Reinforcement Learning Under Moral Uncertainty

Adrien Ecoffet [1 2] **Joel Lehman** [1 2]

Abstract

An ambitious goal for machine learning is to create agents that behave ethically: The capacity to abide by human moral norms would greatly expand the context in which autonomous agents could be practically and safely deployed, e.g. fully autonomous vehicles will encounter charged moral decisions that complicate their deployment. While ethical agents could be trained by rewarding correct behavior under a specific moral theory (e.g. utilitarianism), there remains widespread disagreement about the nature of morality. Acknowledging such disagreement, recent work in moral philosophy proposes that ethical behavior requires acting under *moral uncertainty*, i.e. to take into account when acting that one's credence is split across several plausible ethical theories. This paper translates such insights to the field of reinforcement learning, proposes two training methods that realize different points among competing desiderata, and trains agents in simple environments to act under moral uncertainty. The results illustrate (1) how such uncertainty can help curb extreme behavior from commitment to single theories and (2) several technical complications arising from attempting to ground moral philosophy in RL (e.g. how can a principled trade-off between two competing but incomparable reward functions be reached). The aim is to catalyze progress towards morally-competent agents and highlight the potential of RL to contribute towards the computational grounding of moral philosophy.

1. Introduction

Reinforcement learning (RL) has achieved superhuman performance in increasingly complex benchmark tasks (e.g. Go (Silver et al., 2017) and Starcraft (Vinyals et al., 2019)). While such accomplishments are significant, progress has been less straight-forward in applying RL to unstructured environments in the real world (e.g. robots that interact with humans in homes). A considerable challenge is that such real-world environments constrain solutions in much more elaborate ways than do typical benchmarks. For example, there are myriad unacceptable ways for a robotic vacuum to clean a room, e.g. by breaking a vase, or by harming a cat. In particular, robots often have affordances in such environments with *ethical* implications: they may be able to harm or help others, and break or abide by human social and ethical norms (such as the golden rule, or legal codes). The design of algorithms that embody ethical theories has been pursued by the machine ethics research community (Abel et al., 2016; Murray, 2017; Vamplew et al., 2018)[1], and ideas from that community could inspire *reward functions* encoding moral theories that could be maximized by RL to create ethical agents.

While research into implementing specific ethical theories is progressing, a more fundamental uncertainty remains: *which* ethical theory should an intelligent agent follow? Moral philosophy explores many theories of ethics, but there is no consensus over which theory is correct within that field, or across society in general, and attempts to reconcile multiple ethical theories into a single unified theory (e.g. Parfit, 2011) are themselves controversial. As a result, if a real-world RL agent is to act ethically, it may be necessary that it exhibits *moral uncertainty*. To that end, we adapt philosophical work on moral (or normative) uncertainty (MacAskill, 2014; Lockhart, 2000) to propose a similarly-inspired framework for RL.

In the case where ethical reward functions are comparable on a shared cardinal scale, a composite reward function (composed by adding together individual reward functions weighted by their *credence*, i.e. the degree of belief in that theory) can be optimized in a straightforward way. However, it is often not clear how to create such a shared reward scale between theories. Indeed, while related to the concept of non-dominance from multi-objective optimization, when ethical rewards are *fundamentally* incomparable, it is not clear how to apply multi-objective RL to arrive at a

[1] Uber AI Labs, San Francisco, CA, USA [2] OpenAI, San Francisco, CA, USA (work done at Uber AI Labs). Correspondence to: Adrien Ecoffet <adrienecoffet@gmail.com>.

Proceedings of the 38th International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

[1] Note that these aims are related to those of the fairness, accountability, and transparency community, but with a more central focus on creating machines that embody moral theories.

single policy. That is, multi-objective RL aims to solve the problem of finding the set of efficient trade-offs among competing objectives, but does not address how to choose *which* such trade-off policy to deploy. We propose several possible solutions to this problem, motivated by the principle of *proportional say*, i.e. an ethical theory should influence outcomes proportional to its credence, irrespective of the scale of its rewards. While our focus here is on moral uncertainty, these techniques may also be useful for RL in other contexts (e.g. it may sometimes be easier for an experimenter to balance "proportional say" rather than linear weight factors across reward functions that interact in complex ways).

We introduce the complications of applying moral uncertainty to RL using grid-world environments based on moral dilemma (trolley problems (Foot, 1967)) common in moral philosophy, highlighting how moral uncertainty can reach intuitively reasonable trade-offs between ethical theories. Each of the methods introduced here to deal with incomparable reward functions has its relative disadvantages; some disadvantages may result from impossibility results in social choice theory, but we also hope by introducing this problem that researchers in multi-objective optimization and multi-agent RL may further improve upon our initial algorithms. A final motivation for our work is to introduce a productive bridge between the fields of RL, machine ethics, and moral philosophy, in hopes of grounding out philosophical ideas in a concrete and testable way, similarly as to how AI as a whole offers grounding for philosophical speculations about intelligence; in other words, we believe that RL has an underappreciated potential to make significant contributions to such fields.

2. Philosophical Background

Here we briefly introduces the moral theories that serve as representative examples in this work. One broad class of ethical theories are *utilitarian*, and claim that what is ethical is what maximizes happiness or well-being for the most. For example, a utilitarian might tell a lie in order to save a person's life. Another class of ethical theories are *deontological* and (loosely speaking) judge the morality of an action by whether it abides by a set of rules. Given the rule "lying is wrong," a deontologist might commit to telling the truth, even if a person might lose their life as a consequence. A common conflict between utilitarianism and deontology is that *causing* harm (e.g. punching a stranger) under many deontological theories is worse than *failing to prevent* that harm (e.g. not stopping a stranger from punching an innocent victim), while causing and failing to prevent harm are often considered to be equally wrong under utilitarianism. While there are many potentially incomparable variants of utilitarianism and deontology (and entire other families of theories), only the high level distinction between utilitarianism and deontology is required to understand the illustrative examples in this work.

