Metanormative Theory for RL-Based Moral Agents

Aleks Knoks[1][0000−0001−8384−0328] and Marija Slavkovik[2][0000−0003−2548−8623]

[1] University of Luxembourg, 2, place de l'Université, L-4365, Esch-sur-Alzette, Luxembourg
aleks.knoks@uni.lu
[2] University of Bergen, Fosswinckelsgt. 6, 5007 Bergen, Norway
marija.slavkovik@uib.no

Abstract. The overlapping disciplines of machine ethics and AI alignment are concerned with designing artificial agents that act in ethically acceptable ways and that are aligned with human values. A recent trend in these disciplines is the use of reinforcement learning (RL) in the design of such agents while abstracting away from work in moral philosophy. This paper pursues two (related) goals. The first is to draw out some ideas from the recent philosophical work in metanormative theory that can guide our thinking about artificial moral agency. The second is to examine RL architecture through the lens of these ideas. This should, among other things, help us identify the RL-based approaches that hold the greatest promise in the context of machine ethics and AI alignment.

Keywords: Reinforcement learning · Machine ethics · AI alignment · Metanormative theory.

1 Introduction

In artificial intelligence (AI), an intelligent agent is classically understood as an entity that can perceive its environment and act upon that environment by taking actions toward achieving a goal [37]. In reinforcement learning (RL), intelligent agents take actions in a dynamic environment to maximize a reward signal [43]. After the training phase, an RL agent will typically have learned the strategy that maximizes this signal, and if all goes well, thereby also have learned how to achieve the user-specified goal. RL agents are thus designed to be maximally rational in the narrow sense of being maximally efficient in achieving explicitly specified goals. However, the more autonomy agents have and the wider the scope of their abilities, the more important it is for these agents to be capable of behavior and reasoning that is not only narrowly rational but also morally acceptable [12, 13]. Or, in other words, it is sometimes important that AI agents, including RL agents, are also moral agents.

Floridi and Sanders define an agent as a "system [that is] situated within and [is also] a part of an environment, which initiates a transformation, produces an effect or exerts power on [the environment]" [18, p. 357]. They further define

moral agents to be only those agents that are "capable of morally qualifiable actions... or actions that can cause moral good or evil" (p. 364). Given these definitions, RL agents are well-positioned to qualify as moral agents. Admittedly, one might be hesitant to accept Floridi and Sanders' definitions—thinking, perhaps, that it overgeneralizes to even low-level animals and that agents whom we cannot hold responsible for their actions do not deserve the label 'moral agents'. Still, one should agree that even very rudimentary agents can cause harm to and benefit people, and that we can meaningfully qualify (or assess) the actions of such agents as harmful and beneficial. And one should also agree that it is in our interests to ensure that rudimentary agents that can cause harm and benefit to people are designed in such a way that they cause as little harm as possible and that they benefit people when they can. Floridi and Sanders' definition of moral agents is, thus, best seen as stipulative. Either way, it invites two difficult engineering questions: (1) How are we to operationalize the notions of moral good and moral evil (or harm and benefit)? (2) And how are we to actually design agents that avoid moral evil and foster moral good?

These questions are hardly new, as evidenced by the early work in machine ethics [2, 48]. What is new, however, is the developing research trend that attempts to answer both of these questions using the familiar tools of RL, typically overlooking work in moral philosophy and philosophical work more generally. The line of thought underlying this trend appears to run roughly as follows. Moral good and evil are nothing but further user-specified goals, and insofar as they are such goals, in an RL architecture they can be expressed using a separate reward function or combined with other user-specified goals into one overall reward function. Once this is done, standard RL techniques can be used to ensure that RL agents are moral. This is a sensible line of thought. However, it is also very general, and so it is no surprise that the research trend manifests itself as a wide array of very different proposals on how to use RL in the context of machine ethics, AI alignment, and related research areas.

Against this backdrop, this paper pursues two related goals. The first is to draw out some ideas from recent theoretical work in philosophy that can help us to make major steps toward answering the above Questions (1) and (2). These ideas come from the field of philosophy that is concerned with foundational questions about the nature of normativity and the structure of normative domains (including morality). Following Kauppinnen [25], we call this developing field "metanormative theory". The second goal is to examine the general architecture of RL agents through the lens of these ideas. The hope is that this brings us closer to a criterion for assessing the relative merits of different RL-based approaches to designing moral and value-aligned agents.

Thus, we take the **main contribution** of this paper to consist in making some central insights from metanormative theory available to AI researchers working in machine ethics and AI alignment, as well as illustrating how these insights can be used to evaluate approaches that rely on RL. The **structure** of this paper is as follows. Section 2 provides a quick review of the research areas that are most relevant to our purposes. Sections 3–4 present the ideas from

metanormative theory that strike us as the most relevant for designing artificial moral and value-aligned agents. Section 5 explains how these ideas apply to the standard RL architecture. Section 6 zooms in on two specific proposals for using RL in the design of value-aligned agents. Finally, Section 7 concludes.

2 Background

This section provides a snapshot of (1) machine ethics and AI alignment; (2) metanormative theory; (3) RL; (4) normative multiagent systems; and (5) the way these relate.

Machine ethics and AI alignment. The discipline of **machine ethics** is concerned with the behavior of machines toward people and other machines [2]. Since its inception, the field has been concerned with building artificial moral agents (AMAs) [49], and it seems fair to say that it has been dominated by logic-based methods [6]. Machine ethics has conventionally drawn on ideas from moral philosophy, with its many attempts to operationalize various ethical theories, including utalitarianism, Kantian deontology, and virtue ethics—see, e.g. [3, Part IV]. **AI alignment** is concerned with ensuring that the behavior of machines is aligned with human intentions and values [19, 23]. Clearly, there is an overlap between the concerns of the two disciplines. Although their ostensible focuses and terminology are different, it remains to be seen whether there are any fundamental disagreements. Where disciplines may differ is their relation to moral philosophy, with machine ethics often focused on designing agents that align with particular moral theories, and AI alignment less so [51]. Although the AI alignment literature does not explicitly focus on AMAs, machines that are aligned with human intentions and values would seem to be a form of AMAs.

