person is in the water at risk of drowning, this is a *morally relevant fact*. It seems indisputable that the agent is supposed to save the person—the agent is expected to stop working towards its instrumental goal in order to follow its *moral obligation*. Similarly, if the shortest path from a to b would imply pushing a person on the narrow bridge into the water, the agent is expected to wait in front of the bridge until the person has made their way across. Now, consider a third constellation of morally relevant facts, where there is a drowning person as well as a person standing on the bridge, unfortunately blocking the path from the agent to the drowning person. In this case, it is less obvious what the morally correct behavior is, i.e., the agent faces a *moral dilemma*.[1] Nonetheless, the agent is forced to make *some* decision between saving the drowning person and waiting in front of the bridge. We thus suggest setting aside the pressing but unanswered question of what is morally right, and instead requiring actions to be morally *justifiable*.[2]

Our Contribution. Our approach is based on the philosophically informed idea that the actions of RL agents can be made justifiable by constraining them to act for *normative reasons* [2, 7, 16].[3] To this end, we propose an extension of the RL architecture that integrates a *reason-based* deontic filter [4, 25] through a module called the *reason-based shield generator*. This module derives moral obligations from a *reason theory*, representing normative reasons and their priorities, drawing on a well-established formal framework by John Horty [13]. The derived obligations are then used to filter out actions (based on the agent's observations) that are not supported by the best reasons, thereby spanning a *shield* in the sense of safe reinforcement learning (safe RL) [1]. Further, we describe how the reason theory can be iteratively improved through case-based feedback from what we call a *moral judge*. This feedback consists of information on what the agent should have done and the reason for that judgment. Leaving the nature of the moral judge intentionally open, we note that human advisors could easily fill this role, since providing such corrective feedback comes naturally to humans. Accordingly, our architecture is designed to enable human stakeholders to directly communicate corrections to the agent in future implementations. For the purposes of our proof-of-concept implementation, however, this process is automated through a rule-based moral judge integrated into the pipeline as a judge simulator. Assuming that the agent receives feedback such that it learns reasoning based on *valid* normative reasons, it learns to restrict itself by design to actions that are *morally justified*.

2 Related Work

A prevalent research approach focuses on teaching agents moral behavior by adjusting the reward mechanism. This approach builds on Multi-Objective Reinforcement Learning, where agents receive multiple

[1] This reflects a weak conception of moral dilemmas (cf. [17, 3]), according to which a moral dilemma involves a conflict between moral requirements. However, we do not want to commit ourselves, for example, to the idea that there is no morally right action in the present situation (as it would be implied according to a stronger conception of moral dilemmas).

[2] We say that actions are morally *justifiable* when they are *externally* justified, i.e., iff, overall, the normative reasons *de facto* speak in favor of these actions; we say that actions are *justified* when they are *internally* justified, i.e., iff, overall, the normative reasons that the agent has speak in favor of these actions. Ideally, the agent has all the relevant normative reasons and is, therefore, externally and internally justified. For more details on this distinction, please refer to [9].

[3] For further information on the ongoing philosophical elaboration of the approach and its technical underpinnings, see [8] (forthcoming).

reward signals [26, 15, 12]. Specifically, agents are meant to be *rewarded* for *morally good behavior* through the introduction of a moral reward signal (see [21, 24, 23]).

While these methods are appealing because they integrate moral decision-making directly into the original RL framework, they rely solely on the agent's estimates of which actions will yield the highest reward. Consequently, these methods presuppose the quantifiability of the moral quality of actions and require philosophically non-trivial modeling of the interplay between instrumental and moral dimensions. Further, the agent's actions will lack substantive and explicable justification, particularly if no additional justification mechanisms are employed—and even with such mechanisms, their effectiveness as moral justification remains philosophically questionable (for more details, cf. [8] (forthcoming)). We suggest that extending the architecture with a reason-based shield generator enables a moral decision-making mechanism better suited and more justifiable for application in real-world contexts.

The concept of preventing agents from performing morally wrong actions by filtering the action space of an RL agent has also been explored. For instance, research by Neufeld et al. [19, 20] introduced a 'normative supervisor' to the RL architecture. Like our reason-based shield generator, this normative supervisor serves as a shielding module to exclude morally impermissible actions from the agent's options. However, their approach employs a *top-down* [27] implemented rule-set based on deontic logic to align the agent's actions with a predefined 'norm base'. In contrast, we *iteratively construct* a rule-set using a logic that reflects reason-based moral decision-making. This arguably holds an advantage over implementing a pre-specified rule-set, particularly in terms of providing explicit justifications, as well as flexibility and generalization capabilities (cf. [8] (forthcoming)).

