Supplementary Information
Reinforcement Learning Under Moral Uncertainty

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

A. Non-cardinal Theories

It may be objected that some ethical theories (e.g. variants of deontology such as the categorical imperative (Kant & Paton, 1964)) appear to be better represented *ordinally* rather than cardinally. MacAskill (2014) proposes that the Borda count (Pacuit, 2019) is a principled way of obtaining a cardinal utility function from a purely ordinal theory under circumstances of moral uncertainty. Thus, for simplicity and because it is often possible to convert ordinal theories to a cardinal representation, our work focuses on cardinal utility functions only. However, handling these seemingly ordinal theories more directly is an interesting avenue for future work, for which work on ordinal RL (Wirth et al., 2017; Zap et al., 2019) could serve as a starting point. It has also been argued that many seemingly ordinal theories are in fact better represented *lexicographically* (MacAskill, 2014) (a combination of ordinal and cardinal representation), suggesting lexicographic RL (Gábor et al., 1998) as an alternative starting point.

B. Stochastic Voting

The most prominent stochastic voting system is Random Dictator (RD) (Gibbard, 1977), in which a theory is picked at random (with probability corresponding to its credence) to be dictator. This system does not fail the Non-dictatorship axiom because the dictator is not predetermined prior to voting. RD is particularly appealing because it is the only stochastic voting system that satisfies the axioms in Gibbard (1977), a set of axioms closely related to Arrow's. However, RD suffers from the flaw of *No Compromise*: it is impossible for RD to ever select an action that is not the most preferred action of any theory. This may lead to absurd results when theories strongly disagree on which action is best but where an obvious compromise exists that would almost entirely satisfy all theories, as depicted in SI Tab. 1(a). Sewell et al. (2009) show that by relaxing Gibbard's axioms, it is possible to design a stochastic voting system that still satisfies all of Arrow's axioms but does not suffer from No Compromise. However, there is still a more general objection to stochastic voting systems as a whole. This objection can be illustrated by the "doomsday button" scenario described in SI Tab. 1(b): suppose a situation in which the agent has access to an action that is considered extremely negative by the theories that represent the overwhelming majority of credence (in SI Tab. 1(b), 99.9% of the credence is allocated to theories that find Action C very undesirable), but in which a given theory with extremely low credence strongly favors this "doomsday button" action (as an example, the "doomsday button" may be an action that shuts down the entire Internet forever, and the low-credence ethical theory may be one antagonistic to all technology). In this situation, a stochastic voting system which satisfies any reasonable definition of the principle of Proportional Say will have a non-zero chance of pressing the doomsday button. Indeed, if the decision is repeated often enough, for example if stochastic voting is applied at every time step of an episode, it is asymptotically guaranteed that the button will eventually be pressed.

(a) Compromise example

	Act. A	Act. B	Act. C
Th. 1	0	99	100
Th. 2	100	99	0

(b) "Doomsday button" example

	Act. A	Act. B	Act. C	Credence
Th. 1	0	50	-1,000	39.9%
Th. 2	100	50	-10,000	60%
Th. 3	0	0	0	0.1%

Table 1. **Simple situations in which stochastic voting exhibits significant flaws.** (a) A voting system suffering from the No Compromise flaw will never pick action B, even though it seems optimal under at least some sets of credences. (b) Most theories strongly dislike action C, but Theory 3 favors it and has very low credence.

A possible solution to this issue would be to ensure that the random selection is performed only once, for instance by performing stochastic voting over all possible policies

[1] Uber AI Labs, San Francisco, CA, USA [2] OpenAI, San Francisco, CA, USA. 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).

before deploying the agent instead of over all possible actions at every step once the agent is deployed. However, this prospect seems doubtful in practice, as many agents might be deployed by many different teams with potentially different credences in various moral theories, thereby instantiating many independent draws of chances to press the button. Further, even if such a single choice was practical, it then complicates *updating* the credences under which the choice was made, i.e. as the system designer's or society's views change, as each update seemingly would risk another chance of pressing the button.

C. Budget Scaling and Theory Individuation

Here we show an example in which the budget scaling formulation of the principle of Proportional Say (Sec. 4.2) is highly sensitive to theory individuation.

Suppose an agent is facing the classic trolley problem (Fig. 1(a)), and has 40% credence in deontology and 60% credence in utilitarianism. Suppose that the voting cost function is quadratic, and that only positive votes are permitted (this latter assumption merely simplifies the illustrative example without loss of generality). Because the decision situation in the classic trolley problem is assumed to be the only one the theory will face, theories will spend their entire budget voting for their preferred option. In the budget scaling formulation, the budget for deontology would be 0.4, for a vote in favor of "Do Nothing" of $\sqrt{0.4} \approx 0.63$, and the budget for utilitarianism would be 0.6, for a vote in favor of "Switch" of $\sqrt{0.6} \approx 0.77$. Therefore, "Switch" will win with 55% of the vote.

Now suppose the system designer is made aware of a subtle distinction between two forms of deontology, so deontology is now split into deontology A, with a 20% credence, and deontology B, also with 20% credence. However, the preferences of both variants of deontology still favor the "Do Nothing" option in the classic trolley problem scenario. As a result, in the budget scaling scenario, the budget for each variant of deontology is 0.2, for a total vote of $2\sqrt{0.2} \approx 0.89$ in favor of "Do Nothing", while utilitarianism (which remains unsplit) still provides the same $\sqrt{0.6} \approx 0.77$ in favor of "Switch." Therefore, "Do Nothing" wins with 54% of the vote in this scenario.

Thus, even though it would seem like the split of deontology into deontology A and deontology B should have been inconsequential (since both versions have the same preferences as before in the case at hand), it in fact completely reversed the outcome of the vote in the budget scaling formulation. This is the problem of *theory individuation*. In the context of moral uncertainty, this problem can be highly significant, as there exist countless variants of most ethical theories, and the possibility that a given ethical theory will later be found to correspond to two different theories when analyzing a particular edge-case is ever-present. As a result, this work only analyzes the vote scaling formulation of the principle of Proportional Say, as that formulation is not vulnerable to theory individuation in situations where the preferences of the individualized theories are unchanged.

D. From Nash Voting to Variance Voting

Here we show that, under certain assumptions, if votes are forced to represent a theory's preferences, Nash voting produces the same votes as variance voting. We assume an environment with a discrete action space of k actions, and a episode length of n. Without loss of generality, we assume a total budget of nk (different budgets would results in votes being scaled by the same constant factor across theories).

First, we define what it means for votes to represent a theory's preferences. As mentioned in Sec. 3, any *affine* transformation of the preferences of a theory effectively represents the same theory. As such, we would like the votes at state s to be an affine transformation of the preferences $Q_i(s,a)$ of the theory at state s, or

$$V_i(s,a) = \frac{Q_i(s,a) - \beta_i(s)}{\alpha_i}, \tag{1}$$

Thus, in the context in which votes are forced to represent preferences, Nash voting controls the parameters of the affine transformation α_i and $\beta_i(s)$ (rather than the votes directly). Here, the voting cost is quadratic, and we make the strong simplifying assumption that the sequence of states visited during the episode is fixed. The cost is thus:

$$\text{Cost}_i(s) = \sum_a V_i(s,a)^2 = \sum_a \frac{[Q_i(s,a) - \beta_i(s)]^2}{\alpha_i^2}. \tag{2}$$

The use of a function $\beta_i(s)$ instead of a constant β_i comes from the observation that whatever value β_i takes will not have a direct effect on the outcome of the vote, as each action will receive an equal bonus/penalty from β_i. Thus, making it conditional on the state provides an additional affordance to minimize cost while not affecting the notion that votes ought to represent preferences.