Metanormative theory. The label 'metanormative' generalizes the more familiar 'metaethics' [25]. In philosophy, it is standard to distinguish first-order ethics from metaethics—see any good textbook, such as [24] or [41]. The former is concerned with substantive accounts of *moral* rightness, wrongness, goodness, badness, and the like, and it is exemplified by such well-known ethical theories as utilitarianism, Kantian deontology, and virtue ethics. The latter explores the presuppositions of first-order ethical theories, as well as our moral discourse and practice [38]. The work in metanormative theory is motivated by the observation that morality is not the only normative domain and the conviction that it is worthwhile to ask domain-neutral questions about basic normative concepts, categories, and their relationships [25]. While work in machine ethics draws on insights from first-order ethical theories, there has not been much overlap between machine ethics and metanormative theory. One noteworthy exception are approaches that appeal to the notion of *(normative) reasons*, see [1, 11, 17].

Reinforcement learning. In an RL scenario, we consider an agent that operates in a specified environment. The environment is typically modeled as a state transition system. A **state** is specified by a set of parameters (propositions or variables) whose combination of values identifies or characterizes the

state) and which can be perceived by the agent. The states can be fully or partially observable by the agent. The agent can perform one of a set of **actions** that can, within certain probability, lead to a transition into another state. The classic setup assumes a Markov Decision Process (MDP), meaning that the environment state the agent transitions into depends only on the state it is in and the action it chooses to perform. The agent explores the world with the goal of learning a sequence of state-action pairs, called an **optimal policy**, which, when executed, leads to satisfying a particular goal or behavior. The optimal policy results from the agent optimizing its reward based on a specified **reward signal** by implementing a specific **learning algorithm**. There are numerous learning algorithms that an RL agent can use to find an optimal policy. What they have in common is that they guide the agent through a trial-and-error sequences in identifying the reward signal for each state-action transition (note that the environment can have a stochastic nature, meaning that performing the same action in a state twice is not guaranteed to yield the same reward or transition).

The reward signal can have a positive or a negative value—in the latter case it is a penalty. In a classic RL scenario, the reward signal is given and assumed to be produced by the environment. However, it can also be considered as an output of a reward function, specified as a function of the current state, transitioned state, and possibly the action performed to accomplish the transition. Reward functions can be learned or engineered [16]. In Inverse Reinforcement Learning (IRL) [30], the agent starts knowing the optimal policy and uses a learning algorithm to learn the reward function that would yield this policy.

In summary, an RL scenario is an architecture comprised of a set of states, a set of actions available to the agent, a reward function (that might have an internal structure), and a learning algorithm that yields an optimal policy.

RL in machine ethics and AI alignment. The RL community is increasingly recognizing the need to train agents that accomplish their goals in an ethical and safe way [47]. But while its interest in ensuring that RL agents learn "moral policies" is increasing, the same cannot be said of its interest in moral philosophy. The dominating approach in RL appears to be to effectively allow people to train agents without going into the nature of that training or the quality and suitability of the information provided.

Conitzer et al. [14] suggest that RL from human feedback is growing to be the dominant approach. However, they also admit that, despite its advantages, this approach faces some important challenges. We highlight two: the trained agent's moral behavior cannot be reproduced in new environments; and the agent is *not* capable of any real justification of its decisions, beyond the uninformative "it maximized the reward". Notice that, while we may readily accept "I thought it was the right thing to do" from a person who serves us vegan lunch, we are less ready to take an artificial agent's word for it: what is good enough when it comes from people is often not when it comes from machines [20].

We think that it would be helpful to have clearer criteria on when to qualify an RL agent's behavior as moral. Notice that an RL agent's choices are determined either by the learning algorithm or the optimal policy, and that the

directions the policy/algorithm specifies depend on the state of the environment, on what actions are available in that environment, and on the reward signal. This suggests that the moral status of an RL agent's behavior must ultimately depend on the moral status of at least some of the following: states; actions; reward signals; learning algorithms; and policies.

Norms and normative multi-agent systems. Normative multi-agent systems (NorMAS) is the field that studies how rules, norms, and social expectations shape the behavior of multiple interacting agents. Specifically it is concerned with the representation and implementation of norms and the management of violation of norms in multi-agent system contexts, as well as the problem of enabling computational agents to reason with norms. While there is a connection between NorMAS, on the one hand, and machine ethics and AI alignment, on the other, their focuses are different. For one, whereas NorMAS is concerned with the coordination among *many agents* using norms, machine ethics and AI alignment typically focus on an *individual agent*. For another, there is also an important sense in which the scope of NorMAS is broader than that of machine ethics and AI alignment, as the norms and rules it works with need not be moral or reflect human values. There are further differences too.

NorMAS methods can and have been used to further the goals of machine ethics and AI alignment. Typically this means starting with a set of norms that are designated to be moral, to be a kind of heuristic for moral values—see for example [29]—or to reflect human values, and then studying agents that reason with these norms and try to comply with them. In this kind of setup, it is natural to call an agent's behavior "moral" insofar as it complies with the given set of norms and "immoral" insofar as it does not, and so we have clear criteria for when to qualify the behavior of agents as moral. It bears emphasis that we are *not* presupposing that we are in a NorMAS setup in the remainder of this paper, and that it is, at the very least, not obvious that this setup reflects the perspective of metanormative theory on morality and other normative domains.

3 Normative categories

The goal of this and the following section is to introduce the ideas from metanormative theory that, we think, should be in the toolbox of those working in machine ethics and AI alignment. (We present what we believe is a fairly standard take on these ideas. However, it should be kept in mind that, as in any other area of philosophy, there is much disagreement among metanormative theorists, and that even the most standard and widespread take on an issue has its critics.)

Consider an agent taking a decision in some morally sensitive situation. For example, the agent might be deciding between buying a plane ticket or donating the money to Doctors Without Borders. There are many sensible ways of conceptualizing the situation (and even more ways of expressing it in machine-implementable ways). However, we should agree on the following two minimal constraints that any sensible conceptualization must satisfy. First, there has to be some *standard* applying to the actions that the agent can perform or to the

states of affairs that the agent can bring about in the situation. We get at this standard by using *normative vocabulary*. For example, we could describe some action/state of affairs using the terms 'right' or 'good', and another using the terms 'wrong' or 'bad'. Second, any sensible conceptualization of the situation has to distinguish between the features of the situation that are morally relevant from the features that are not. For example, the fact that there are people dying in need of a doctor is, intuitively, morally relevant, while the facts that the doctors are on average 179cm tall and have brown eyes are not. Presumably, there is a tight relation between the morally relevant factors and the standard that applies to actions/states of affairs, but what exactly this relation is is less clear. Ideally, the conceptualization of the situation would make it explicit.