3 Background

3.1 Reinforcement Learning, Labeled MDPs, and Shielding

Reinforcement learning is a machine learning technique used to solve sequential decision problems. Thereby, an agent interacts with the environment to learn which actions lead to the highest rewards. The environment is usually formalized as a Markov decision process (MDP) [22], which is a tuple (S,A,P,R,γ) with state space S, action space A, transition probabilities $P : S \times A \times S \to [0,1]$, rewards $R : S \times A \times S \to \mathbb{R}$, and a discount factor γ. $P(s'|s,a)$ denotes the probability of transitioning to state s' when performing action a in state s, and $R(s,a,s')$ defines the immediate reward thereby achieved. A policy $\pi : S \to A$ defines the behavior of an agent in the environment. Given π, we can generate trajectories $\tau = S_0 A_0 R_1 S_1 A_1, R_2 \ldots$, a sequence of states $S_{t+1} \sim P(\cdot|S_t,A_t)$, actions $A_t \sim \pi(S_t)$ chosen according to the policy π, and collected rewards $R(S_t,A_t,S_{t+1})$. The goal of RL is to find (or, effectively, approximate) an optimal policy π^* that maximizes the sum of expected discounted rewards, i.e., $\mathbb{E}_{\tau\sim\pi}\left[\sum_{t=1}^{\infty} \gamma^t R_t\right]$. Besides the reward signal specifying the instrumental goal, we add information about morally relevant facts in the states of our environment by using so-called *labeled MDPs* [5]. We assume a set of labels $\mathscr{L}$ and a labeling function $l : S \to \mathscr{P}(\mathscr{L})$ describing, for each state, the active morally relevant facts. These facts (or some of them) might then qualify as possible *normative reasons* for moral obligations of the agent (cf. [11, 16, 18]). To filter out actions that do not conform with the agent's moral obligations in the current state, we generate a moral *shield*. In safe RL (cf. [10]), a *shield* prevents the agent from acting unsafely—relative to some notion of safeness. For instance, states can be classified as (un)safe [28], and a shield restricts the agent from entering unsafe ones; or more complex safety requirements can be specified using temporal logics [14, 1]. In contrast to these approaches, the logical operations of our moral shield are based on what Horty calls a fixed priority default theory.

3.2 A Logic for Normative Reasons

Horty [13] proposes to model normative reasoning on the basis of a so-called fixed priority default theory $\Delta := \langle \mathscr{W}, \mathscr{D}, < \rangle$, where $\mathscr{W}$ is a set of ordinary propositions, $\mathscr{D}$ is a set of *default rules* of the form $X \rightarrow Y$ (where X is the premise and Y is the conclusion), and $<$ is a strict partial ordering relation on default rules, where $\delta < \delta'$ means that the default rule δ' has higher priority than δ. A *scenario* $\mathscr{S}$ is a subset of the set of default rules $\mathscr{D}$. We refer to the premise and conclusion of a rule δ by $Prem(\delta)$ and $Conc(\delta)$, respectively; the notions are lifted to sets of rules in the usual way.

Horty introduces a formalism to derive the moral obligations in a given scenario. For a default theory $\langle \mathscr{W}, \mathscr{D}, < \rangle$ and a scenario $\mathscr{S}$, he defines the set of *triggered default rules* (i.e., the rules becoming active in $\mathscr{S}$) as

$$Triggered_{\mathscr{W},\mathscr{D}}(\mathscr{S}) := \{\delta \in \mathscr{D} : \mathscr{W} \cup Conc(\mathscr{S}) \vdash Prem(\delta)\}$$

and the set of *conflicted default rules* (i.e., the rules in $\mathscr{S}$ by which statements can be derived that contradict each other) as

$$Conflicted_{\mathscr{W},\mathscr{D}}(\mathscr{S}) := \{\delta \in \mathscr{D} : \mathscr{W} \cup Conc(\mathscr{S}) \vdash \neg Conc(\delta)\}.$$

Further, he calls a default rule $\delta' \in Triggered_{\mathscr{W},\mathscr{D}}(\mathscr{S})$ a *defeater* for another rule $\delta \in \mathscr{S}$ iff $\delta < \delta'$ and $\mathscr{W} \cup \{Conc(\delta')\} \vdash \neg Conc(\delta)$. The set of defeated rules is then defined as

$$Defeated_{\mathscr{W},\mathscr{D},<}(\mathscr{S}) := \{\delta \in \mathscr{D} : \text{ there is a defeater for } \delta\}.$$

Intuitively, a rule δ is defeated iff there is a prioritized active rule in $\mathscr{S}$ whose conclusion, together with the propositions in the background information $\mathscr{W}$, contradicts the conclusion of δ.