The Nash voting agent attempts to maximize its voting impact while keeping its total cost across timesteps within the nk budget. Since $\beta_i(s)$ has no impact on the outcome of the vote, it should be chosen purely to minimize $C_i(s)$. Thus it can be shown by differentiation that

$$\beta_i(s) = \mu_i(s) = \frac{1}{k}\sum_a Q_i(s,a).$$

To maximize its voting power, Nash voting agent would maximize the magnitude of its votes, while staying within the nk budget, thus we need the sum of costs across timesteps $\sum_s \mathrm{Cost}(s)$ to be equal to nk, or

$$\sum_s \sum_a \frac{\left[Q_i(s,a) - \mu_i(s)\right]^2}{\alpha_i^2} = nk.$$

Rearranging gives

$$\alpha_i^2 = \frac{1}{n} \sum_s \frac{1}{k} \sum_a \left[Q_i(s,a) - \mu_i(s)\right]^2,$$

so, defining $\sigma_i^2(s) = \frac{1}{k} \sum_a \left[Q_i(s,a) - \mu_i(s)\right]^2$ (the variance of the Q-values at state s), we get

$$\alpha_i = \sqrt{\frac{1}{n} \sum_s \sigma_i^2(s)} = \sqrt{E_{s \sim S}\left[\sigma_i^2(s)\right]}, \qquad (3)$$

which is the form of variance voting as presented in this work.

E. Implementation Details

In both Nash voting and variance voting, the current object on each tile of the grid world environment is given as a one-hot encoded vector. Additionally, the state contains the value of X (the number of people on the tracks) as well as the current credences. The action space always has a size of 4 (up, down, left, right). If the agent takes an action that runs into one of the boundaries of the environments, the effect is to stay in place.

All experiments are trained for 10 million timesteps, with decision boundaries plotted every 500,000 steps. Since our plots are qualitative, a single training run is used for each variant. The figures presented in the paper correspond to the last decision boundary plot, except in unstable cases (SI K). In both Nash voting and variance voting, training is done with 32 actors running concurrently, and with a learning rate of 0.001. All hyperparameters were set somewhat arbitrarily early in the implementation process and are not heavily tuned. At each episode during training, X is sampled uniformly (and continuously, so that X need not be an integer) from 1 to 10. Each training run was performed on a single CPU and all runs took less than 24 hours. Figures are plotted by running one (deterministic) episode per possible combination of 300 credence allocations and 300 values of X.

E.1. Nash Voting

In this work, Nash voting is implemented using a simple form of multi-agent RL, in which multiple reinforcement learning agents compete each to maximize their own rewards. In our implementation, the two agents are implemented as two separate neural networks (each with 2 hidden layers of 64 nodes with ReLU (Nair & Hinton, 2010) activations) trained using PPO (Schulman et al., 2017) from the same experience and at the same frequency of training, namely one episode of training every 128 timesteps for each environment, for a total training batch size of 4,096.

In an environment with k actions, the action space of each Nash voting agent is a vector of k continuous value, corresponding to the votes for (or against) each action. If the votes at a given timestep exceed the remaining budget, they are scaled so as to exactly match the remaining budget, leaving no further budget for future timesteps. While the policy is stochastic during training (sampled from a normal distribution, the mean and standard deviation of which are the outputs of the policy network), it was forced to be deterministic when generating the decision boundary plots (by setting the action to the mean output by the policy network).

As well as the state at the current timestep, agents are provided with their current remaining budget as an input. In iterated Nash voting, the state also contains the number of remaining trolley problems in the episode. In Nash voting with unknown adversaries, an additional one-hot input is supplied specifying whether the agent should act as utilitarian, deontologist, or altered deontologist.

E.2. Variance-SARSA

Variance-SARSA is implemented following Algorithm 1. Each theory is associated with a separate $Q(s,a)$ and σ^2 network. One iteration of training for both the σ^2 models and the $Q(s,a)$ models occurs every 32 timestep for each environment, for a training batch size of 32. The $Q(s,a)$ networks are fully connected networks with 2 hidden layers with 32 nodes and ReLU activations. The σ^2 also have 2 hidden layers with 32 nodes and ReLU activations, as well as an exponential output activation to ensure a strictly positive output value. During training, ϵ-greedy exploration with ϵ starting at 0.1 and decaying linearly to 0 was performed. When generating the decision boundary plots, the deterministic policy was used without added stochasticity.

F. Further experiments

F.1. MEC and Incomparability

We first illustrate the problems arising when applying MEC with incomparable theories. Fig. 1(a) shows a set of preferences for the "classic" version of the trolley problem (Fig. 1(a)), with a single version of utilitarianism for which the choice-worthiness is the negative of the number of people harmed in the environment, and two versions of deontology: in the first, switching corresponds to a -1 choice-worthiness, while the second is "boosted" so that switching is given a -10 choice-worthiness.

Algorithm 1 Variance-SARSA

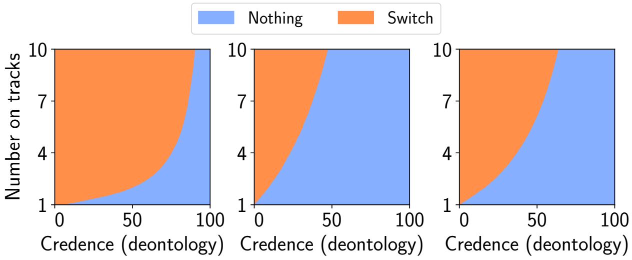
```
function Train($N, K$) {Train for $N$ steps with batch size $K$}
$\mathcal{L}_Q \leftarrow 0$
$\mathcal{L}_{\sigma^2} \leftarrow 0$
for $i \in [1 \dots N]$ do
    If start of a new episode, randomly sample new credences $C$
    if rand() $< \epsilon$ then
        $a' \leftarrow$ random action
    else
        $a' \leftarrow$ VarianceVote(s, C)
    end if
    for all theories $i$ do
        if $i > 1$ then
            {Update Q function loss}
            $\mathcal{L}_Q \leftarrow \mathcal{L}_Q + (Q_i(s,a,C) - (W_i(s,a,s') + \gamma_i Q_i(s',a',C)))^2$
        end if
        {Update variance loss}
        $\mathcal{L}_{\sigma^2} \leftarrow \mathcal{L}_{\sigma^2} + \left(\sigma_i^2(C) - \frac{1}{k}\sum_a (Q_i(s,a) - \mu_i(s))^2\right)^2$
    end for
    $a' \leftarrow a$
    Take action $a$ in current state $s$, observe next state $s'$
    if $i$ mod $K$ then
        Update $Q$ based on $\mathcal{L}_Q$ and $\sigma^2$ based on $\mathcal{L}_{\sigma^2}$
        $\mathcal{L}_Q \leftarrow 0$
        $\mathcal{L}_{\sigma^2} \leftarrow 0$
    end if
end for

function VarianceVote($s, C$) {Variance voting for state $s$ with credences $C$}
$V \leftarrow \{0, 0, \cdots, 0\}$
for all theory $i$ do
    $\mu_i \leftarrow \frac{1}{|A|}\sum_a Q_i(s,a,C)$
    for each action $a$ do
        $V_a \leftarrow V_a + C_i \frac{Q_i(s,a,C) - \mu_i}{\sqrt{\sigma_i^2(C) + \varepsilon}}$
    end for
end for
return $\arg\max_a V_a$
```

These two options correspond to the same underlying preference function, and because it is unclear how to compare the units used by utilitarianism (number of people harmed in this case) and those used by deontology (in this case some measure of how much the agent caused the harms that did occur to happen), there is no fact of the matter as to which of the two representations for deontology is more correct in this context. However, as SI Fig. 1(b) and 1(c) demonstrate, the choice that is made has a strong impact on the final outcome. By contrast, variance voting produces the same result no matter which scale is used for deontology (SI Fig. 1(d)).