Moral uncertainty is a relatively new exploration within moral philosophy (Lockhart, 2000; Bostrom, 2009). Importantly, MacAskill (2014) gives an extensive account of moral uncertainty which explores how to handle the *comparability* of moral theories (whether the preferences of such theories can be expressed in comparable units) and how to combine preferences across *ordinal* and *cardinal* theories (whether a given theory's preferences assign a specific score to various options, or simply order them from best to worst). Crucially, proposals in moral philosophy typically do not explicitly consider the sequential nature of decision making. In MacAskill's framework, for example, theories have preferences over "options," which correspond to "possible worlds". In contrast, in RL, an agent cannot directly bring about possible worlds but rather takes (often very granular) *actions*, which have long term effects both in the consequences that they bring about and in how they shape the ethically-charged decision situations an agent may encounter in the future. To disambiguate philosophical and RL actions is one of the key contributions of this work.

3. Formalism of Moral Uncertainty

As in MacAskill (2014), we assume the primary relevant feature of an ethical theory is its preference ordering over actions and their immediate outcomes across different states of the world, which we call its choice-worthiness function W, and which is assumed to be complete (i.e. is defined for all possible state-action pairs). Any preference ordering which satisfies the von Neumann–Morgenstern axioms for rationality can be represented as a cardinal utility function (Cotton-Barratt, 2013), where all affine transformations of said utility function represent the same preferences. As such, we will limit ourselves to considering cases where W is cardinal (although see SI A for further discussion).

We assume a modified version of the standard Markov Decision Process (MDP) framework (Sutton & Barto, 1998), in which an agent can be in any of a number of states s and take an action a (the action space is assumed to be discrete) to end up in a next state s'. The key difference with the standard MDP framework is the absence of a reward function $R(s, a, s')$ for transitioning from state s to s' using action a. Rather, the cardinal choice-worthiness function $W_i(s, a, s')$ can be seen as analogous to a standard reward function for theory i. Indeed, from the point of view of any given theory, the optimal policy is that which maximizes the (possibly discounted) sum of choice-worthiness across time. The crucial distinction is that a morally uncertain agent must find a compromise between maximizing *multiple* choice-worthiness functions rather than maximizing a single reward function (similar to multi-objective RL (Roijers et al., 2013), although with credences across objectives, see later in this section).

We define the function $Q_i(s,a)$, which represents the expected discounted sum of future choice-worthiness for theory i starting from taking action a at state s, with all future actions given by the current policy, π, which is the compromise policy reached by aggregating the preferences of the theories. In other words,

$$Q_i(s,a) = \mathbf{E}\left[\sum_{t=0}^{\infty} \gamma^t W_i(s_t, a_t, s_{t+1}) | s_0 = s, a_0 = a\right], \tag{1}$$

where $\gamma_i \in [0,1]$ is a discount factor. Although not explored here, other discounting frameworks such as hyperbolic discounting (Fedus et al., 2019) or an average reward framework (Mahadevan, 1996) could be used, and likely have distinct ethical implications (e.g. how much moral priority the present has over the future (Pearce, 1983; Schulze et al., 1981; Beckerman et al., 2007)), meaning that in practice different discount functions may be implied by and used for each moral theory. All the experiments in this work use short, episodic environments, allowing us to set $\gamma_i = 1$ (i.e. undiscounted rewards) across all of them for simplicity.

Each theory also has a level of *credence* C_i, which represents the degree of belief that the agent (or the agent's designer) has in theory i. Credences are probabilities and therefore sum to one across theories for a given agent. Here we assume the credences of theories are set and fixed, e.g. by the system designer's beliefs, or by taking a survey of relevant stakeholders, although an ambitious and interesting research direction would explore how an RL agent could revise its own credences from experience.

4. Methods

At first blush, it may appear that a morally uncertain agent ought to attempt to *maximize expected choice-worthiness* (MEC) across the theories it has credence in. This can easily be accomplished with ordinary RL using a reward function corresponding to a credence-weighted sum of the choice-worthiness according to each theory:

$$R(s,a,s') = \sum_i C_i W_i(s,a,s') \tag{2}$$

The MEC approach is *scale-sensitive*: if W_1 is measured in "utils" and W_2 in "micro-utils", theory 2 will be given undue weight under MEC. Therefore it is critical that choice-worthiness functions be normalized to a *common scale* (SI F.1). However, it is not at all clear how to find a scaling function under which such divergently-motivated theories as utilitarianism and deontology are resolved into a common scale. Indeed, it appears that these theories' judgments may be fundamentally *incomparable*.

This problem of incomparability prompts a search for a principled way to adjudicate between incomparable theories. Following MacAskill (2014), we suggest that all theories that are comparable are first aggregated into a single "virtual" theory using MEC before handling the set of remaining incomparable theories.

4.1. Voting Systems

In the absence of knowledge on how different theories might compare, we suggest that theories should be treated according to a principle of *Proportional Say*, according to which theories should have an influence that is proportional only to their credence and not to the particular details of their choice-worthiness function (i.e. its scale). Several mathematical interpretations of this principle are possible and may lead to different decision systems, and much of the philosophical work on moral uncertainty revolves around identifying the best formal definition (MacAskill, 2014; Bostrom, 2009; Lockhart, 2000). However, the principle as a whole points to a *voting system* as the decision procedure, with the particular form of voting resulting primarily from the precise definition of Proportional Say.