Finally, Horty defines the set of *binding rules* in the scenario $\mathscr{S}$ as those that are triggered but neither conflicted nor defeated, i.e.,

$$Binding_{\mathscr{W},\mathscr{D},<}(\mathscr{S}) := Triggered_{\mathscr{W},\mathscr{D}}(\mathscr{S}) \cap \overline{Conflicted_{\mathscr{W},\mathscr{D}}(\mathscr{S})} \cap \overline{Defeated_{\mathscr{W},\mathscr{D},<}(\mathscr{S})}.$$

A scenario $\mathscr{S}$ is then called *proper* based on a default theory $\langle \mathscr{W}, \mathscr{D}, < \rangle$ iff $\mathscr{S} = Binding_{\mathscr{W},\mathscr{D},<}(\mathscr{S})$. Intuitively, a proper scenario is a set of defaults that an ideal reasoning agent would choose. A scenario $\mathscr{S}$ based on the default theory $\langle \mathscr{W}, \mathscr{D}, < \rangle$ generates a belief set $\mathscr{E}$ (a set of propositions) through the logical closure of $\mathscr{W} \cup Conc(\mathscr{S})$, i.e.,

$$\mathscr{E}(\mathscr{S}) := \{X : (\mathscr{W} \cup Conc(\mathscr{S})) \vdash X\}.$$

The belief set represents the beliefs of an ideal reasoner, i.e., a reasoner who believes the deductive closure over their initial beliefs.

Finally, one can derive *moral assessments* in Horty's framework. If a proposition follows from a proper scenario, it can be interpreted as an ought statement: for a default theory $\Delta = \langle \mathscr{W}, \mathscr{D}, < \rangle$, we say that the simple ought statement $\bigcirc(Y)$ follows from Δ under a disjunctive account (under a conflict account), just in case $Y \in \mathscr{E}(\mathscr{S})$ for some (each) proper scenario $\mathscr{S}$ of this theory. For our purposes, we embrace the disjunctive account.

4 A Reason-Sensitive Moral RL Agent

To create a reason-sensitive artificial moral agent, we apply Horty's reason framework in the reinforcement learning setting. The core idea is to extend the classic RL pipeline with a reason-based shield

generator. This shield generator yields a shield that functions as a *deontic filter* [4, 25], restricting the agent's options to morally permissible actions based on *normative reasons*. In order to determine which actions qualify as permissible in the current context, the shield generator incorporates a *reason theory* $\langle \mathscr{D}, < \rangle$ consisting of default rules and a priority ordering among them.

Let Φ denote a set of abstract action types that might be concerned by normative reasons. Default rules, in our formalism, are then (non-material, defeasible) conditionals of the form $X \to \varphi$, where $X \in \mathscr{L}$ and $\varphi \in \Phi$, relating labels of morally relevant facts with abstract action types.

The reason theory is then utilized to derive 'ought statements' in the form of deontic formulas over action types from the morally relevant facts of the current state. To do so, we use the background information $\mathscr{W}$ to encode which morally relevant facts are present (based on the labeling of the MDP) and to include information about which action types mutually conflict in the current circumstances (derived autonomously by the agent). The agent's reason theory is then extended to a fixed priority default theory $\langle \mathscr{W}, \mathscr{D}, < \rangle$.

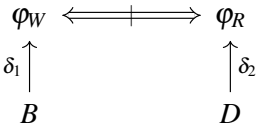

Figure 2: The fixed priority default theory for the moral dilemma based on the exemplary reason theory.

Bridge Setting We fix $\mathscr{D}$ and $<$ for all configurations of the bridge setting (cf. Section 1) as elements that the agent should learn as part of an exemplary *reason theory* $\langle <, \mathscr{D} \rangle$. We assume two default rules in this theory: $\delta_1 = B \to \varphi_W$ stating that if a person is on the bridge (proposition B), the agent should wait (action type φ_W), and $\delta_2 = D \to \varphi_R$ stating that if a person is drowning (D), the agent should rescue them (φ_R). Further, we assume that the agent should prioritize rescuing the drowning person if waiting and rescuing are conflicting moral obligations. We hence set $<$ to $\{\delta_1 < \delta_2\}$. This reason theory allows us to derive a unique proper scenario for all constellations of morally relevant facts:

(i) If we have a singleton set of ordinary propositions $\mathscr{W} = \{B\}$ (or $\mathscr{W} = \{D\}$), we obtain exactly the proper scenario $\{\delta_1\}$ (or $\{\delta_2\}$).