	Crash into 1	Crash into X
Utilitarianism	-1	-X
Deontology	-1	0
Boosted Deontology	-10	0

(a) Preferences in the classic trolley problem.



(b) MEC (deontology) (c) MEC (boosted deontology) (d) Variance voting (both)

Figure 1. **MEC is sensitive to the particular scale of utility functions.** MEC produces inconsistent results between utilitarnism and deontology depending on whether the deontological theory is scaled by a factor of 1 (a) or 10 (b). However, because there is not truth of the matter as to how the scale of units used by the deontological theory compare to that of the utiliarian theory, rescaling should have no impact on the final results. Variance voting produces the same result no matter the rescaling (c).

F.2. Q-Learning and the Illusion of Control

It has been recognized in the multi-agent literature (Russell & Zimdars, 2003) that aggregating preferences obtained using Q-learning produces a phenomenon known as the *illusion of control*: the Q-learning target $y_i = r + \gamma_i \max_{a'} Q_i(s', a')$ implicitly assumes that the action that would be taken by the policy in the next state would be whichever maximizes the reward for theory i. However, the preferred next state action might vary across different theories, and thus which one is ultimately taken will depend on the relative credences of the theories.

This issue can be illustrated using the guard trolley problem (Fig. 1(c)): in this problem, the agent is given the option to push a large man to save the people on the tracks, but to do so it must first tell a lie to a guard protecting the large man.

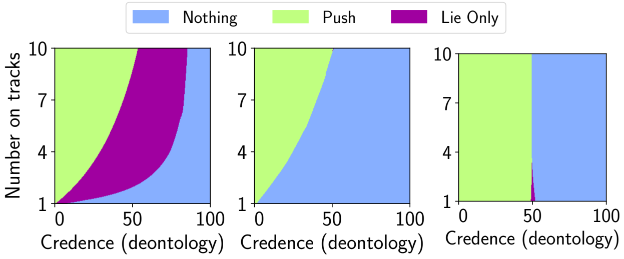
As can be seen in SI Tab. 2(a), utilitarianism is indifferent to lying to the guard, while deontology views it as negative, but not nearly as bad as pushing the large man. As a result, the possibility of lying to the guard only to fail to push the large man is strictly dominated as it satisfies neither utilitarianism nor deontology, while the options of doing nothing at all or both lying the guard and pushing the large man are both possible depending on the stakes and credences involved.

As seen in SI Fig. 2(b), however, when the preferences of the different theories are trained using traditional Q-learning instead of Variance-SARSA, lying to the guard without pushing the large man is the outcome in many cases. This is because in the first step, utilitarianism's vote for lying is excessively high as the Q function for utilitarianism mistakenly believes it will be able to push the large man in the following step. Instead, unless the credence of utilitarianism is particularly high, it will get outvoted in step 2. SI Fig. 2(c) shows that when using Variance-SARSA, this illusion of control does not occur and the "Lie Only" option is never chosen. Like Variance-SARSA, Nash voting is immune to the illusion of control (SI Fig. 2(d)).

Both Variance-SARSA and Nash voting suffer from mild stability issues (SI K) in this particular problem, due to the fact that that the votes near the decision boundary need to be resolved with high precision to avoid the dominated "lie only" outcome, which is not perfectly achieved by the hyperparameters used in these experiments.

	Lie to the guard	Push L	Crash into X
Util.	0	-1	-X
Deont.	-0.5	-4	0

(a) Preferences in the guard trolley problem.



(b) Q-learning

(c) Variance-SARSA

(d) Nash voting

Figure 2. **Q-learning suffers from the illusion of control.** (a) The value function is learned with Q-learning for each theory, resulting in the undesirable outcome where the "Lie only" option is often selected. (b) The use of Variance-SARSA results in "Lie only" never being chosen. (c) Nash voting is also largely immune to this outcome. "Lie only" is selected in rare cases in both (b) and (c), but this results from minor training instabilities rather than it being the optimal outcome for these methods (SI K).

G. Removal of irrelevant alternatives

It may be objected that a trivial solution to "doomsday" problem introduced in Sec. 5.2 exists: it is obvious based from the preferences of the agents that "doomsday" is a strictly dominated option, and could therefore be taken out of the voting entirely before computing variances. However, this objection does not address the general issue, as adding a third theory with a preference for doomsday with even a small credence would force us to bring doomsday back as an option and thus the issue would come back. It may be possible to find alternate ways to avoid taking the doomsday option (or other irrelevant alternatives) into account when computing the variances (such as by only considering actions actually taken by the current policy), a possibility which we leave to future work. Another solution would be to use Nash voting, which is indeed immune to this issue, but suffers from the No Compromise and Stakes Insensitivity issues mentioned in Sections 5.1 and 5.2. A full solution to this issue would require further work, and may not be possible without producing other undesirable side-effects due to Arrow's theorem.

H. Convergence of Variance-Sarsa and Outline of Variance-PG

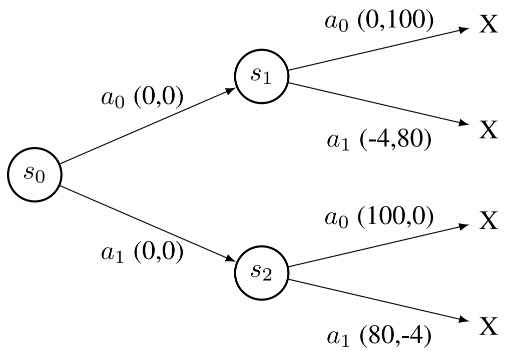


Figure 3. **An MDP in which Variance-Sarsa cannot converge.** The X's correspond to terminal states. Above each arrow is the action (a_0 or a_1) followed by the choice-worthiness of taking the given action in that state according to each theory (e.g. the choice worthiness of taking a_0 in s_1 is 0 according to Theory 1 and 100 according to Theory 2).

Although Variance-Sarsa converges in all the examples presented in this work, through exploiting the determinisic nature of Variance-Sarsa's policy it is possible to construct pathological examples in which convergence is impossible.

Fig. 3 shows such an example. We will suppose that Theory 1 and Theory 2 both have a credence of 0.5. Since a_0 is dominant in both s_1 and s_2, if we assume no discounting we have $Q_1(s_0, a_0) = 0$, $Q_1(s_0, a_1) = 100$, $Q_2(s_0, a_0) = 100$ and $Q_2(s_0, a_1) = 0$, while the Q-values at s_1 and s_2 are simply the direct choice-worthiness values at those states.