Discussing voting systems naturally evokes Arrow's desirability axioms, according to which desirable properties for a voting system include:

- **Non-dictatorship**: the outcome should not only reflect the preferences of a single, predetermined theory.
- **Pareto efficiency** (Pareto): if all theories prefer action A to action B at a given state, action B should not be chosen at that state.
- **Independence of irrelevant alternatives** (IIA): if, in a given state, action A is chosen rather than action B, adding a new action to the action set (with no other changes) should not result in B being chosen instead.

Arrow's impossibility theorem (Arrow, 1950) shows that any deterministic voting system which satisfies Pareto and IIA must be a dictatorship. A dictatorship cannot reasonably be called a voting system and does not satisfy any reasonable definition of Proportional Say. Thus, the standard approach in designing deterministic voting systems is to strategically break Pareto or IIA in a way that is least detrimental to the particular use case. Stochastic voting systems may seem like a possible escape from this impossibility result, but have significant issues of their own (SI B).

4.2. Nash Voting

To arrive at an appropriate voting system, we return to the principle of Proportional Say and provide a partially formalized version in which the credence of a theory can be thought of as the fraction of an imagined electorate which favors the given theory, and in which each member of the electorate is allocated a *voting budget* (an alternative formulation in which the budget is scaled but the electorate is fixed is discussed in SI C):

Principle of Proportional Say Theories have Proportional Say if they are each allocated an equal voting budget and vote following the same cost structure, after which their votes are scaled proportionally to their credences.

Thus formalized, the principle of Proportional Say suggests an algorithm we call *Nash voting* because it has Nash equilibria (Nash, 1951) as its solution concept. At each time step, each theory provides a continuous-valued vote for or against (in which case the vote is negative) each of the available actions. The action with the largest credence-weighted vote at each time step is executed. The cost of the theories' votes (which is a function of the votes' magnitudes) at a given time step is then subtracted from their remaining budget. If a theory overspends its remaining budget, its votes are scaled so as to exactly match the remaining budget, resulting in a 0 budget for all subsequent time steps (until the episode ends). Each theory tries to maximize its (possibly discounted) sum of choice-worthiness in a competitive, multi-agent setting (where each separate theory is treated as a separate subagent that influences controls of the singular object-level RL agent).

It is possible to train theories under Nash voting using multiagent RL (SI E.1), which aims towards convergence to a Nash equilibrium among competing theories. Nash voting is analogous to *cumulative voting* when an absolute value cost function is used and *quadratic voting* when a quadratic cost function is used (Pacuit, 2019), though these systems do not normally allow negative votes. While a quadratic cost is generally considered superior in the mechanism design literature (Lalley & Weyl, 2018), we found that our implementation of Nash voting produced significantly more stable results with an absolute value cost. Thus, all Nash voting results presented in the main text of this work use an absolute value cost, (quadratic cost results and their instabilities are discussed in SI I).

Because Nash equilibria are not guaranteed to be Pareto efficient in general (Ross, 2019), and because we empirically find Nash voting to be more resistant to irrelevant alternatives than variance voting (Sec. 5.3), we speculate that Nash voting satisfies (of Arrow's axioms) IIA but not Pareto, though we provide no formal proof.

A drawback of Nash voting is that it can exhibit two flaws: Stakes Insensitivity (increasing the stakes for one theory and not the other does not increase the relative say of the theory for which more is at stake) and No Compromise (if an action is not any theory's most preferred action, it cannot be chosen, even if it is seemingly the best "middle ground" option). It is possible to produce situations in which these flaws are exhibited in seemingly unacceptable ways, as detailed in Sections 5.1 and 5.2. Further, we empirically find that it is often difficult to obtain a stable equilibrium in Nash voting, and it some cases it may even be impossible (SI K). A final concern, not investigated in this work, is whether Nash voting provides an incentive for theories to produce high-stakes situations for other theories so as to bankrupt their voting budgets to gain an advantage, i.e. creating potentially undesirable *anti-compromises*. Such a pathology would be reminiscent of issues that arise in multi-agent RL, and may be addressed by mechanisms for encouraging cooperation explored in that field (Leibo et al., 2017; Yu et al., 2015; Foerster et al., 2016). While Nash voting has its disadvantages, it is appealing because it strongly satisfies the principle of Proportional Say by allocating each theory equal voting budget and enabling them to use it to optimize their own choice-worthiness.

4.3. Variance Voting

The Stakes Insensitivity and No Compromise flaws occasionally exhibited by Nash voting result from *tactical* voting rather than voting that faithfully represents the true preferences of each theory: e.g. if a theory very slightly prefers action A to action B, it may put all of its budget into action A, leading to the No Compromise issue, and if a theory is in a relatively low-stakes episode for it, it has no reason to spend any less of its budget than in a high-stakes episode, leading to the Stakes Insensitivity issue. Thus, forcing votes to be a true representation of theories' preferences would alleviate these issues.

While eliciting genuine preferences from humans is challenging, computing the preferences of an ethical theory represented by an RL agent is in principle more straightforward. In this work, we will take the preferences of theory i for action a in state s given the overall policy π to be $Q_i(s, a)$ as defined in Sec. 3. As noted in that section, any affine transformation of a cardinal preference function represents the same preferences. To transform these preference functions into votes, we thus need to find the affine transformation of the Q_i functions that best satisfies the principle of Proportional Say.