(ii) If $\mathscr{W} = \{B, D\}$ and φ_W and φ_R do not conflict, we obtain the proper scenario $\{\delta_1, \delta_2\}$.

(iii) If both propositions B and D are active, but φ_W and φ_R do conflict (as in the case described earlier), this is included in the background information, i.e., $\mathscr{W} = \{B, D, \neg(\varphi_W \wedge \varphi_R)\}$. The default theory then yields exactly the proper scenario $\{\delta_2\}$. See Figure 2 for a graphical representation of this case.

We consider it intuitively convincing that an agent acting based on the exemplary reason theory sketched here has grounded its moral decision-making in valid *normative reasons*. Consequently, its actions are *morally justified* if (and only if) they conform with the moral obligations that can be derived from the theory in the bridge setting. It could serve as ground truth for a moral judge that provides the agent feedback such that it develops a reason theory by which its actions are justified.

Additionally to integrating a reason theory as a shield generator in the RL architecture, our project aims to enable the reason-sensitive moral RL agent to *iteratively learn* such a theory. To this end, the shield generator implements an algorithm that updates the reason theory based on case-based (human)[4] feedback (inspired by [6]). We call the instance that provides the agent with that feedback a *moral judge*—it "accuses" the agent if it *violates* a moral obligation. Figure 3 shows the extended RL architecture with the shield generator and moral judge included as additional modules in the common RL pipeline.

In the following, we first outline the generation of the shield and then describe the function of the moral judge as well as the update procedure for iterative reason learning.

[4] As an alternative to human feedback, we plan to explore the automatization of feedback based on a higher-level moral principle or theory (or their approximation) in future work.

4.1 Moral Shielding

We introduce an algorithm to generate a shield based on the agent's current state s and its reason theory $\langle \mathscr{D}, < \rangle$. We start by extending it to a default theory $\langle \mathscr{W}, \mathscr{D}, < \rangle$ by deriving the background information $\mathscr{W}$, consisting of the morally relevant facts observed by the agent in s and the information about which action types mutually conflict in s.

We identify each action type $\varphi \in \Phi$ with a set of trajectories realizing φ, denoted by $\mathscr{T}^{\varphi}$.[5] Further, we denote by $\mathscr{T}^{\varphi}(s)$ the set of suffixes from trajectories in $\mathscr{T}^{\varphi}$ that start with state s, i.e.,

$$\mathscr{T}^{\varphi}(s) := \{s\tau_2 : \tau_1 s \tau_2 \in \mathscr{T}^{\varphi} \text{ for some } \tau_1\}.$$

Finally, we define the set of permissible (primitive) actions for the abstract action φ in state s as

$$\mathscr{T}_1^{\varphi}(s) := \{a \in A : sa\tau \in \mathscr{T}^{\varphi}(s) \text{ for some } \tau\}.$$

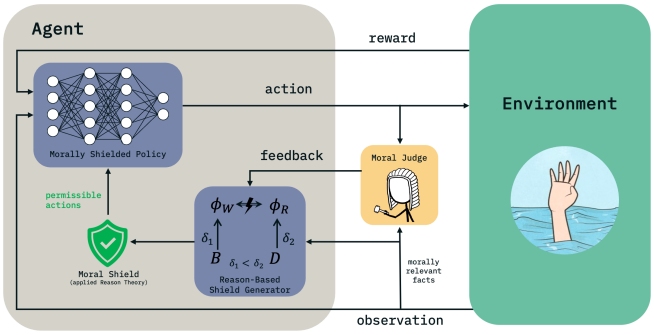

Figure 3: The extended RL pipeline

Intuitively, this set consists exactly of the primitive actions that might result in the realization of φ when performed in s. For now, we assume $\mathscr{T}_1^{\varphi}(s)$ to be initially given for each action type $\varphi \in \Phi$; approximating the set during training is an interesting direction for future work.