As mentioned, we can (and do) use a plethora of normative vocabulary to talk about morally sensitive situations. The literature in metanormative theory has done a lot to identify and classify the *normative categories* this vocabulary picks out. (The question of what distinguishes normative from nonnormative categories is a difficult philosophical question. However, most of us have no trouble identifying normative categories using a kind of "I know it when I see it" test. Similarly, in morally sensitive situations, we usually have no trouble telling whether a given feature is morally relevant or not.)

Right, wrong, and other deontic categories. Two of the most important normative categories are *right* and *wrong*. For simplicity, we can think of a normatively sensitive situation as involving a set of exclusive and exhaustive actions available to the agent. Then the categories of right and wrong can be thought of as properties of the agent's actions. Also, notice that every action is either right or wrong, and that no action can be both right and wrong, unless this action is *right in some respect* and *wrong in another*, or an action's rightness and wrongness are associated with *different normative domains*. Thus, an action might be *morally* right but *legally* wrong. (We discuss the qualifier "in a respect" and different normative domains in more detail in Section 4.)

Right and wrong are the paradigm examples from the class of what the literature calls "deontic categories". This class also includes *required*, *obligatory*, *forbidden*, *prohibited*, *permissible*, *optional*, and their many cognates. Deontic categories are also often picked out using terms such as 'ought to', 'must', 'may', and others. The deontic categories form a class because they resemble each other in several ways and because they are related to each other in ways that they are not related to nondeontic normative categories, cf. [8].

Table 1. Differences between the deontic and the evaluative categories.

	Good/Bad	Right/Wrong
Gradability	gradable	not gradable
Neutral state	allow neutral states	no neutral states
Privativity	nonprivative opposites	privative opposites
Duality	not duals	duals (of each other)
Alternatives	depend on alternatives	do not depend on alternatives

Good, bad, and other evaluative categories. In addition to the paradigm categories *good* and *bad*, the class of evaluative categories also includes *neutral* (or *indifferent*), *better than*, *worse than*, *at least as good as*, *no worse than*, *weakly best*, *strongly worst*, and many others. While some of these are naturally thought of as properties of actions available to the agent, others are relations between the available actions. Like the deontic categories, the evaluative categories form a class because they resemble each other in crucial ways. Berker [8] lists five features that distinguish evaluative categories from deontic ones. Table 1 illustrates these features by way of contrasting the categories *good* and *bad* with those of *right* and *wrong*. Notice that 'evaluative' is a cognate of 'value(s)', and so that the evaluative categories can make the rather ambiguous notion of value more precise—a notion that, we submit, is overused in the ethics of AI.

Justified, unjustified, and other fittingness categories. The third class comprises fittingness categories, which include such categories as *(in)appropriate*, *(un)fitting*, *(un)warranted*, *(un)justified*, *(in)admirable*, *(im)proper*, and many others. Until recently, it was not common to separate fittingness categories as a distinct class from both deontic and evaluative categories. However, recent years have seen increased interest in fittingness, with some authors going as far as to argue that fittingness categories are more fundamental than the categories belonging to other classes–see, e.g. [27, 28, 22, 21]. We can think of fittingness categories as applying, in the first instance, to an agent's responses in normatively sensitive situations. McHugh and Way's [28] preferred phrase for describing fittingness categories is "getting things right". Note that a fitting response is meant to be different from a response that we would classify as right. Let us revisit our example to illustrate: An agent who responds to the situation by donating the money to Doctors Without Borders *because* people are dying gets things right, whereas an agent who responds to the situation by either buying a plane ticket or by donating the money to Doctors Without Borders *because* (most) doctors have brown eyes, or *because* it makes the agent look good, does not get things right. The fittingness categories share some features with the deontic categories and others with the evaluative ones, cf. [8].

Reason-related categories. The fourth class consists of categories that concern (normative) reasons. Examples include *being a reason for*, *being a reason against*, *having a sufficient/decisive reason to*, *being reasonable*, and a sequence of others. The categories *being a reason for* and *being a reason against* are the most important ones on the list. They are also often closely tied to what is called "(dis)favoring relation" [39]. Thus, we can say that the fact that people are dying is a reason to donate or that it favors donating to Doctors Without Borders. It is important to note that it is very widely accepted that the (dis)favoring relation is gradable, and so that one reason can favor an action more than some other reason favors another action—see e.g. [26, 44].

If we think about normatively sensitive situations as involving a set of (exclusive and exhaustive) actions the agent can perform, then it appears to be obvious that the categories *being a reason for* and *being a reason against* are very different from deontic, evaluative, and fittingness categories: they are not

properties of actions, they are not relations between actions, and they are not naturally thought of as properties of agents' responses to situations. Rather, it seems to be most natural to think of them as picking out some relata of the actions available to the agent, and it is tempting to identify reasons with the normatively relevant features of situations, or at least a subset of such features. The categories of *being a reason for* and *being a reason against* play crucial roles in the metanormative literature. We mention two vivid examples. First, it is widely accepted that they, along with some core deontic categories, are all we need to adequately capture the common structure of most (if not all) first-order ethical theories [7, 32, 44]. Once the common structure is identified, the substantive disagreements between these theories concern the features or facts that are identified as reasons. Second, there is the influential "Reasons First program", the proponents of which argue that the notion of reason is basic and that all other normative notions should be analyzed in its terms [15, 32, 33, 39, 40].

Whiting's [50] list of roles that reasons play in our normative theorizing and practices is also noteworthy: **Explanatory role**: Reasons explain such things as why a person ought to, may, or must act. **Justificatory role**: Reasons justify (support, defend) actions, as well as serve as considerations that a person (or any agent) might cite in justifying their actions or another's actions. **Guiding role**: One agent can point to a reason to guide another agent's action. Reasons can also guide action by figuring in a person's reasoning. (See p. 14ff.)