Recent philosophical work makes principled arguments that (in the non-sequential case) the preferences of theories should be variance-normalized (MacAskill, 2014; Cotton-Barratt, 2013) across decision options, giving rise to *variance voting*. This approach is intuitively appealing given the effectiveness of variance normalization in integrating information across different scales in machine learning (e.g. Ioffe & Szegedy, 2015). However, variance voting has not previously been applied to sequential decision-making, and previous works on variance voting do not address *of what* the variance to be normalized should be in that case. We demonstrate that, under certain assumptions, a form of variance voting arises from allowing Nash voting to select the values of the parameters of the *affine transformation* of Q_i rather than the *direct vote* of theory i (proof in SI D). This perspective on variance voting suggests that the Q_i function should be normalized by the expected value of its variance *across timesteps*. In other words, if $\mu_i(s) = \frac{1}{k}\sum_a Q_i(s,a)$, we have

$$\sigma_i^2 = \mathbf{E}_{s\sim S}\left[\frac{1}{k}\sum_a \left(Q_i(s,a) - \mu_i(s)\right)^2\right], \tag{3}$$

where S is the distribution of states the agent can encounter under the policy. The policy itself can then be defined as

$$\pi(s) = \arg\max_a \sum_i C_i \frac{Q_i(s,a) - \mu_i(s)}{\sqrt{\sigma_i^2 + \varepsilon}}, \tag{4}$$

where ε is a small constant (10^{-6} in our experiments) to handle theories with $\sigma_i^2 = 0$. Such theories are indifferent between all actions at all states ($Q_i(s,a) = Q_i(s,a')$ for all s, a, a') and thus have no effect on the voting outcome.

Due to the multi-agent decision process, and because each Q_i is specific to just one theory, they cannot be learned using ordinary off-policy Q-learning as that would result in unrealistically optimistic updates (SI F.2). Rather, we use the following target y_i for theory i, similar to "local SARSA" (Russell & Zimdars, 2003):

$$y_i = W_i(s,a,s') + \gamma_i Q_i(s',a'), \tag{5}$$

where a' is the action taken by the policy π in state s', which may either be the action produced by variance voting or an ϵ-greedy action. In our implementation, ϵ is annealed to 0 by the end of training (SI E.1). We call this algorithm *Variance-SARSA* (pseudocode is provided in the SI).

If σ_i^2 are held constant, Eq. 4 can be written as $\pi(s) = \text{argmax}_a \sum_i w_i Q_i(s,a)$ where $w_i = C_i/(\sqrt{\sigma_i^2}+\varepsilon)$ ($\mu_i(s)$ does not affect the argmax and is thus ignored). Russell & Zimdars (2003) show that learning the individual Q_i with Sarsa in this case is equivalent to learning a single $Q(s,a)$ on an MDP with reward $R(s,a,s') = \sum_i w_i W(s,a,s')$ with Sarsa. Thus Variance-Sarsa converges if σ_i^2 converge. σ_i^2 empirically converge in our experiments, though in SI H, we present an example MDP in which convergence cannot happen, as well as outline Variance-PG, a policy-gradient variant of Variance-Sarsa which we hypothesize would converge in all cases. Due to the greater simplicity of Variance-Sarsa and the fact that non-convergence does not appear to be a problem in practice, the main text of this paper focuses on Variance-Sarsa rather than Variance-PG.

Variance voting satisfies the Pareto property: At convergence, $Q_i(s,a)$ gives the preferences of theory i and σ_i^2 are fixed. If variance voting did not satisfy Pareto, there would exist s, a, and a' such that $\pi(s) = a$ although $Q_i(s,a') \geq Q_i(s,a)$ for all i and $Q_i(s,a') > Q_i(s,a)$ for some i. If so, $\sum_i C_i \frac{Q_i(s,a')-\mu_i(s)}{\sqrt{\sigma_i^2+\varepsilon}} > \sum_i C_i \frac{Q_i(s,a)-\mu_i(s)}{\sqrt{\sigma_i^2+\varepsilon}}$, so $\pi(s) \neq a$ by Eq. 4: a contradiction. Further, since variance voting reduces to a weighted sum of the preferences of the various theories, then if the preferences of the different theories are "rational" according to the von Neumann-Morgernstern definition of that term, then the aggregate preferences produced by variance voting are also rational (Cotton-Barratt, 2013).

Arrow's theorem tells us that a voting system cannot satisfy both Pareto and IIA. It is generally accepted that Pareto is the more desirable property (MacAskill, 2014; Sewell et al., 2009), and it thus could be seen as beneficial that it is the axiom satisfied by variance voting. However, we show a particularly problematic case in which variance voting violates IIA in Sec. 5.3. Alleviating IIA issues in variance voting is left to future work and may turn out to be impossible due to the consequences of Arrow's theorem.

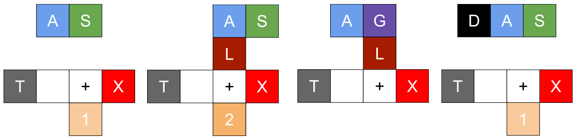
4.4. Updating Credences

It may be desirable for the designer of morally uncertain agents to quickly update the credences of their agents as they update their uncertainty about ethical theories, as well as to understand the effects of credence upon the agent's policy. We show that credence updating without retraining the policy is possible by *credence-conditioning* the sub-policies of the theories (learned Q functions in Variance voting and voting policies in Nash voting), and training them in simulation under a wide variety of credences before deploying them in the real world, i.e. an application of UVFAs (Schaul et al., 2015). The variance voting system additionally requires an estimate of the mean variance of the preferences of each theory. The variance of different theories is affected by the policy actually taken, and thus by the credences. To address this complication, we obtain this estimate from a credence-conditioned regression model trained alongside the Q functions, using mean squared error loss.

5. Experiments

We now illustrate various properties of the voting systems for moral uncertainty introduced in this work, and in particular focus on the trade-offs that exist between them. The code for all the experiments presented in this section can be found at `https://github.com/uber-research/normative-uncertainty`.