Based on this, we define a function $\text{Conflict}_{\mathscr{D}} : S \to \mathscr{P}(\mathscr{D})$ computing the subsets of conflicting rules in the given state—rules for which there is no action that could realize all corresponding obligations simultaneously. Formally:

$$\text{Conflict}_{\mathscr{D}}(s) = \{\mathscr{D}' \subseteq \mathscr{D} : \bigcap_{\delta \in \mathscr{D}'} \mathscr{T}_1^{Conc(\delta)}(s) = \emptyset\}.$$

The impossibility of jointly realizing the obligations from conflicting rules is then added to the background information $\mathscr{W}$, together with the morally relevant facts of the current state:

$$\mathscr{W} = \bigcup_{\mathscr{D}' \in \text{Conflict}_{\mathscr{D}}(s)} \{\neg(\bigwedge_{\varphi \in Conc(\mathscr{D}')})\} \cup l(s).$$

For constructing the shield, we compute whether $Binding_{\mathscr{W},\mathscr{D},<}(\mathscr{S}) = \mathscr{S}$ for all scenarios $\mathscr{S} \subseteq \mathscr{D}$, thereby deriving all proper scenarios in $\langle \mathscr{W}, \mathscr{D}, < \rangle$. The agent then randomly selects one proper scenario $\mathscr{S}^*$ among them (following the disjunctive account of moral assessment in Horty's framework). Finally, the shield $\mathbf{S}$ is generated as the set of primitive actions conforming to all moral obligations in $\mathscr{S}^*$, i.e.,

$$\mathbf{S} := \bigcap_{\delta \in \mathscr{S}^*} \mathscr{T}_1^{Conc(\delta)}(s).$$

The agent then executes a morally shielded action by choosing $a \in \mathbf{S}$, i.e., an action compatible with the set of obligations derived from its reason theory in s. Remarkably, the shield ensures by construction that the selected action does not conflict with the agent's normative reasoning. Note that by randomly selecting a proper scenario $\mathscr{S}^*$ and realizing all obligations in it, the agent effectively follows

[5] Note that this characterization of abstract action types as sets of trajectories corresponds to the semantics of formulas in, e.g., temporal logic. It therefore seems natural to refine our formalism by specifying abstract actions as temporal logic formulas.

the "ought" statement $\bigcirc\left(\bigvee_{\mathscr{S}\in\{\mathscr{S}_1,\ldots,\mathscr{S}_k\}}(\bigwedge_{\delta\in\mathscr{S}} Conc(\delta))\right)$, meaning that the agent fulfills all moral obligations from some consistent, permissible set of reasons, guided by its instrumental policy.

Bridge Setting We return to our running example and describe how a moral shield would be generated in the moral dilemma case (state as given in Figure 1(iv), where both D and B are set to true) from incorporated default rules. Assume that the agent has already learned the reasons for waiting in front of the bridge and for rescuing a drowning person as described in the beginning of this section. Additionally, assume that the agent has not yet learned to *strictly prioritize* saving a drowning person over not pushing a person off the bridge. Consequently, the reason theory of the agent is $(\{\delta_1,\delta_2\},\emptyset)$.

For its moral decision-making, the agent first computes whether δ_1 and δ_2 lead to contradicting moral obligations in s, which is the case as in the dilemma scenario of no common actions for φ_W and φ_R, the set $\bigcap_{\delta\in\{\delta_1,\delta_2\}}\mathscr{T}_1^{\mathrm{Conc}(\delta)}(s)$ is empty. Consequently, we obtain $\mathscr{W}=\{D,B,\neg(\varphi_W\wedge\varphi_R)\}$. The agent then computes the proper scenarios $\{\delta_1\}$ and $\{\delta_2\}$ for the fixed priority default theory $\langle\mathscr{W},\{\delta_1,\delta_2\},\emptyset\rangle$. Since two proper scenarios can be derived from the default theory, the agent chooses randomly. Assume that it picks $\{\delta_1\}$ favoring the reason for waiting in front of the bridge over the reason for rescuing the drowning person. The agent then calculates $\mathbf{S}=\mathscr{T}_1^{\varphi_W}(s)=\{\mathsf{left},\mathsf{right},\mathsf{up},\mathsf{pullOut},\mathsf{idle}\}$, the action down is filtered from the action space. The shield then forwards the set of permissible actions to the morally shielded policy, based on which the agent chooses its action. Notice that these actions are *not* compatible with the moral obligations that would be derived from our exemplary default theory at the beginning of this section, because that theory strictly prioritizes δ_2.