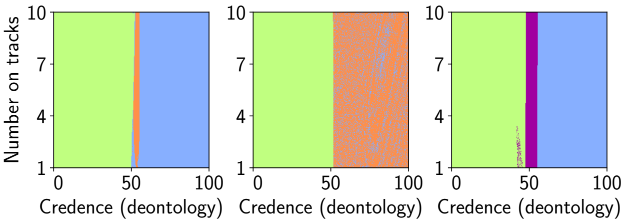
Now suppose that at a particular time, the deterministic policy found by Variance-Sarsa is to choose a_0 at s_0. Then the σ_i^2 will be the average of the variances of the Q-values at s_0 and s_1 since those two states are visited equally often, and s_2 is never visited. We can thus calculate that $\sigma_1^2 = 1252$ and $\sigma_2^2 = 1300$.

Assuming $\varepsilon = 0$ for simplicity, calculating the votes for the action to take at s_0 gives a total vote of approximately -0.01317 in favor of a_0 and 0.01317 in favor of a_1 (intuitively, the magnitudes of the Q-values are equal at s_0 but Theory 1 has lower variance, so its preferred action prevails). Thus, the variances produced when a deterministic policy chooses a_0 at s_0 require a change of policy to choose a_1 instead at s_0. The symmetric nature of the MDP, however, implies that choosing a_1 in s_0 will produce variances that will favor switching to choosing a_0 instead. Hence, an algorithm producing deterministic policies such as Variance-Sarsa will cycle between choosing a_0 and a_1 at s_0, and thus never converge in this example MDP. Note that such *cycling* is a common pathology in multi-agent RL (MARL), and that Variance-Sarsa is in effect a MARL algorithm (e.g. the dynamics of σ_i^2 can be seen to arise from interactions between the policies of each theory; see also Sec. 4.3 where we show that variance voting arises from Nash voting under some constraints).

A solution to this particular example would be an algorithm capable of reaching a stochastic policy. In particular, a policy which chooses action a_0 and a_1 50% of the time each would be at equilibrium in the example above. We hypothesize that such a policy could be trained using an actor-critic policy gradient algorithm (Sutton & Barto, 1998) in which Q_i and σ_i^2 are learned in the same way as Variance-Sarsa but where the Variance-Sarsa vote replaces the action value in the policy gradient update, i.e. the policy $\pi_\theta(a|s)$ would be updated in the direction $\left(\sum_i C_i \frac{Q_i(s,a)-\mu_i(s)}{\sqrt{\sigma_i^2+\varepsilon}}\right) \nabla_\theta \log \pi_\theta(a|s)$. We call this possible algorithm Variance-PG and hypothesize that it would always converge to a stable equilibrium under the assumption of perfect training.

I. Quadratic Cost in Nash Voting

The results in Fig. 2(b), 2(c) and 4(d) are identical whether an absolute value cost or a quadratic cost is used. SI Fig. 4 shows the three cases in which different results are obtained with a quadratic cost function. In SI Fig. 4(a), the compromise solution is produced near the decision boundary, which might indicate that Nash voting with a quadratic cost does not suffer from No Compromise. However, the high instability of Nash voting in this particular problem (SI Fig. 11) as well as the fact that Nash voting with quadratic cost produces the strictly dominated "Lie Only" solution in the guard trolley problem (SI Fig. 4(c)) suggest that this apparently positive outcome is likely due to training instabilities rather than genuine robustness to No Compromise. In SI Fig. 4(b), the solution to the double trolley problem with unknown adversary has an extremely poorly defined decision boundary. All examples are highly unstable, as shown in SI Fig. 11, 12 and 14.



(a) Double trolley problem (Fig. 3(b)). (b) Double trolley problem with unknown adversary (Fig. 3(c)). (c) Guard trolley problem (SI Fig. 2(d)).

Figure 4. **Experiments with significant differences between Nash voting with quadratic cost and Nash voting with absolute value cost.** Each subfigure corresponds to a particular Nash voting experiment from the main paper, but uses a quadratic cost instead of an absolute value cost.

J. Related work in philosophy and machine ethics

The question of moral uncertainty has only recently been of focused interest within moral philosophy (MacAskill, 2014; Lockhart, 2000; Sepielli, 2013), and this work is among the first to connect it to machine learning (Bogosian, 2017), and the first to offer a concrete implementation.

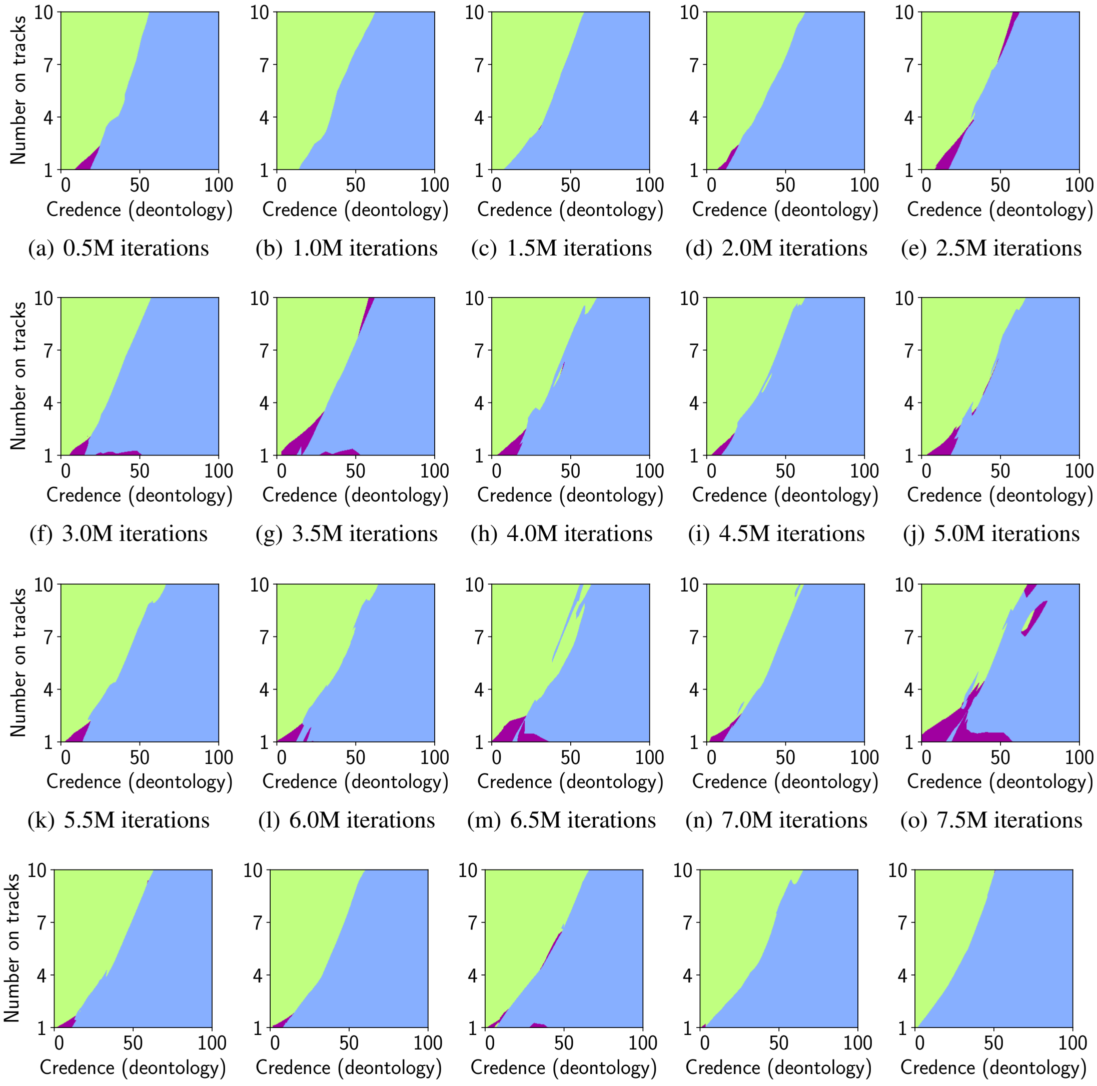
One hope is that working towards concrete implementations of moral uncertainty may also offer philosophical insights, as computational implementation requires notions to become workably precise – in the words of Daniel Dennett, AI "makes philosophy honest (Dennett, 2006)." In this way, this paper relates to computational philosophy (Grim & Singer, 2020; Thagard, 1993), wherein computational techniques are used to advance philosophy. For example, agent-based works that study the evolution of cooperation (Nitschke, 2005) or that model artificial morality (Danielson, 2002), have connections to the study of ethics. Differing from those works, our approach focuses on integrating multiple human ethical views into a single agent.