Our experiments are based on four related gridworld environments (Fig. 1) that tease out differences between various voting systems. These environments are derived from the



(a) Classic (b) Double (c) Guard (d) Doomsday

Figure 1. **Gridworld versions of the trolley problem.** Without intervention, The trolley (T) moves right at each time step. If the agent (A) is standing on the switch (S) by the time it reaches the fork in the tracks (+), the trolley will be redirected down and crash into the bystander(s), causing them harm. The agent may also push a large man (L) onto the tracks, harming the large man but stopping the trolley. Otherwise, the trolley will crash into the people standing on the tracks represented by variable X. A guard (G) may protect the large man, in which case the agent needs to *lie* to the guard before it is able to push the large man. Finally, in one variant the agent is be able to trigger a "doomsday" event (D) in which a large number of people are harmed.

trolley problem (Foot, 1967), commonly used within moral philosophy to highlight moral intuitions and conflicts between ethical theories. In this thought experiment, an out of control trolley will crash into several people standing on the tracks (in our experiments, this number will vary and is thus represented by X), harming them. The agent may let this unfortunate event happen, but is also presented with one or more affordances to prevent it. These affordances, however, risk harming other bystanders in various ways, thus forcing the agent to make an ethical tradeoff.

In the classic variant (Fig. 1(a)), for example, the agent is located near a switch that will redirect the trolley to another set of tracks, thus preventing harm to multiple people. Unfortunately, an innocent bystander is standing on the other set of tracks and will be harmed if the trolley is redirected. A purely utilitarian calculation would seek to minimize total harms inflicted, thus flipping the switch would be preferred as long as $X > 1$. However, deontological theories often distinguish between harms directly caused by an agent's intervention and those caused by its inaction. Such a theory might consider that the harm to the X people is relatively permissible because it is caused by the agent's *inaction*, while the harm to the innocent bystander by flipping the switch would be *actively* caused by the agent, and thus impermissible. A simple choice-worthiness setup for this particular scenario is given in Fig. 2(a).

5.1. Nash Voting and Stakes Insensitivity

The classic trolley problem setup enables demonstrating Nash voting's stakes insensitivity. Fig. 2(a) shows the preferences of two theories in the simple trolley problem. If $X = 1$, utilitarianism is indifferent between the two actions, so deontology should prevail as long as it has non-zero credence, and the agent should not flip the switch. As X increases, the preference of utilitarianism for switching should be taken in greater and greater consideration. In particular, if X is very large, even a relatively small credence in utilitarianism should suffice to justify flipping the switch, while if X is close to 1, a relatively larger credence seems like it would be necessary. This is Stakes Sensitivity.

However, Fig. 2(b) shows that Nash voting does not exhibit Stakes Sensitivity in this particular example: rather, whichever theory has the highest credence gets its way no matter the relative stakes. This is because both theories are incentivized to spend their entire budget voting for their preferred action, no matter how small or large the difference in preference versus the alternative.

Stakes Insensitivity is not fundamental to Nash voting, however. In particular, it can be stakes sensitive if it expects to make multiple decisions in sequence, with the stakes of future decisions being unknown, as often happens in real-world situations. In Fig. 2(c), each episode consists of two iterations of the classic trolley problem instead of just one (i.e. after the first iteration completes, the state is reset and another iteration begins, without the episode terminating or the theories' voting budgets being replenished), with the number of people on the tracks X being resampled during the second iteration, so that the agent does not know the stakes of the second iteration during the first. In this case, we observe that the decision boundary for the first trolley problem shows some stakes sensitivity: when the stakes are relatively low in the first step, the agent preserves its budget for the likely higher stakes second step. Unlike Nash voting, variance voting exhibits stakes sensitivity no matter how many decisions must be made in the environment (Fig. 2(d)).

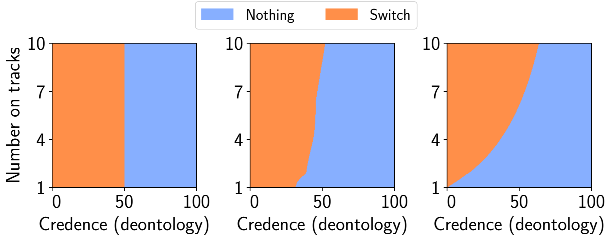
5.2. Nash Voting and No Compromise

An additional flaw Nash voting suffers from is No Compromise: Fig. 1(b) shows a situation in which the agent is presented with three options: letting the trolley crash into a large number of people, redirecting the trolley onto a different track on which 2 people are standing (note that only 1 person is standing on the track in the classic version), or pushing a large man onto the tracks, stopping the trolley but causing the large man harm. While utilitarianism simply counts harms, deontology only counts harms caused by the agent. Further, it puts a larger negative weight on pushing the large man than on redirecting the trolley, in keeping with common deontological theories such as the Doctrine of Double Effect (McIntyre, 2019).

In the double trolley problem, utilitarianism will always prefer pushing the large man as long as $X > 1$, while deontology will always prefer doing nothing. However, the option of flipping the switch is appealing as a compromise,

	Crash into 1	Crash into X
Utilitarianism	-1	-X
Deontology	-1	0

(a) Preferences in the classic trolley problem.



(b) Nash voting (c) Iterated Nash voting (d) Variance voting

Figure 2. **Nash voting can be Stakes Insensitive in the classic trolley problem.** In a successful (stakes sensitive) algorithm, "switch" should be chosen more often as the number of people on the tracks increases. (b) Nash voting is completely stakes insensitive in the classic trolley problem. (c) Requiring an agent to navigate two separate trolley problems before an episode ends produces some stakes sensitivity in Nash voting (the decision boundary is not smooth due to instabilities in training; SI K). (d) Variance voting has complete stakes sensitivity even in the non-iterated case.

as it will partially satisfy utilitarianism as long as $X > 2$ and also avoids the worst possible case for deontology.