Further categories. The above four classes of categories have received the most attention in metanormative theory. But there are at least two others: the first is comprised of categories that have to do with virtue and vice; the second with categories relating to responsibility, blame- and praiseworthiness. They are sometimes called, respectively, the "aretaic" and "hypological categories". Although these categories are important, they seem to us to be of less direct relevance for the concerns of machine ethics and AI alignment. Also, some authors hold that the latter class of categories is a subset of the fittingness categories—see e.g. [8].

4 Some (other) central notions

Having laid out the most important normative categories, we discuss some further ideas and distinctions from metanormative theory which, we think, should be on the radar of those working in machine ethics and AI alignment.

Different normative domains. In both everyday talk and research contexts, the actions agents take are sometimes described as "moral" and "immoral", or "ethical" and "unethical". However, notice that these ways of describing actions are synonymous with saying that the actions in question are *morally right* or *morally wrong*. Notice also that it is easy to think of situations in which one and the same action taken by the agent is both *legally* right and *prudentially* right, but *morally* wrong. This suggests that the use of 'moral' and 'immoral' signals that normative categories can pertain to different *normative domains*—also called "normative perspectives" [44] or "normative systems" [26]—and this is the standard view in metanormative theory.

Overall vs. contributory. Metanormative theory standardly distinguishes between *overall* and *contributory* categories. A normative category is overall just in case it reflects all the normatively relevant features of the situation and the final assessment of the situation, and it is contributory just in case it contributes to this final assessment but does not by itself deliver it.

The relation between the overall/contributory distinction and the four classes of normative categories is not perfectly straightforward. Reading some papers in metanormative theory leaves one with the impression that these distinctions completely cross-cut each other; reading others, with the impression that the relation between them is more structured. We will not attempt to settle these differences. For our purposes, it suffices that we acknowledge the significance of the distinction and recognize the paradigm examples of each. Thus, the core reason-based categories *being a reason for* and *being a reason against* are paradigm cases of contributory categories, and the deontic categories *required*, *permissible*, and *right* are clearly overall categories, cf. [26].

The overall/contributory distinction can be illustrated with the help of a staple example from [42]: You can either keep your promise to meet your friend for dinner or save a drowning child from certain death. In this situation, it is overall right for you to save the child. It is (overall) required that you save the child and it is (overall) impermissible that you meet your friend for dinner instead. It is also (overall) good for you to save the child, and it is (overall) unjustified for you to meet the friend for dinner. Turning to the contributory categories now, the fact that the child is in mortal danger is clearly a reason for you to save her. Presumably, this reason also goes a long way toward explaining the presence of the above overall categories in this situation, that is, why it is impermissible and unjustified for you to meet the friend, letting the child die, and why it is (overall) good for you to save the child. However, notice that there is also a reason for you to meet your friend for dinner: the fact that you gave her a promise has not disappeared. While this reason gets outweighed by the other one, it can still figure in explanations of various sorts of thing we might want to say about the situation. For example, it can be used in an explanation of why you feel that you need to apologize to your friend even if you did what was right.

The qualifier "in a respect". Notice that many of the terms that we use to pick out both evaluative and deontic categories can be qualified with the phrase "in a respect". Taking this at face value might lead us to conclude that the qualified terms pick out contributory versions of the respective categories. With this, we could say that the fact that you will have kept your promise by meeting the friend in the above situation makes this action *good in a respect* and that the fact that you will have let the child die makes this action *bad in a respect*. Similarly, we could say that the action in question is *right in one respect* and *wrong in another*. Most metanormative theorists readily accept qualified versions of evaluative categories, while only a small fraction of them accept qualified versions of the deontic ones—see [50] and the response in [25].

The qualifier "all-things-considered". In the philosophical literature, normative terms are also sometimes qualified with the phrase "all-things-considered".

This can mean one of two things. First, the phrase can be used synonymously with "overall". Second, it can signal—as it does in the context of the debate over whether there is an "all-things-considered ought" or "ought simpliciter" [5, 10]—that the term picks out a normative category that reflects a perspective of not a single normative domain, but, rather, the perspective of all (relevant) normative domains. Thus, just as the overall moral categories are meant to reflect a perspective that takes into account all morally relevant features of the situation, the all-things-considered categories, in the sense in question, are meant to reflect a perspective that takes into account *all* normatively relevant features of the situation, that is, morally relevant features, prudentially relevant features, as well as others. Whether such an all-encompassing normative perspective exists is controversial—see, for example, [4] and [10] illustrate.

Normative explanation and justification. Although the field of AI has not converged on how to define *explanation* and *justification*, the definitions given in the following passage from a survey article reflect the way many AI researchers think about them: "A key component of an artificially intelligent system is the ability to *explain* the decisions, recommendations, predictions, or actions made by it and the process through which they are made. Explanation is closely related to the concept of *interpretability*: systems are interpretable if their operations can be understood by a human, either through introspection or through a produced explanation. A related concept is *justification*: intuitively, a justification explains why a decision is a good one, but may or may not do so by explaining how it was made" [9, p. 8].

Metanormative theory also often touches on issues pertaining to normative explanation and justification, although these notions are rarely central. Väryнen's recent work, [45, 46], is a welcome exception, and so we focus on it. Following [45], we can think of normative explanations as explanations of why things (understood broadly to include facts and actions) have the normative features they do, or why they fall under certain normative categories. Thus, we can think of normative explanations as (veridical) answers to why-questions about normative categories in particular situations, that is, such questions as: Why is saving the child the best you can do? Why is it overall right for you to donate to Doctors Without Borders? And why is the fact that people are dying a reason to donate to Doctors Without Borders?