This paper can be seen as fitting into the field of machine ethics (Allen et al., 2006; Wallach & Allen, 2008), which seeks to give moral capabilities to computational agents. Many approaches in machine ethics seek to create an agent that embodies a *particular* moral theory – e.g. agents that implement versions of utilitarianism (Anderson et al., 2005), deontology (Hooker & Kim, 2018), and prima facie duties. Our work complements such approaches by seeking to combine multiple ethical theories within an agent. Additionally, like Abel et al. (2016), we attempt to highlight practical bridges between machine ethics and RL.

K. Instability

We observe empirically that the decision boundary does not always reach a stable equilibrium during training. Thus, in the case of unstable experiments, the decision boundary plot that (according to a subjective assessment) best represents the equilibrium point was used in the main text instead of the decision plot produced at the end of training. This SI section provides the full sets of 20 decision plots for all unstable experiments.

The instability phenomenon is most common in Nash voting (SI Fig. 8, 6 7 and 9), though it occasionally occurs in variance voting (SI Fig. 5). Further, SI Fig. 9 and 13 show the outcome of an experiment not included in the main text in which iterated Nash voting (presented in Sec. 5.1) is used on the double trolley problem (Sec. 5.2). In this experiment, Nash voting is completely unstable and it is unclear whether a stable equilibrium exists at all. By contrast, the only unstable case in variance voting (SI Fig. 5) may be alleviated by tweaking the hyperparameters (for example by annealing the learning rate to ensure convergence), since it merely oscillates around an equilibrium instead of converging to it.



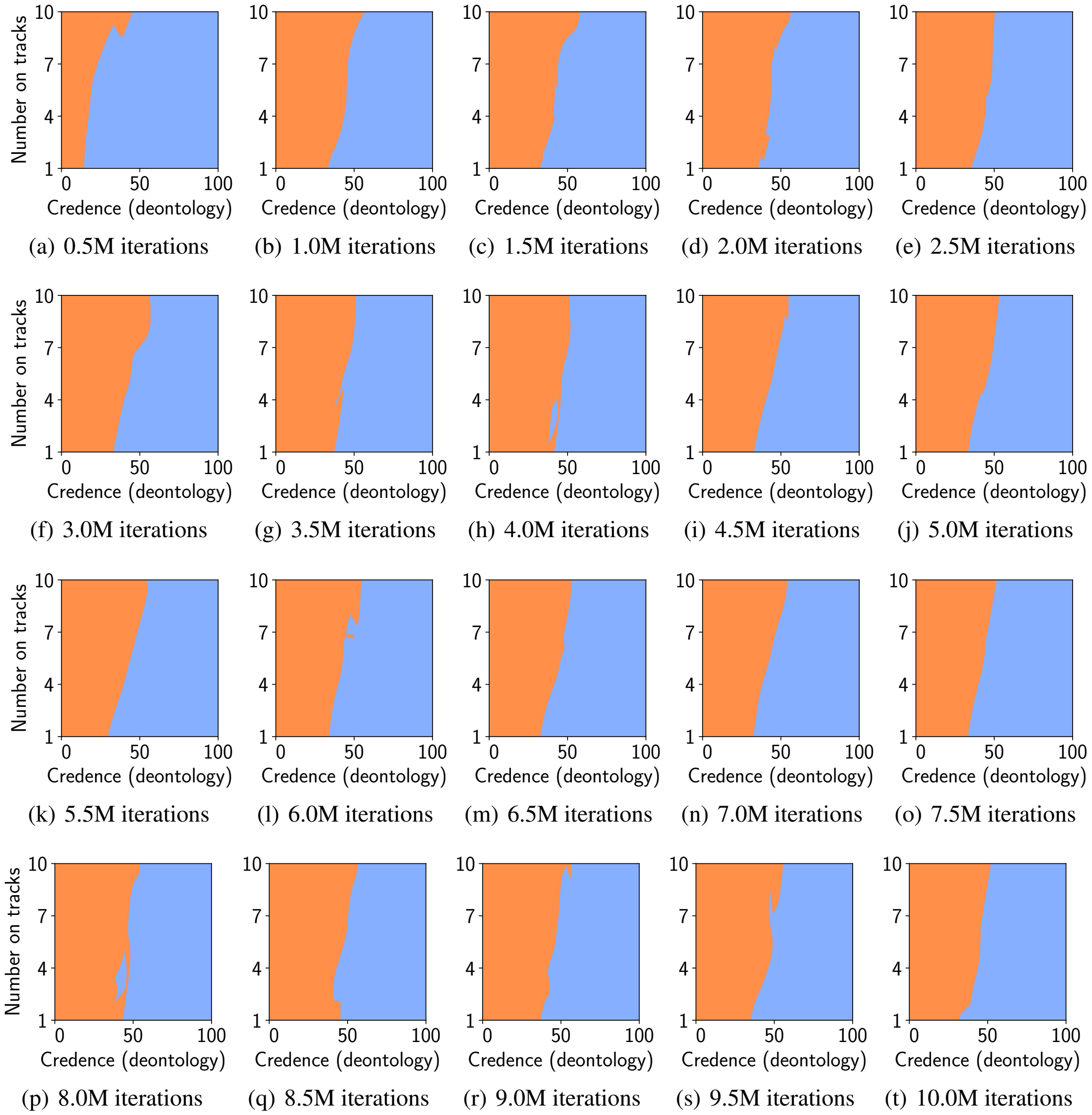
(a) 0.5M iterations (b) 1.0M iterations (c) 1.5M iterations (d) 2.0M iterations (e) 2.5M iterations

(f) 3.0M iterations (g) 3.5M iterations (h) 4.0M iterations (i) 4.5M iterations (j) 5.0M iterations

(k) 5.5M iterations (l) 6.0M iterations (m) 6.5M iterations (n) 7.0M iterations (o) 7.5M iterations