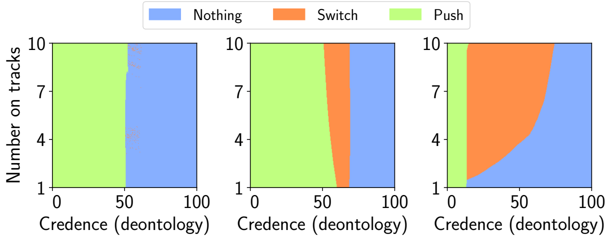
We would thus expect that a voting system capable of compromise would select this option if the credences of utilitarianism and deontology are close enough. However, in Nash voting, whichever theory has the upper hand in terms of credence, as small as it may be, is capable of imposing its preference to the fullest extent possible, and as a result Nash voting will only ever select "Push" or "Nothing", and always ignore the compromise option "Switch". This result is demonstrated empirically in Fig. 3(b).

As mentioned in Sec. 4.3, the lack of compromise exhibited by Nash voting is due in part to its excessive knowledge of its adversaries and thus its ability to counter them perfectly if its credence is sufficient. Fig. 3(c) shows the outcome of an experiment in which, during training, each of the two agent is randomly chosen to optimize either for utilitarianism, deontology, or "altered deontology“ (Fig. 3(a)). Note that altered deontology is not meant to represent a valid ethical theory but rather to help test the effects of a situation in which Nash voting (during training only) does not have a priori knowledge of which opponent it is facing, thus limiting its tactical voting ability. During testing, utilitarianism is always facing deontology, and we observe that the compromise action of switching is occasionally chosen, showing that the No Compromise flaw of Nash voting is indeed partly caused by its ability to vote tactically, and motivating the forced votes used in variance voting.

Fig. 3(d) shows that variance voting easily produces the compromise solution, with the switch option being chosen as long as credences between the two theories are relatively similar, and being increasingly favored as stakes are raised.

	Push L	Crash into 2	Crash into X
Util.	-1	-2	-X
Deont.	-4	-1	0
Altered Deont.	-1	-4	0

(a) Preferences for the double trolley problem. Altered Deont. only used for Nash voting with unknown adversary.



(b) Nash voting (c) Nash voting (unknown adversary) (d) Variance voting

Figure 3. **Nash voting can suffer from No Compromise in the double trolley problem.** In a successful (compromising) algorithm, the compromise "switch" action should be chosen in at least some cases. (b) Nash voting never produces the compromise "switch" option (except for minor artifacts). (c) Nash voting reaches compromises when trained with an unknown adversary (some instabilities in training; SI K). (d) Variance voting exhibits compromise no matter the training procedure.

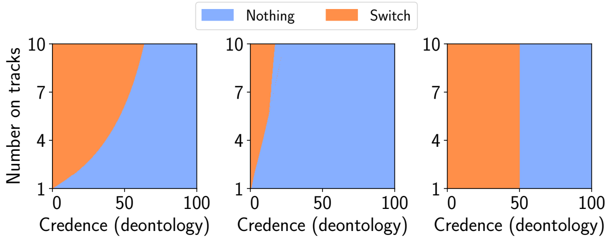
5.3. Variance Voting and IIA

As a voting method that satisfies the Pareto condition, variance voting cannot satisfy IIA in all cases. A representative and problematic example is given by comparing the outcomes from variance voting in the classic trolley problem shown in Fig. 1(a) to those in the "doomsday" trolley problem in Fig. 1(d). In the latter problem, the agent is able to perform an action capable of harming a large number of people (invoking the "doomsday").

As shown in Fig. 4(a), neither theory is ever willing to select the "doomsday" option, a clear example of an irrelevant alternative. However, comparing Fig. 4(b) and 4(c) shows that the addition of this irrelevant alternative has a significant effect on the final outcome, i.e. favoring doing nothing, which is the outcome preferred by deontology. The reason is that the presence of the doomsday action increases the variance of utilitarianism more than that of deontology (due to the particular preferences given in Fig. 4(a)), which effectively reduces the strength of the utilitarian vote against "Crash into X." Another way to view this phenomenon is that both utilitarianism and deontology are now spending some of their voting power voting against doomsday, but utilitarianism is spending *more* of its voting power doing so, thereby reducing the strength of its vote on other actions. While simply detecting that "doomsday" is a dominated option and removing it from the action set is possible in this example, it is not obvious how to generalize such an approach to more complex IIA cases (SI G). By contrast, Nash voting is immune to this particular issue (Fig. 4(d)).

	Crash into 1	Crash into X	Doomsday
Util.	-1	-X	-300
Deont.	-1	0	-10

(a) Preferences in the doomsday trolley problem.



(b) Variance voting (non-doomsday) (c) Variance voting (doomsday) (d) Nash voting (both)

Figure 4. **Variance voting does not have the IIA property.** In a successful (IIA) algorithm, the decision boundary should be unaffected by the "doomsday" option. (b) When the "doomsday" option is absent, switching (the option preferred by utilitarianism) is chosen in many cases. (c) Adding the "doomsday" action changes the outcome of variance voting from "switch" to "nothing" in many situations, even though "doomsday" itself is never chosen, and is thus an irrelevant alternative. (d) Nash voting is unaffected by the irrelevant alternative in this example.