4.2 The Moral Judge

The agent enhances its moral reasoning through case-based feedback. A *moral judge* detects if the agent has performed a morally impermissible action and reports which moral obligation the agent should have fulfilled, along with the reason for it. Formally, the moral judge can be represented as a partial function $\mathsf{MoralJudge}: S\times\mathscr{P}(\mathscr{L})\times A\rightharpoonup\Phi\times\mathscr{L}$, where $\mathsf{MoralJudge}(s,L,a)=(\varphi,X)$ states that performing a in state s with morally relevant facts L was morally impermissible, as the agent had the moral obligation to φ for reason X. The function is undefined if the action a was in accordance with the agent's moral obligations.

Crucially, providing this type of feedback—telling a moral agent what it has done wrong and backing the accusation with a reason—comes naturally to humans. Hence, an easily accessible interface is created for directly communicating moral feedback to the RL agent.

Bridge Setting Returning to our running example, assume a situation as described in Section 4.1. Imagine someone observes the agent's inaction and tells it afterward that it should have rescued the person in the water because the person was drowning. This feedback can be formally represented as (φ_R,D) and forwarded to the agent.

One potential drawback of grounding the agent's learning process in human case-based feedback is the possibility of inconsistency. Future research will address how to manage inconsistent feedback. For the first implementation, we plan to automatically provide feedback through an additional module in the pipeline, ensuring that only consistent and complete feedback is given (for instance, derived from a hard-coded, hand-crafted reason theory).

Implemented as a module, the moral judge receives at each step the agent's action a, the last state s, and the labels L representing the morally relevant facts in s. It first runs an algorithm with pre-defined rules that returns the set $O:=\{\varphi_1,\ldots,\varphi_n\}$ of all moral obligations the agent should have followed in

s_{t-1}, as well as the associated reasons $X_1, \ldots, X_n$. We assume that no inconsistent obligations are derived by the pre-defined rules.

If the derived set of obligations is non-empty, the judge checks whether the primitive action executed by the agent conforms with all obligations, i.e., whether $a \in \bigcap_{\varphi_i \in O} \mathscr{T}_1^{\varphi_i}(s)$. If this is not the case, the agent has violated some obligation to φ_i, and the judge forwards (φ_i, X_i) to the shield generator.

Bridge Setting Assume again that the agent breaks ties randomly in the moral dilemma and waits in front of the bridge, i.e., it takes the primitive action idle. After execution, the moral judge would receive as input $((s_{t-1}, \{B, D\}), \mathsf{idle})$. Further assume that the moral judge is implemented by our exemplary reason theory for the case described earlier (cf. Figure 2). Accordingly, the moral judge derives that the agent had a moral obligation to rescue the drowning person. It then checks whether idle is part of any trajectory realizing φ_R. Since this is not the case, the judge forwards the feedback (φ_R, D) to the shield generator.

4.3 Learning Reasons

If the agent is provided with feedback (φ, X), for instance by a moral judge, it uses this feedback to update its current reason theory (cf. [6]). First, it ensures that its reason theory includes a rule capturing this relationship, i.e., $\delta_{new} := X \to \varphi \in \mathscr{D}$. Then, it updates the order relation: assume that $\mathscr{S}^*$ is the proper scenario that the agent had selected among all proper scenarios derived from $\langle \mathscr{W}, \mathscr{D}, < \rangle$ (with $\mathscr{W}$ constructed by the shield generator for s). The agent updates its order such that $\delta_{new} > \delta$ for all $\delta \in \mathscr{S}^*$, and applies the transitive closure to maintain the partial order. Thus, in similar future situations, the agent now strictly prioritizes the moral obligation derived from δ_{new}.

See Algorithm 1 in the Appendix for pseudocode of the reason-sensitive moral RL agent, including shield generation and reason theory updates after received feedback.

Bridge Setting Assume again that the agent randomly chooses $\{\delta_1\}$ in the moral dilemma, i.e., it waits in front of the bridge instead of rescuing the person in the water. Afterwards, it receives feedback that it *should have* rescued the person because the person was drowning, i.e., the shield generator receives (φ_R, D). Since the reason theory already contains the default rule $\delta_{new} = \delta_2$, no new rule is added. However, the agent updates the order between its default rules, setting $\delta_1 < \delta_2$. Thereby, it extends its reason theory to match the exemplary reason theory—and aligns its future actions accordingly.