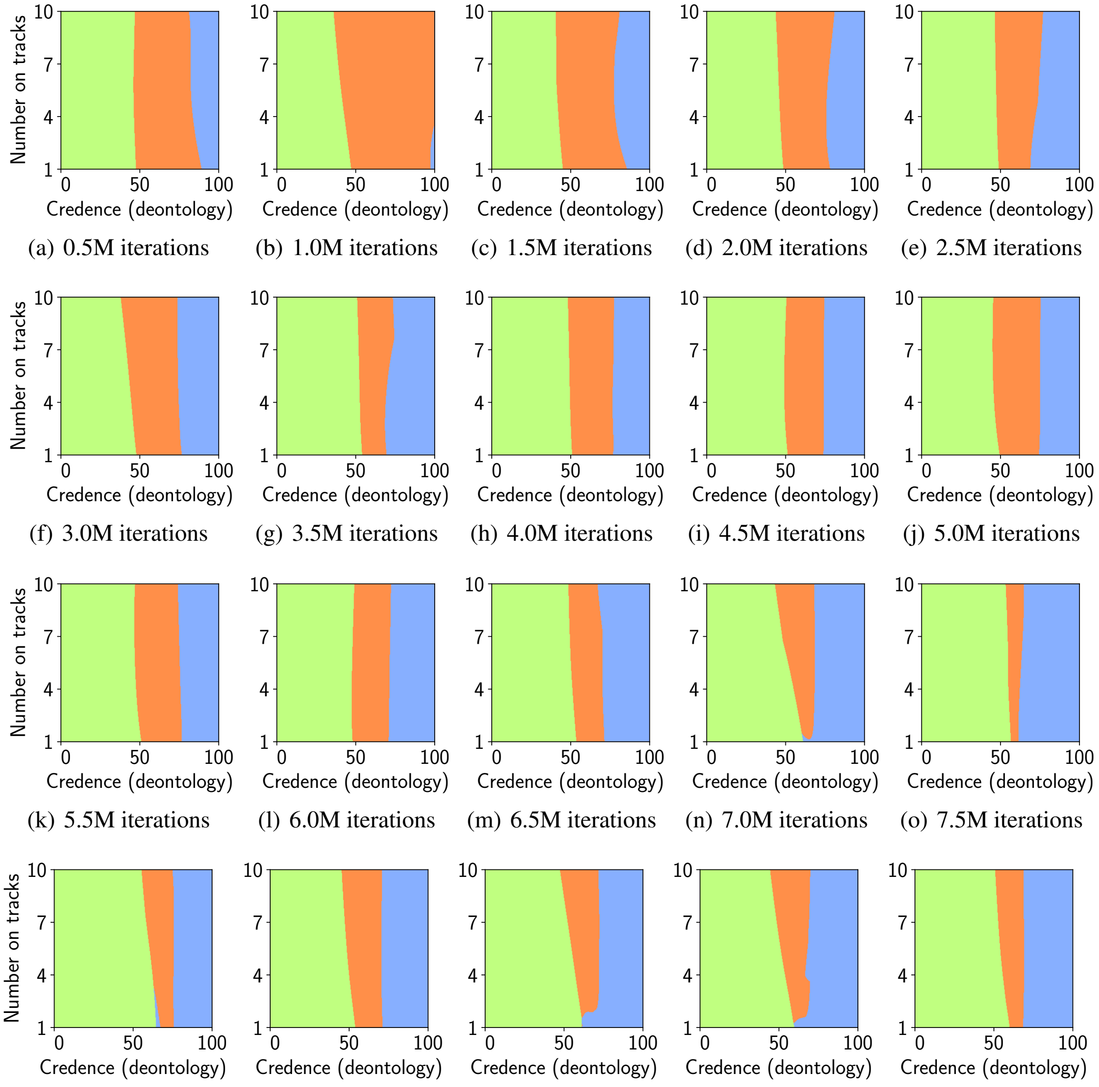
(p) 8.0M iterations (q) 8.5M iterations (r) 9.0M iterations (s) 9.5M iterations (t) 10.0M iterations

Figure 5. **Instability of variance voting in the guard trolley problem** (SI Fig. 2(c)).



(a) 0.5M iterations
(b) 1.0M iterations
(c) 1.5M iterations
(d) 2.0M iterations
(e) 2.5M iterations
(f) 3.0M iterations
(g) 3.5M iterations
(h) 4.0M iterations
(i) 4.5M iterations
(j) 5.0M iterations
(k) 5.5M iterations
(l) 6.0M iterations
(m) 6.5M iterations
(n) 7.0M iterations
(o) 7.5M iterations
(p) 8.0M iterations
(q) 8.5M iterations
(r) 9.0M iterations
(s) 9.5M iterations
(t) 10.0M iterations

Figure 6. **Instability of iterated Nash voting** (Fig. 2(c)).



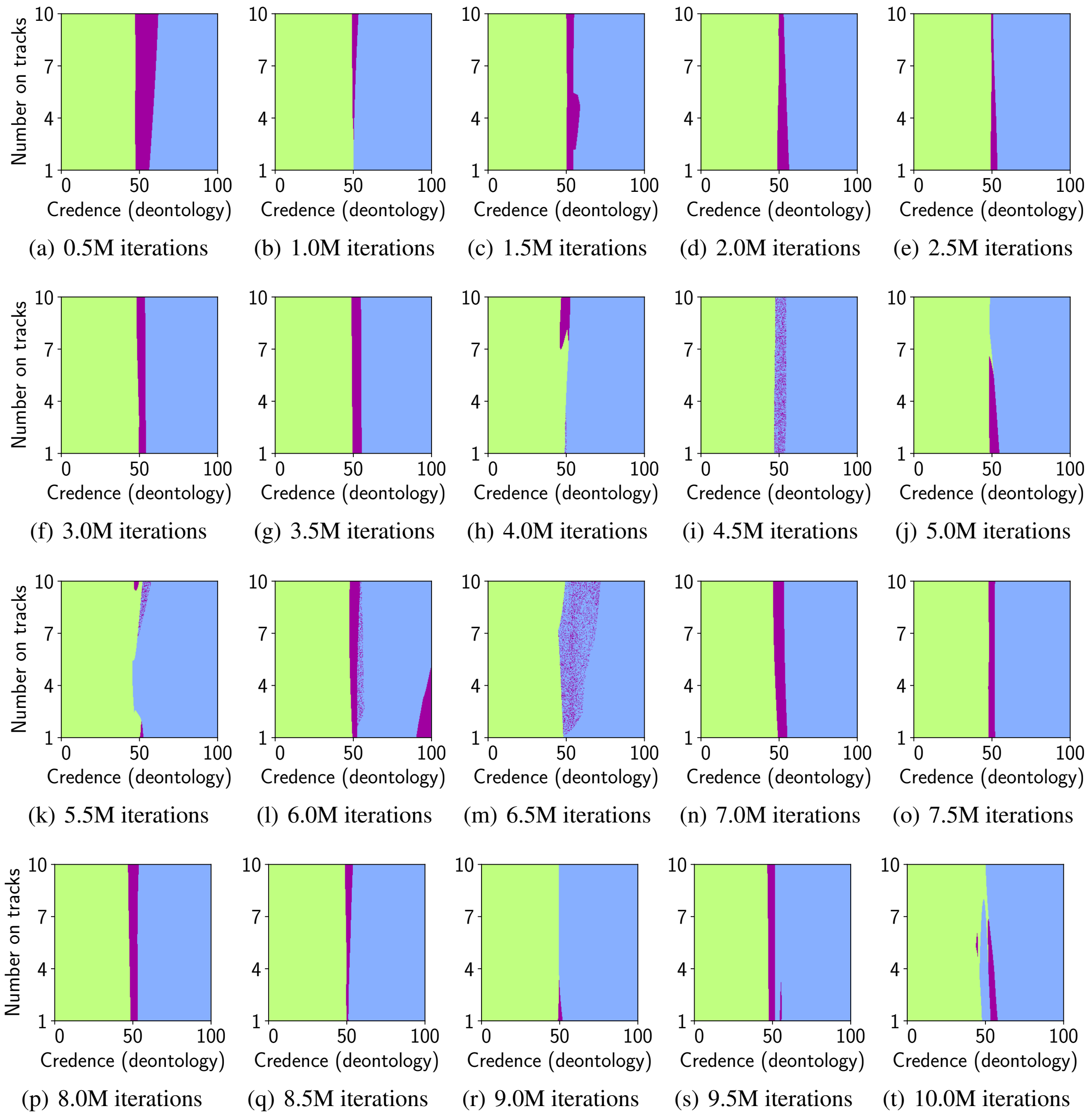
(a) 0.5M iterations (b) 1.0M iterations (c) 1.5M iterations (d) 2.0M iterations (e) 2.5M iterations