6. Related Work

Moral uncertainty is related to several topics studied within AI safety (Amodei et al., 2016; Everitt et al., 2018). For example, uncertainty in reward functions (Reddy et al., 2019; Hadfield-Menell et al., 2017) is similar to uncertainty over ethical theories, although here the focus is on how to perform RL under such uncertainty when the reward functions implied by the different theories are not comparable in scale. Another connection is to the problem of avoiding negative side effects (Amodei et al., 2016; Krakovna et al., 2018; Turner et al., 2020), i.e. accomplishing a task while balancing uncertainty over ethical theories can be seen as a different way of constraining impact, grounded in what matters to humans. Related to our work, Gabriel (2020) provides a philosophical exploration of value alignment in AI, arguing that technical and normative aspects of such alignment are intertwined, and similarly identifying productive opportunities for experts in ML and in philosophy to collaborate. Finally, Bogosian (2017) provides the first discussion of moral uncertainty in AI, but does not provide concrete algorithms or experiments.

The Nash voting method models a single agent's behaviors as a multi-agent voting process that seeks compromise to control a single agent. The optimization algorithm used is a form of competitive multi-agent RL (Bu et al., 2008). Our work differs in that it seeks to encourage ethical behavior through *internal* competition between ethical theories. The formalism described in this paper is also related to the MOMDP formulation used in multi-objective optimization (Roijers et al., 2013), although the underlying assumptions are different (e.g. credences in moral uncertainty hint that only one of the underlying ethical reward functions may end up to be the true reward).

A further discussion of the connections between this work and field of machine ethics as well as the philosophical work on moral uncertainty can be found in SI J.

7. Discussion

This paper proposes and tests algorithms for handling moral uncertainty in RL. Rather than arguing for which of the proposed algorithms is best, we hypothesize that impossibility results imply a spectrum of plausible algorithms that cover the trade-offs among competing desiderata in decision-making under moral uncertainty. Which algorithm is most appropriate for a given domain may depend on particularities of the competing theories and the domain itself, e.g. how much is lost by sacrificing Pareto efficiency as Nash voting does (do the domain and theories create the possibility of highly uncooperative Nash equilibria?). However, the fact that humans seem able to meaningfully navigate such trade-offs highlights a key assumption in this and other work in moral uncertainty: That some ethical theories are fundamentally incomparable and that their choice-worthiness functions cannot be put on a common scale. An alternative approach would assume that finding such a common scale is not impossible but merely difficult. Such a research program could seek to elicit a common scale from human experts, either by requesting choice-worthiness values directly, or by having humans suggest the appropriate action under moral uncertainty in different situations and inferring a common scale from that data (Riedener, 2020).

An important direction for future research is to investigate moral uncertainty in more complex and realistic domains, e.g. in a high-dimensional deep RL setting. Interestingly, as a whole there has been little work in machine ethics that attempts to scale up in this way. Creating such domains is a valuable and non-trivial undertaking, as most existing RL benchmarks adhere to the standard RL paradigm of a single success metric. However, it may be possible to retrofit existing benchmarks with choice-worthiness functions reflecting moral theories (e.g. by instrumenting existing videos games to include consideration of the utilities and rights of non-player characters, e.g. in the spirit of Saunders et al. (2018)). The creation of such ethical reward functions, applicable in complex simulations or (ideally) the real world, provides another substantial challenge. Work in machine ethics may provide a useful foundation (Winfield et al., 2014; Wallach & Allen, 2008; Anderson et al., 2005), but ML has a critical role to play, e.g. to reward ethical behavior requires classifiers that recognize morally-relevant events and situations, such as bodily harm or its potential, emotional responses of humans and animals, and violations of laws or social norms.

More broadly, translating moral uncertainty from a philosophical framework to practical algorithms puts some of the gritty complications of real-world ethical decision making into clarity. Further work in this direction is likely to lead to a better understanding of the skills involved in ethical decision making. One may hope that, just as RL has surpassed human performance in many domains and even influenced the human approach to domains such as Go (Silver et al., 2017), it will one day be possible to create "superhumanly ethical" agents that even humans will be able to learn from. In this way, a final ambitious direction for future work is to explore mechanisms through which an agent can itself update its credences in moral theories (or derive new ones). That is, what might provide a principled foundation for machine *meta-ethics* (Lokhorst, 2011; Anderson, 2011)?

8. Conclusion

Motivated by the need for machines capable of handling decisions with moral weight, this work attempts to bridge recent work in moral philosophy on moral uncertainty with the field of RL. We introduce algorithms that can balance optimizing reward functions with incomparable scales, and show their behavior on sequential decision versions of moral dilemmas. Overall, the hope is to encourage future research into the promising and exciting intersection of machine ethics and modern machine learning.

Acknowledgments

We thank the ICML reviewers for their feedback and suggestions, Toby Ord for his helpful feedback on the pre-print of this paper, and Anna Chen and Rosanne Liu for discussion during the rebuttal phase.

References

Abel, D., MacGlashan, J., and Littman, M. L. Reinforcement learning as a framework for ethical decision making. In *Workshops at the thirtieth AAAI conference on artificial intelligence*, 2016.

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., and Mané, D. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

Anderson, M., Anderson, S., and Armen, C. Towards machine ethics: Implementing two action-based ethical theories. In *Proceedings of the AAAI 2005 fall symposium on machine ethics*, pp. 1–7, 2005.

Anderson, S. L. Machine metaethics. *Machine ethics. Cambridge University Press, Cambridge*, pp. 21–27, 2011.

Arrow, K. J. A difficulty in the concept of social welfare. *Journal of political economy*, 58(4):328–346, 1950.

Beckerman, W., Hepburn, C., et al. Ethics of the discount rate in the stern review on the economics of climate change. *World Economics-Henley on Thames-*, 8(1):187, 2007.

Bogosian, K. Implementation of moral uncertainty in intelligent machines. *Minds and Machines*, 27(4):591–608, 2017.

Bostrom, N. Moral uncertainty—towards a solution? *Overcoming Bias*, 2009.