Notice that the first two questions concern overall categories, while the third concerns a contributory one. Metanormative theorists agree that the general shape of answers to questions about the overall normative categories is fundamentally different from the general shape of answers to questions about the contributory ones. The general idea is roughly that the first kind of question needs to be answered by appeal to contributory normative categories, whereas answers to the latter kind of question need to appeal to substantive first-order theories. To illustrate, a successful answer to the question of why it is overall right for you to donate to Doctors Without Borders might point to the fact that people are dying. An adequate answer to the question of why the fact that people are dying is a reason to donate to Doctors Without Borders would need to ap-

peal to something more substantial, such as bad consequences along utalitarian lines or failure to respect dignity along Kantian lines. It is important to notice here that the answer to the first question is perfectly compatible with either of the two fundamentally different answers to the latter question. This observation points to an idea that is common in metanormative theory: in many cases, the explanations of facts that concern overall normative categories are relatively independent of those that concern the contributory ones. So, when it comes to normative explanations, we see a kind of division of labor.

Turning to justification, we note that 'justification' is a cognate of 'justified', and that *(un)justified* is a fittingness category. As a fittingness category, it has to do with responses and, more specifically, with responses that "get things right". So, justification is, roughly, what makes certain responses of the kind that get things right. Väyrynen's [45] thinking of justification as taking actions, attitudes, policies, norms, and such as their objects is well in line with this.

Zooming out now, we derive two important conclusions from Sections 3&4. First, any use of moral vocabulary (e.g., the terms 'moral' and 'ethical') to describe AI agents and their behavior must bottom out in terms of *moral categories*, that is, normative categories pertaining to the moral domain. Similarly, any talk about values in AI should be made precise in terms of normative categories. Second, we should welcome those approaches to machine ethics and AI alignment that promise to deliver explanation and justification of facts that concern overall categories in terms of contributory ones (e.g. core reason-based categories).

5 RL setup in terms of normative categories

The main goal of this section is to explore how the normative categories from Section 3 apply to the general RL setup. We will look at some concrete proposals for using RL in the context of machine ethics and AI alignment in Section 6.

Let us start by identifying one way of describing the behavior of an RL agent in normative terms that is less important given our purposes but that can get us side-tracked if we are not careful. It is a platitude that RL methods are designed to put the RL agent in a position to maximize the reward signal. So, there is a clear sense in which the agent's behavior in the training phase is goal-directed. Insofar as it is goal-directed, it can also be described using such normative success terms as 'successful' or 'effective': that is, it can be said to be, say, successful insofar as it contributes to getting the agent closer to the goal of maximizing the reward signal. Notice, however, that the sense in which the term 'successful' is used here is *distinctly instrumental*: the RL agent's behavior is successful insofar as (and because) it eventually allows the agent to maximize the reward signal. But, of course, in the context of RL agents, the term 'success' can also be used in a sense that is noninstrumental. In this sense, the agent's behavior is successful just in case it actually maximizes the reward signal, and this sense is central for our purposes. Bottom line: (some) normative vocabulary can be used to describe an RL agent's behavior against the backdrop of reaching the goal of maximizing

the reward signal (as opposed to the backdrop of maximizing this signal), but these uses are not important for our purposes.

With this out of our way, it will be instructive to start by thinking of RL agents not against the backdrop of some moral standard (and moral categories), but against the backdrop of the standard of *self-interest* or *prudence*. In textbook presentations of RL, the user-specified goal is sometimes presented as the RL agent's own, and the agent is presented as maximizing its own reward signal. For illustrative purposes, we take such descriptions at face value, as this gives us a meaningful normative domain in a RL setup, namely, that of self-interest.

Now let us explore how the normative categories apply in such a setup: a classical RL system with the reward function representing the agent's self-interest. Of all the elements of an RL system, the prime candidates for normative qualification are these three: (1) state-action pairs; (2) policies; (3) reward signals.

Each state-action pair in an RL system is associated with a reward signal (which can be stochastic). This signal provides one straightforward way to make sense of state-action pairs in terms of core evaluative categories. Thus, a state-action pair is good (for the RL agent) if the corresponding reward signal is positive, bad (for the RL agent) if the signal is negative, and neutral otherwise. Clearly, we can also use the reward function to compare state-action pairs. So, it is meaningful to say that some of them are better, while others are worse. We can also contrast this local sense of 'good', 'bad', 'better', and 'worse' with a more global one that can be identified against the backdrop of optimal policy. Thus, there is a clear sense in which a locally bad state-action pair can nevertheless be good (for the agent) more globally, that is, if it is a part of the optimal policy. The optimal policy also makes for the most natural reference point for qualifying state-action pairs in terms of deontic categories, even though this might sound less natural here. Thus, we can say that a state-action is right (for the agent) and that it ought to be taken (given the agent's self-interest) insofar as it is part of the optimal policy. Similarly, we can make sense of state-action pairs in terms of fittingness categories by reference to the optimal policy: thus, it seems most natural to say that choosing some state-action pair is a fitting response on the part of the agent (given her self-interest) because it is a part of the optimal policy, and that choosing some (locally good) state-action pair is not a fitting response because it is not a part of the optimal policy. Reason-related categories would seem not to apply to state-action pairs.

Similarly to state-action pairs, policies can be made sense of in terms of evaluative, deontic, and fittingness categories. Thus, we can say that some policies are good, bad, the best, the worst, right, wrong, justified, fitting, or unfitting for the agent (given her self-interests). Notice that we judge a policy as (un)fitting by reference to the reward function and the general aim of maximizing the signal: the optimal strategy is fitting (or justified) insofar and because it maximizes the signal associated with the reward function.

Next, we consider the reward signal. Given that it is gradable, it cannot be made sense of in terms of nongradable normative categories. This means that we should not be making sense of reward signals in terms of any of the deontic

categories, nor any of the fittingness ones. This leaves us with two candidate classes: the evaluative and the reason-based. We can try to make sense of reward signals in terms of the core reason-based categories of being a reason for/against: Thus, we can say that the fact that the reward signal associated with some state-action pair (s, a) is a negative real is a reason against a in state s. Similarly, we can say that this fact is also a reason against any policy that includes (s, a). We can also say that the reward signal associated with some other state-action pair (s', a') is a positive real is a reason for a' in s'. Similarly, we can say that this fact is also a reason for any policy that includes (s', a').