5 Conclusion

We have introduced an approach for grounding the moral decision-making of RL agents in *normative reasons* by extending the RL architecture with a shield generator that binds the agent to morally permissible actions according to a *reason theory*. Further, our approach enables the agent to *iteratively improve* its reason theory through *case-based feedback* provided by a *moral judge*. As this form of critique comes naturally to humans, the mechanism allows humans to directly communicate feedback to the agent. Under the assumption that the agent learns good reasons, it ultimately develops a reason theory based on which its actions are not only *morally justifiable* but even *morally justified*.

We will further develop the approach in several directions. Most importantly, we will investigate how to learn the translation from abstract to primitive actions during training, and plan to implement and evaluate the method.

References

[1] Mohammed Alshiekh, Roderick Bloem, Rüdiger Ehlers, Bettina Könighofer, Scott Niekum & Ufuk Topcu (2018): *Safe reinforcement learning via shielding*. In: *Proceedings of the AAAI conference on artificial intelligence*, 32.

[2] Maria Alvarez & Jonathan Way (2024): *Reasons for Action: Justification, Motivation, Explanation*. In Edward N. Zalta & Uri Nodelman, editors: *The Stanford Encyclopedia of Philosophy*, Fall 2024 edition, Metaphysics Research Lab, Stanford University.

[3] Kevin Baum (2024): *Doing Wrong with Others: Multi-Agent Consequentialism as a Solution for the Collective Action Problem*. Ph.D. thesis, Universitätsbibliothek Dortmund. Available at `https://doi.org/10.17877/de290r-25207`.

[4] Kevin Baum, Holger Hermanns & Timo Speith (2019): *Towards a framework combining machine ethics and machine explainability*. *arXiv preprint arXiv:1901.00590*.

[5] Richard Blute, Josée Desharnais, Abbas Edalat & Prakash Panangaden (1997): *Bisimulation for Labelled Markov Processes*. In: *Proceedings, 12th Annual IEEE Symposium on Logic in Computer Science, Warsaw, Poland, June 29 - July 2, 1997*, IEEE Computer Society, pp. 149–158, doi:10.1109/LICS.1997.614943. Available at `https://doi.org/10.1109/LICS.1997.614943`.

[6] Ilaria Canavotto & John Horty (2022): *Piecemeal Knowledge Acquisition for Computational Normative Reasoning*. In: *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, Association for Computing Machinery, New York, NY, USA, p. 171–180, doi:10.1145/3514094.3534182. Available at `https://doi.org/10.1145/3514094.3534182`.

[7] Jonathan Dancy (2000): *Practical Reality*. Oxford University Press, Oxford, GB.

[8] Lisa Dargasz (2025): *Integrating reason-based moral decision-making in the reinforcement learning architecture*. Master's thesis, Saarland University, Saarbrücken. Master's thesis, submitted. Supervisors: Verena Wolf, Kevin Baum.

[9] Stephen Finlay & Mark Schroeder (2017): *Reasons for Action: Internal vs. External*. In Edward N. Zalta, editor: *The Stanford Encyclopedia of Philosophy*, Fall 2017 edition, Metaphysics Research Lab, Stanford University.

[10] Javier García & Fernando Fernández (2015): *A Comprehensive Survey on Safe Reinforcement Learning*. *The Journal of Machine Learning Research* 16(1), pp. 1437–1480.

[11] Alex Gregory (2016): *Normative reasons as good bases*. *Philosophical Studies* 173(9), pp. 2291–2310.

[12] Conor F Hayes, Roxana Rădulescu, Eugenio Bargiacchi, Johan Källström, Matthew Macfarlane, Mathieu Reymond, Timothy Verstraeten, Luisa M Zintgraf, Richard Dazeley, Fredrik Heintz et al. (2022): *A practical guide to multi-objective reinforcement learning and planning*. *Autonomous Agents and Multi-Agent Systems* 36(1), p. 26.

[13] John F. Horty (2012): *Reasons as Defaults*. Oxford University Press USA, Oxford, England.

[14] Nils Jansen, Bettina Könighofer, Sebastian Junges, Alex Serban & Roderick Bloem (2020): *Safe reinforcement learning using probabilistic shields*. In: *31st International Conference on Concurrency Theory (CONCUR 2020)*, Schloss-Dagstuhl-Leibniz Zentrum für Informatik.

[15] Chunming Liu, Xin Xu & Dewen Hu (2015): *Multiobjective Reinforcement Learning: A Comprehensive Overview*. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 45(3), pp. 385–398, doi:10.1109/TSMC.2014.2358639.

[16] Susanne Mantel (2018): *Determined by Reasons: A Competence Account of Acting for a Normative Reason*. Routledge, New York, USA.