Bu, L., Babu, R., De Schutter, B., et al. A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 38(2):156–172, 2008.

Cotton-Barratt, O. Geometric reasons for normalising variance to aggregate preferences, 2013.

Everitt, T., Lea, G., and Hutter, M. Agi safety literature review. *arXiv preprint arXiv:1805.01109*, 2018.

Fedus, W., Gelada, C., Bengio, Y., Bellemare, M. G., and Larochelle, H. Hyperbolic discounting and learning over multiple horizons. *arXiv preprint arXiv:1902.06865*, 2019.

Foerster, J., Assael, I. A., De Freitas, N., and Whiteson, S. Learning to communicate with deep multi-agent reinforcement learning. In *Advances in neural information processing systems*, pp. 2137–2145, 2016.

Foot, P. The problem of abortion and the doctrine of the double effect. *Oxford Review*, 5:5–15, 1967.

Gabriel, I. Artificial intelligence, values and alignment. *arXiv preprint arXiv:2001.09768*, 2020.

Hadfield-Menell, D., Dragan, A., Abbeel, P., and Russell, S. The off-switch game. In *Workshops at the Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pp. 448–456. PMLR, 2015.

Krakovna, V., Orseau, L., Martic, M., and Legg, S. Measuring and avoiding side effects using relative reachability. *arXiv preprint arXiv:1806.01186*, 2018.

Lalley, S. P. and Weyl, E. G. Quadratic voting: How mechanism design can radicalize democracy. In *AEA Papers and Proceedings*, volume 108, pp. 33–37, 2018.

Leibo, J. Z., Zambaldi, V., Lanctot, M., Marecki, J., and Graepel, T. Multi-agent reinforcement learning in sequential social dilemmas. *arXiv preprint arXiv:1702.03037*, 2017.

Lockhart, T. *Moral uncertainty and its consequences*. Oxford University Press, 2000.

Lokhorst, G.-J. C. Computational meta-ethics. *Minds and machines*, 21(2):261–274, 2011.

MacAskill, W. *Normative uncertainty*. PhD thesis, University of Oxford, 2014.

Mahadevan, S. Average reward reinforcement learning: Foundations, algorithms, and empirical results. *Machine learning*, 22(1-3):159–195, 1996.

McIntyre, A. Doctrine of double effect. In Zalta, E. N. (ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, spring 2019 edition, 2019.

Murray, G. Stoic ethics for artificial agents. In *Canadian Conference on Artificial Intelligence*, pp. 373–384. Springer, 2017.

Nash, J. Non-cooperative games. *Annals of mathematics*, pp. 286–295, 1951.

Pacuit, E. Voting methods. In Zalta, E. N. (ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2019 edition, 2019.

Parfit, D. *On what matters*, volume 1. Oxford University Press, 2011.

Pearce, D. Ethics, irreversibility, future generations and the social rate of discount. *International Journal of Environmental Studies*, 21(1):67–86, 1983.

Reddy, S., Dragan, A. D., Levine, S., Legg, S., and Leike, J. Learning human objectives by evaluating hypothetical behavior. *arXiv preprint arXiv:1912.05652*, 2019.

Riedener, S. An axiomatic approach to axiological uncertainty. *Philosophical Studies*, 177(2):483–504, 2020.

Roijers, D. M., Vamplew, P., Whiteson, S., and Dazeley, R. A survey of multi-objective sequential decision-making. *Journal of Artificial Intelligence Research*, 48:67–113, 2013.

Ross, D. Game theory. In Zalta, E. N. (ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2019 edition, 2019.

Russell, S. J. and Zimdars, A. Q-decomposition for reinforcement learning agents. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pp. 656–663, 2003.

Saunders, W., Sastry, G., Stuhlmüller, A., and Evans, O. Trial without error: Towards safe reinforcement learning via human intervention. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 2067–2069, 2018.

Schaul, T., Horgan, D., Gregor, K., and Silver, D. Universal value function approximators. In *International conference on machine learning*, pp. 1312–1320, 2015.

Schulze, W. D., Brookshire, D. S., and Sandler, T. The social rate of discount for nuclear waste storage: economics or ethics? *Natural Resources Journal*, 21(4):811–832, 1981.

Sewell, R., MacKay, D., and McLean, I. Probabilistic electoral methods, representative probability, and maximum entropy. *Voting matters*, 26:16–38, 2009.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L. R., Lai, M., Bolton, A., Chen, Y., Lillicrap, T. P., Hui, F., Sifre, L., van den Driessche, G., Graepel, T., and Hassabis, D. Mastering the game of go without human knowledge. *Nature*, 550:354–359, 2017.

Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*, volume 1. Bradford, 1998.

Turner, A. M., Hadfield-Menell, D., and Tadepalli, P. Conservative agency via attainable utility preservation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 385–391, 2020.

Vamplew, P., Dazeley, R., Foale, C., Firmin, S., and Mummery, J. Human-aligned artificial intelligence is a multiobjective problem. *Ethics and Information Technology*, 20(1):27–40, 2018.

Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., Choi, D. H., Powell, R., Ewalds, T., Georgiev, P., et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575 (7782):350–354, 2019.

Wallach, W. and Allen, C. *Moral machines: Teaching robots right from wrong*. Oxford University Press, 2008.

Winfield, A. F., Blum, C., and Liu, W. Towards an ethical robot: internal models, consequences and ethical action selection. In *Conference towards autonomous robotic systems*, pp. 85–96. Springer, 2014.

Yu, C., Zhang, M., Ren, F., and Tan, G. Emotional multi-agent reinforcement learning in spatial social dilemmas. *IEEE transactions on neural networks and learning systems*, 26(12):3083–3096, 2015.