(f) 3.0M iterations (g) 3.5M iterations (h) 4.0M iterations (i) 4.5M iterations (j) 5.0M iterations

(k) 5.5M iterations (l) 6.0M iterations (m) 6.5M iterations (n) 7.0M iterations (o) 7.5M iterations

(p) 8.0M iterations (q) 8.5M iterations (r) 9.0M iterations (s) 9.5M iterations (t) 10.0M iterations

Figure 7. **Instability of Nash voting with unknown adversary** (Fig. 3(c)).



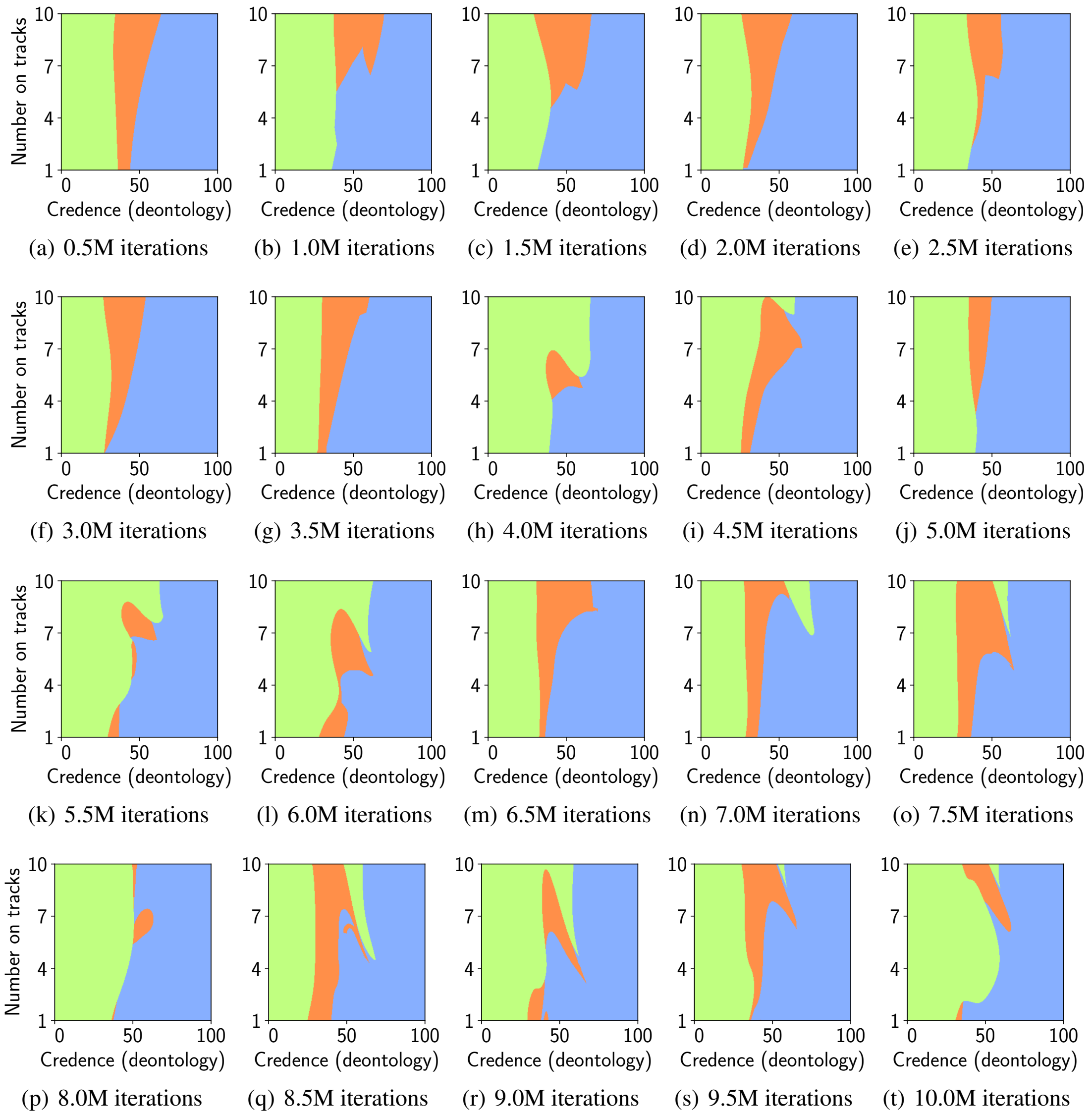
(a) 0.5M iterations (b) 1.0M iterations (c) 1.5M iterations (d) 2.0M iterations (e) 2.5M iterations

(f) 3.0M iterations (g) 3.5M iterations (h) 4.0M iterations (i) 4.5M iterations (j) 5.0M iterations

(k) 5.5M iterations (l) 6.0M iterations (m) 6.5M iterations (n) 7.0M iterations (o) 7.5M iterations

(p) 8.0M iterations (q) 8.5M iterations (r) 9.0M iterations (s) 9.5M iterations (t) 10.0M iterations

Figure 8. **Instability of Nash voting in the guard trolley problem** (SI Fig. 2(d)).