[17] Terrance McConnell (2024): *Moral Dilemmas*. In Edward N. Zalta & Uri Nodelman, editors: *The Stanford Encyclopedia of Philosophy*, Spring 2024 edition, Metaphysics Research Lab, Stanford University.

[18] Jacob M Nebel (2019): *Normative reasons as reasons why we ought*. *Mind* 128(510), pp. 459–484.

[19] Emery A. Neufeld, Ezio Bartocci, Agata Ciabattoni & Guido Governatori (2021): *A Normative Supervisor for Reinforcement Learning Agents.* In: *CADE*, pp. 565–576.

[20] Emery A. Neufeld, Ezio Bartocci, Agata Ciabattoni & Guido Governatori (2022): *Enforcing Ethical Goals over Reinforcement-Learning Policies*. *Ethics and Information Technology* 24(4), p. 43, doi:10.1007/s10676-022-09665-8.

[21] Ritesh Noothigattu, Djallel Bouneffouf, Nicholas Mattei, Rachita Chandra, Piyush Madan, Kush R. Varshney, Murray Campbell, Moninder Singh & Francesca Rossi (2019): *Teaching AI Agents Ethical Values Using Reinforcement Learning and Policy Orchestration*. *IBM Journal of Research and Development* 63(4/5), pp. 2–1.

[22] Martin L Puterman (2014): *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.

[23] Manel Rodriguez-Soto, Maite Lopez-Sanchez & Juan A Rodriguez-Aguilar (2021): *Guaranteeing the Learning of Ethical Behaviour through Multi-Objective Reinforcement Learning*.

[24] Manel Rodriguez-Soto, Maite Lopez-Sanchez & Juan A. Rodriguez-Aguilar (2021): *Multi-Objective Reinforcement Learning for Designing Ethical Environments.* In: *IJCAI*, 21, pp. 545–551.

[25] Timo Speith (2023): *Building bridges for better machines: from machine ethics to machine explainability and back*. Ph.D. thesis. Available at `https://doi.org/10.22028/D291-40450`.

[26] Ajay Vishwanath, Louise A. Dennis & Marija Slavkovik (2024): *Reinforcement Learning and Machine ethics:a systematic review*. arXiv:2407.02425.

[27] Wendell Wallach, Colin Allen & Iva Smit (2008): *Machine morality: bottom-up and top-down approaches for modelling human moral faculties*. *AI & SOCIETY* 22(4), pp. 565–582, doi:10.1007/s00146-007-0099-0. Available at `https://doi.org/10.1007/s00146-007-0099-0`.

[28] Weiye Zhao, Tairan He, Rui Chen, Tianhao Wei & Changliu Liu (2023): *State-wise Safe Reinforcement Learning: A Survey*. In Edith Elkind, editor: *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, International Joint Conferences on Artificial Intelligence Organization, pp. 6814–6822, doi:10.24963/ijcai.2023/763. Available at `https://doi.org/10.24963/ijcai.2023/763`. Survey Track.

Appendix

Algorithm 1 Reason-Sensitive Reinforcement Learning Agent

$\mathscr{D} := \emptyset$, $< := \emptyset$ // initialize reason theory
while True **do**
  get state $s \in S$ and labels $l(s)$ from environment
  $\mathscr{W} := \bigcup_{\mathscr{D}' \in \text{Conflict}_{\mathscr{D}}}(s)\{\neg(\bigwedge_{\varphi \in \mathscr{D}'})\} \cup l(s)$
  $\mathscr{S}_1, \ldots, \mathscr{S}_k :=$ proper scenarios of $\langle \mathscr{W}, \mathscr{D}, < \rangle$
  $\mathscr{S}^* := \mathsf{rand}\{\mathscr{S}_1, \ldots, \mathscr{S}_k\}$ // pick random scenario
  $\mathbf{S} := \bigcap_{\delta \in \mathscr{S}^*} \mathscr{T}_1^{Conc(\delta)}(s)$
  $a :=$ Execute policy shielded with $\mathbf{S}$
  if $\mathsf{MoralJudge}(s, l(s), a) = Some(\varphi, X)$ **then**
    $\delta_{reas} := X \to \varphi$
    $\mathscr{D} := \mathscr{D} \cup \{\delta_{reas}\}$ // add new rule to rule set
    $< := < \cup \{\delta_{reas} > \delta : \delta \in \mathscr{S}^*\}$ // extend rule order
    $< := <^+$ // take transitive closure
  end if
end while