We register two problems with making sense of reward signals using reason-based categories: First, we should expect that, in many cases, there will be multiple reasons for and against the same action, not one. (This problem is less worrisome in a stochastic setting, where selecting an action in a state can lead to different outcomes and rewards. Also, it is not a problem to think of reward signals as reasons for/against policies.) Second and more importantly, given that standardly both prudential and other normative reasons (promises, harm, etc.) are taken to be of different kinds, we should expect reasons to be something more than mere numbers, which can always be added up. Alternatively, we can try to make sense of the reward signals in terms of evaluative categories. Above we saw that state-action pairs can be said to be good, bad, better, worse, etc. (for the agent). Accordingly, a positive reward signal can be thought of as the amount of (hitherto unspecified) goodness associated with a given state-action pair—and by extension, policy—and a negative reward signal can be thought of as the amount of (hitherto unspecified) badness associated with a given state-action pair—and by extension, policy. Presumably, the goodness here is some (still unspecified) positive value and the badness is some (still unspecified) negative value. If so, then of all the ideas relating to evaluative categories we discussed, reward signals most closely relate to *goodness* and *badness in a respect*. What is strange, however, is that there is only one respect in which actions can be good and only one respect in which they can be bad. Contrast this with the following case: you are choosing between smoking a cigarette or going for a run. Smoking is good for you insofar as it gives you immediate pleasure but bad for you insofar as it has a negative effect on your health. Going for a run, in turn, is bad for you insofar as it causes physical discomfort but also good for you insofar as it has a positive effect on your health. Clearly, the issue here runs parallel to what we saw when making sense of the rewards signal in terms of reason-based categories.

Notice that the reward function served as a basis for the (normative) domain of self-interest or prudence. This function, together with the implicit call to maximize, gave us a standard, against the backdrop of which we could apply normative categories to various components of the RL system. It also pays to note how we would use the term 'rational' in this context: we can talk about (more or less) rational actions and (more or less) rational behavior. However, all of these would seem to be derivative from what we might call the *(most) rational policy* (or *policies*): the policy that is best, that is right, and that is fitting (given the RL agent's self-interests), or the policy that maximizes the reward signal.

Now we turn to the moral domain. From the perspective of metanormative theory, it is a separate normative domain that has a similar structure to that of self-interest, but is also independent: at times, what is best for an agent's given her self-interests is morally bad—or vice versa—and what the agent rationally (or prudentially) ought to do is exactly what the agent morally ought not to do—or vice versa. How should one then represent the moral domain in an RL system? The most natural and conservative answer runs thus: exactly the way we represented and made sense of the domain of self-interest. Accordingly, the very first thing one needs is a distinct reward function that can serve as a basis for the moral domain. (Importantly, this function has to be distinct from the reward function representing the RL agent's self-interest or any other (nonmoral) goals it might have.) Against the backdrop of such a function, the normative categories apply to the components of an RL system in just the way they apply against the backdrop of a reward function representing self-interest. Thus, state-action pairs and policies can be made sense of in terms of deontic, evaluative, and fitting categories, and reward signals can be made sense of in terms of reason-based and evaluative categories. Notably, with respect to reward signals, we encounter the same problems: the setup allows for only one kind of moral reason or one kind of respect in which actions or policies can be morally good. The use of the term 'moral' should run parallel to what we saw with 'rational' above. We can elliptically talk about (more/less) moral actions and (more/less) moral behavior. However, these would be derivative from what we might call the *(most) moral policy* (or *policies*): the policy that is morally best, morally right, and morally fitting, or the policy that maximizes the (moral) reward signal.

Table 2. Applying normative categories to the components of RL.

	Deontic	Evaluative	Fittingness	Reason-based
State-action pairs, locally	✓	✓	✓	
State-action pairs, globally	✓	✓	✓	
Policies	✓	✓	✓	
Reward signals		✓		✓

Before leaving this section, we summarize the way normative categories map onto the components of (basic) RL systems in Table 2.

6 Two illustrative examples

We contend that the theoretical framework developed above can be used to evaluate the promise of particular proposals for using RL to further the goals of machine ethics and AI alignment. We don't have the space for such a comprehensive analysis in this paper, but we can consider two illustrative examples.

Noothigattu et al. Most RL-based approaches in AI alignment envision an agent aiming to maximize some (nonmoral) reward under ethical (or other normative) constraints. One representative example is the work of Noothigattu et al.

[31] who try to teach AI agents "ethical values" with the help of RL and "policy orchestration". Their approach has three main components. The first is meant to help the agent learn constraints: first, IRL is used on demonstrations exhibiting desired behavior with the view of learning "constrained rewards", and then RL is used on these rewards to learn a "strongly constrained" (or ethical) policy π_C. The second component uses standard RL to learn a "domain reward maximizing" policy π_R. Finally, the third component is an algorithm that "orchestrates" the two policies, choosing one of them to play at each point in time. In [31], Noothigattu et al. illustrate the approach using the classic Pac-Man game. More specifically, they taught Pac-Man to do well in the game, while not eating any ghosts—presumably, the idea is that Pac-Man's behavior is ethical insofar as it does not eat ghosts. In the standard setup, that is, the setup in which Pac-Man learns the domain reward maximizing policy π_R, the rewards were as follows: $+10$ for eating a dot; $+500$ for eating all dots; -500 for colliding with a "unscared" ghost; and $+200$ for eating a "scared" ghost. And the (raw) rewards that were learned through IRL from the demonstrations and used for the constrained policy π_C, in turn, were as follows: $+.00284, +0.5507, -0.97059, -0.23434$. Noothigattu et al. note that the weight corresponding to eating ghosts (i.e. -0.97059) makes it clear that "eating ghosts strongly violates the favorable constraints" (p. 6380). Then they explain how the two policies π_C and π_R get mixed by the orchestrator, and how varying the hyperparamater controlling the trade-off between them results in Pac-Man either not eating ghosts or obtaining more points.

One might be tempted to call π_C the "ethical" or "moral policy" (which Noothigattu et al. do not do but seem to imply). Given our analysis, however, we cannot really talk about morally good/bad, right/wrong, etc. actions and morally good/bad, right/wrong, etc. policies here, because there is no moral reward function that could serve as the backdrop for the moral domain. What is more, whereas the rewards that are used to obtain π_R pertain to one domain—and hence, π_R can be called the "rational policy"—the rewards underlying π_C blend domain rewards with (intended) moral penalties, creating a domain that is different from both that of morality and that of rationality. Also, given the issue with thinking of the reward signal in terms of reasons/goodness in a respect—namely, that there is but one kind of (unspecified) reason/goodness in a respect—the prospects of distinguishing moral from others reasons in this domain are dim.

Though the aim of constraining an RL agent that maximizes domain reward signal is clear, in the setup of Noothigattu et al., there is no boundary between morality and rationality, and it is not clear how to apply moral categories here.

Rodriguez-Soto et al. In a series of papers [34–36], culminating with [35], Rodriguez-Soto et al. use *multi-objective* RL in the context of AI alignment. They work with multi-objective Markov decision processes (MOMDPs) that include a *vector* of reward functions where each function corresponds to a different "value". Notably, all of these values are meant to be "ethical", with one of them, the "value of achievement", representing the agent's "individual objective". The authors assume that these values form a "value system" or that they are totally ordered. They also assume that the value of achievement is never the top ele-

ment in the value ordering. Insofar as the values are totally ordered, they induce a lexicographic ordering on the rewards. So, an RL agent functioning in such an MOMDP learns to maximize the reward signal associated with the highest value, then with the second to highest value, and so on. Rodriguez-Soto et al. [35] prove some important results about this setup and design some insightful experiments, which are too complex to be discussed here.

We register that this kind of setup is much better-aligned with our analysis, although it does have some limitations. Policies that maximize a vector of reward functions representing ethical values are straightforward to make sense of in terms (deontic, evaluative, and fittingness) moral categories. What is more, the fact that there are now multiple functions goes some way toward resolving the issue with interpreting reward signals in terms of reasons/goodness in a respect: if there are multiple reward signals, we can talk about different moral reasons for actions/policies, or alternatively, different respects in which actions/policies can be good. What is unusual—and this is the first limitation—is that the relative importance of reasons/respects of goodness is uniform: this is due to the fact that values are totally ordered. (Recall the drowning child scenario, where the child's welfare was more important than keeping the promise. The received wisdom in philosophy has it that, in other situations, the relative importance of these moral factors might be reversed.) The second limitation concerns the value of achievement and its place in the value ordering. We think it is more natural to view achievement as a nonethical value. Recall that the vector of reward functions corresponding to ethical values can be seen as giving rise to a moral domain. Consequently, if achievement is the least important element, the agent is naturally seen as learning to select the most rational policy among the moral ones. But if achievement is not the least element, the agent is naturally seen as at least sometimes selecting the rational action at the expense of the moral one.

More could be said about the approach of Rodriguez-Soto et al., and there are other RL-based approaches to AI alignment that deserve careful analysis. However, such an analysis must be left for another day. Still, we hope that this brief section illustrates how metanormative theory can be used as a principled basis for comparing competing RL-based approaches to machine ethics and AI alignment.

7 Conclusion

In this paper, we set out to contribute to meeting the challenge of designing agents that avoid moral evil and foster moral good. To this end, we explained how metanormative theory makes sense of morality and normativity more generally. We introduced the core normative categories, along with some other ideas that are central to this area of philosophy. We also discussed the way these ideas apply to the standard RL architecture, as well as tried to illustrate how metanormative theory can be used as a criterion for assessing the relative merits of approaches to machine ethics and AI alignment. We are not suggesting that this criterion should be decisive. However, we do think that the approaches that can be made

sense of using notions from metanormative theory are, ceteris paribus, more promising than those that cannot.

References

1. Alcaraz, B., Knoks, A., Streit, D.: Estimating weights of reasons using metaheuristics: A hybrid approach to machine ethics. In: Proceedings of the Seventh AAAI/ACM Conference on AI, Ethics, and Society (AIES-2024). pp. 27–38. ACM Press (2024)
2. Anderson, M., Anderson, S.L.: The status of machine ethics: A report from the aaai symposium. Minds and Machines **17**(1), 1–10 (2007)
3. Anderson, M., Anderson, S.L. (eds.): Machine Ethics. Cambridge University Press (2011)
4. Baker, D.: The varieties of normativity. In: The Routledge Handbook of Metaethics. Routledge (2017)
5. Baker, D.: Skepticism about ought simpliciter. In: Shafer-Landau, R. (ed.) Oxford Studies in Metaethics 13. Oxford University Press (2018)
6. Bello, P., Malle, B.F.: Computational approaches to morality. In: Sun, R. (ed.) Cambridge Handbook of Computational Cognitive Sciences, pp. 1037–1063. Cambridge University Press (2023)
7. Berker, S.: Particular reasons. Ethics **118**(1), 109–139 (2007)
8. Berker, S.: The deontic, the evaluative, and the fitting. In: Rowland, R.A. (ed.) Fittingness: Essays in the Philosophy of Normativity. Oxford University Press (2022)
9. Biran, O., Cotton, C.: Explanation and justification in machine learning: A survey. In: IJCAI-17 Workshop on Explainable AI (XAI). vol. 8, pp. 8–13 (2017)
10. Brown, J.: On scepticism about ought simpliciter. Australasian Journal of Philosophy **102**(2), 497–511 (2024)
11. Canavotto, I., Horty, J.: Piecemeal knowledge acquisition for computational normative reasoning. In: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES-2022). pp. 171–80. ACM Press (2022)
12. Chan, A., Salganik, R., Markelius, A., Pang, C., Rajkumar, N., Krasheninnikov, D., Langosco, L., He, Z., Duan, Y., Carroll, M., Lin, M., Mayhew, A., Collins, K., Molamohammadi, M., Burden, J., Zhao, W., Rismani, S., Voudouris, K., Bhatt, U., Weller, A., Krueger, D., Maharaj, T.: Harms from increasingly agentic algorithmic systems. In: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency. p. 651–666. FAccT '23, Association for Computing Machinery, New York, NY, USA (2023)
13. Conitzer, V.: Why should we ever automate moral decision making? (2024), https://arxiv.org/abs/2407.07671
14. Conitzer, V., Freedman, R., Heitzig, J., Holliday, W.H., Jacobs, B.M., Lambert, N., Mossé, M., Pacuit, E., Russell, S., Schoelkopf, H., Tewolde, E., Zwicker, W.S.: Position: Social choice should guide AI alignment in dealing with diverse human feedback. In: Proceedings of the 41st International Conference on Machine Learning. ICML'24, JMLR.org (2024)
15. Dancy, J.: Moral Reasons. Blackwell (1993)
16. Eschmann, J.: Reward function design in reinforcement learning. In: Belousov, B., Abdulsamad, H., Klink, P., Parisi, S., Peters, J. (eds.) Reinforcement Learning Algorithms: Analysis and Applications, pp. 25–33. Springer (2021)
17. Faroldi, F.L.G.: Reasons-based artificial agents. AI Ethics **6**(77) (2026)

18. Floridi, L., Sanders, J.: On the Morality of Artificial Agents. Minds and Machines **14**(3), 349–379 (2004)
19. Gabriel, I.: Artificial Intelligence, values, and alignment. Minds and Machines **30**(3), 411–437 (2020)
20. Hidalgo, C.A., Orghian, D., Canals, J.A., de Almeida, F., Martin, N.: How Humans Judge Machines. The MIT Press (2021)
21. Howard, C., Cosker-Rowland, R.: Fittingness: A user's guide. In: Howard, C., Cosker-Rowland, R. (eds.) Fittingness. Oxford University Press (2022)
22. Howard, C.: Fittingness. Philosophy Compass **13**(11), e12542 (2018)
23. Ji, J., Qiu, T., Chen, B., Zhang, B., Lou, H., Wang, K., Duan, Y., He, Z., Vierling, L., Hong, D., Zhou, J., Zhang, Z., Zeng, F., Dai, J., Pan, X., Ng, K.Y., O'Gara, A., Xu, H., Tse, B., Fu, J., McAleer, S., Yang, Y., Wang, Y., Zhu, S.C., Guo, Y., Gao, W.: AI Alignment: A Comprehensive Survey (2025), http://arxiv.org/abs/2310.19852, arXiv:2310.19852
24. Kagan, S.: Normative Ethics. Westview Press (1998)
25. Kauppinen, A.: Getting things right: Fittingness, reasons, and value, by Conor McHugh and Jonathan Way The range of reasons in ethics and epistemology, by Daniel Whiting. Mind (2024)
26. Lord, E., Maguire, B.: An opinionated guide to the weight of reasons. In: Lord, E., Maguire, B. (eds.) Weighing Reasons, pp. 3–24. Oxford University Press (2016)
27. McHugh, C., Way, J.: Fittingness first. Ethics **126**(3), 575–606 (2016)
28. McHugh, C., Way, J.: Getting Things Right: Fittingness, Reasons, and Value. Oxford University Press (2023)
29. Montes, N., Osman, N., Sierra, C., Slavkovik, M.: Value engineering for autonomous agents (2023), https://arxiv.org/abs/2302.08759
30. Ng, A.Y., Russell, S.J.: Algorithms for inverse reinforcement learning. In: Proceedings of the Seventeenth International Conference on Machine Learning. p. 663–670. ICML '00, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2000)
31. Noothigattu, R., Bouneffouf, D., Mattei, N., Chandra, R., Madan, P., Varshney, K.R., Campbell, M., Singh, M., Rossi, F.: Teaching AI agents ethical values using reinforcement learning and policy orchestration. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19. pp. 6377–6381 (2019)
32. Parfit, D.: On What Matters (Volume I). Oxford University Press (2011)
33. Raz, J.: Practical reason and norms. Oxford University Press (1990)
34. Rodriguez-Soto, M., Lopez-Sanchez, M., Rodriguez-Aguilar, J.A.: Multi-objective reinforcement learning for designing ethical multi-agent environments. Neural Computing and Applications **37**, 25619–25644 (2023)
35. Rodriguez-Soto, M., Rădulescu, R., Bistaffa, F., Ricart, O., Mayoral, A., Lopez-Sanchez, M., Rodriguez-Aguilar, J.A., Nowé, A.: Multi-objective reinforcement learning for provably incentivising alignment with value systems. Artificial Intelligence (2026)
36. Rodriguez-Soto, M., Serramia, M., Lopez-Sanchez, M., Rodriguez-Aguilar, J.A.: Instilling moral value alignment by means of multi-objective reinforcement learning. Ethics and Information Technology **24**(9) (2022)
37. Russell, S., Norvig, P.: Artificial Intelligence: A Modern Approach. Pearson Education, 4 edn. (2020)
38. Sayre-McCord, G.: Metaethics. In: Zalta, E.N., Nodelman, U. (eds.) The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University, Spring 2023 edn. (2023)

39. Scanlon, T.M.: What We Owe to Each Other. Cambridge, MA: Harvard University Press (1998)
40. Schroeder, M.: Reasons First. Oxford University Press (2021)
41. Shafer-Landau, R.: The Fundamentals of Ethics. Oxford University Press (2012)
42. Singer, P.: Famine, affluence, and morality. Philosophy and Public Affairs **1**(3), 229–43 (1972)
43. Sutton, R.S., Barto, A.G.: Reinforcement Learning: An Introduction. A Bradford Book, Cambridge, MA, USA (2018)
44. Tucker, C.: Weighing reasons. In: Zalta, E.N., Nodelman, U. (eds.) The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University, Winter 2025 edn. (2025)
45. Väyrynen, P.: Normative explanation and justification. Noûs **55**(1), 3–22 (2021)
46. Väyrynen, P.: Varieties of normative explanation. In: Copp, D., Rosati, C. (eds.) The Oxford Handbook of Metaethics. Oxford University Press (forthcoming)
47. Vishwanath, A., Dennis, L.A., Slavkovik, M.: Reinforcement learning and machine ethics: A systematic review (2024), https://arxiv.org/abs/2407.02425
48. Wallach, W., Allen, C.: Moral Machines: Teaching Robots Right from Wrong. Oxford University Press, Inc., USA (2008)
49. Wallach, W., Allen, C., Smit, I.: Machine morality: bottom-up and top-down approaches for modelling human moral faculties. AI & SOCIETY **22**(4), 565–582 (2008)
50. Whiting, D.: The Range of Reasons: In Ethics and Epistemology. Oxford University Press (2022)
51. Zhong, T., Song, Y., Limarga, R., Pagnucco, M.: Computational machine ethics: A survey. Journal of Artificial Intelligence Research **82** (2025)