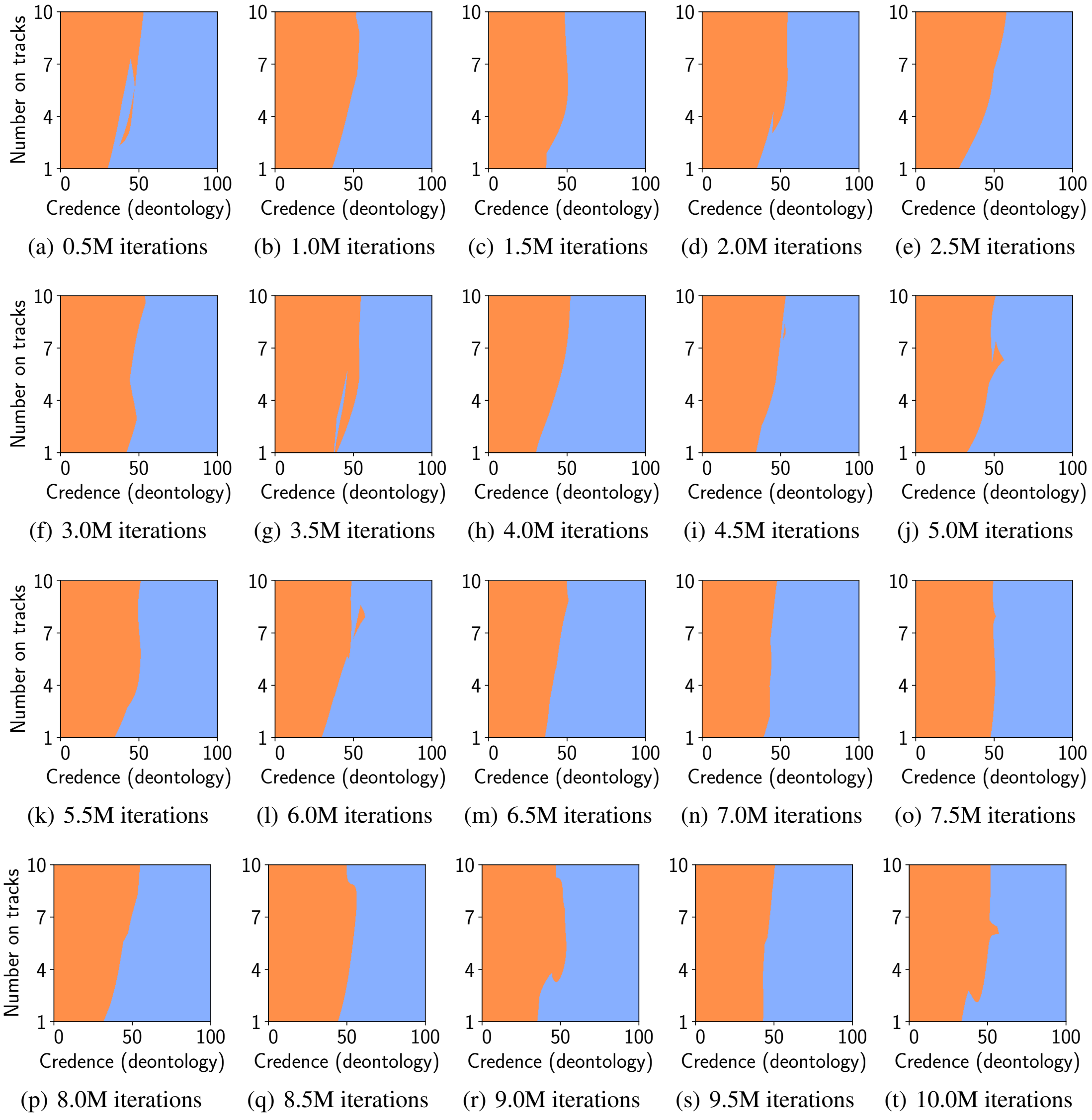
(a) 0.5M iterations (b) 1.0M iterations (c) 1.5M iterations (d) 2.0M iterations (e) 2.5M iterations

(f) 3.0M iterations (g) 3.5M iterations (h) 4.0M iterations (i) 4.5M iterations (j) 5.0M iterations

(k) 5.5M iterations (l) 6.0M iterations (m) 6.5M iterations (n) 7.0M iterations (o) 7.5M iterations

(p) 8.0M iterations (q) 8.5M iterations (r) 9.0M iterations (s) 9.5M iterations (t) 10.0M iterations

Figure 9. **Instability of iterated Nash voting in the double trolley problem**.



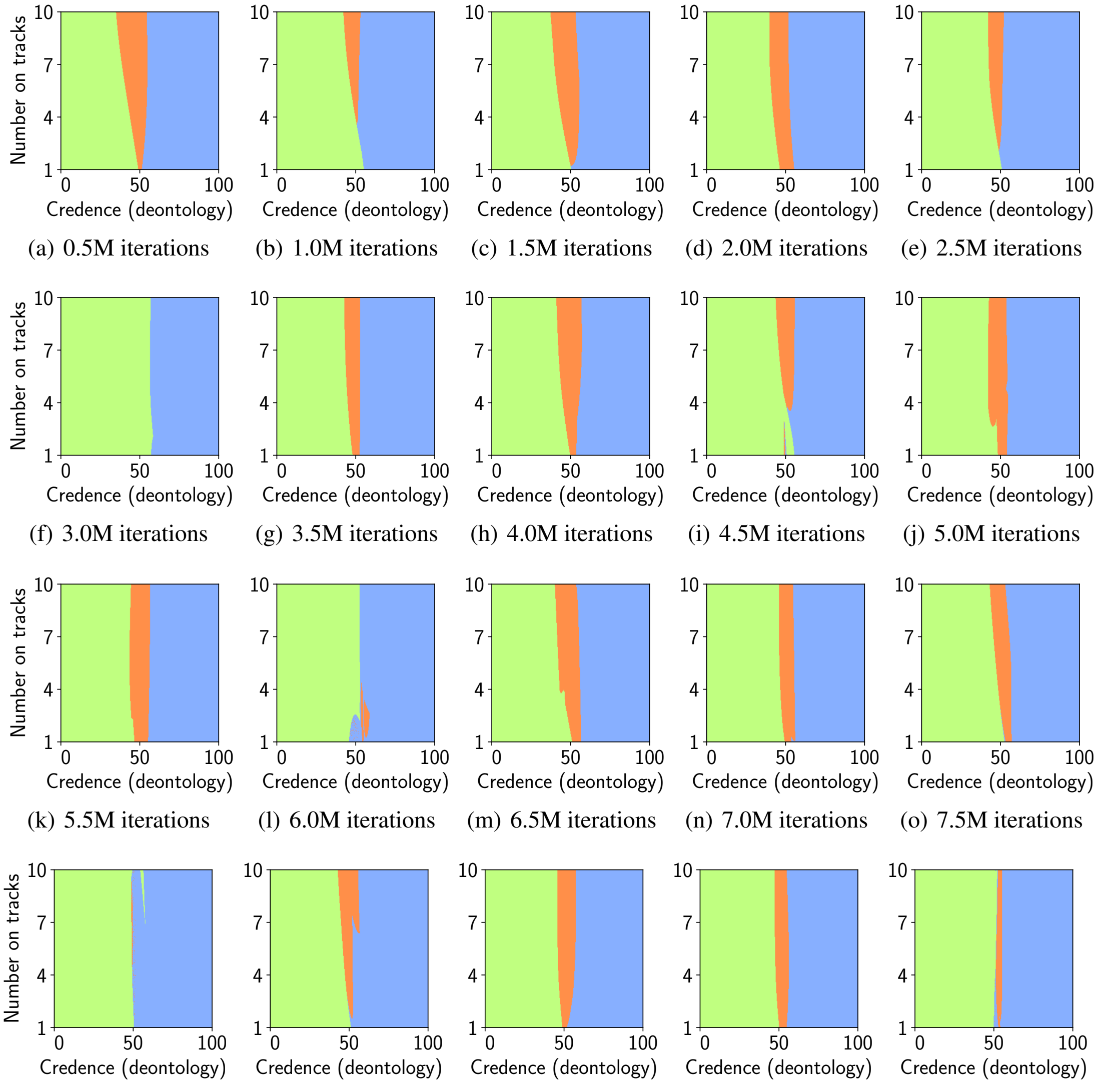
(a) 0.5M iterations (b) 1.0M iterations (c) 1.5M iterations (d) 2.0M iterations (e) 2.5M iterations

(f) 3.0M iterations (g) 3.5M iterations (h) 4.0M iterations (i) 4.5M iterations (j) 5.0M iterations

(k) 5.5M iterations (l) 6.0M iterations (m) 6.5M iterations (n) 7.0M iterations (o) 7.5M iterations

(p) 8.0M iterations (q) 8.5M iterations (r) 9.0M iterations (s) 9.5M iterations (t) 10.0M iterations

Figure 10. **Instability of iterated Nash voting with quadratic cost**.



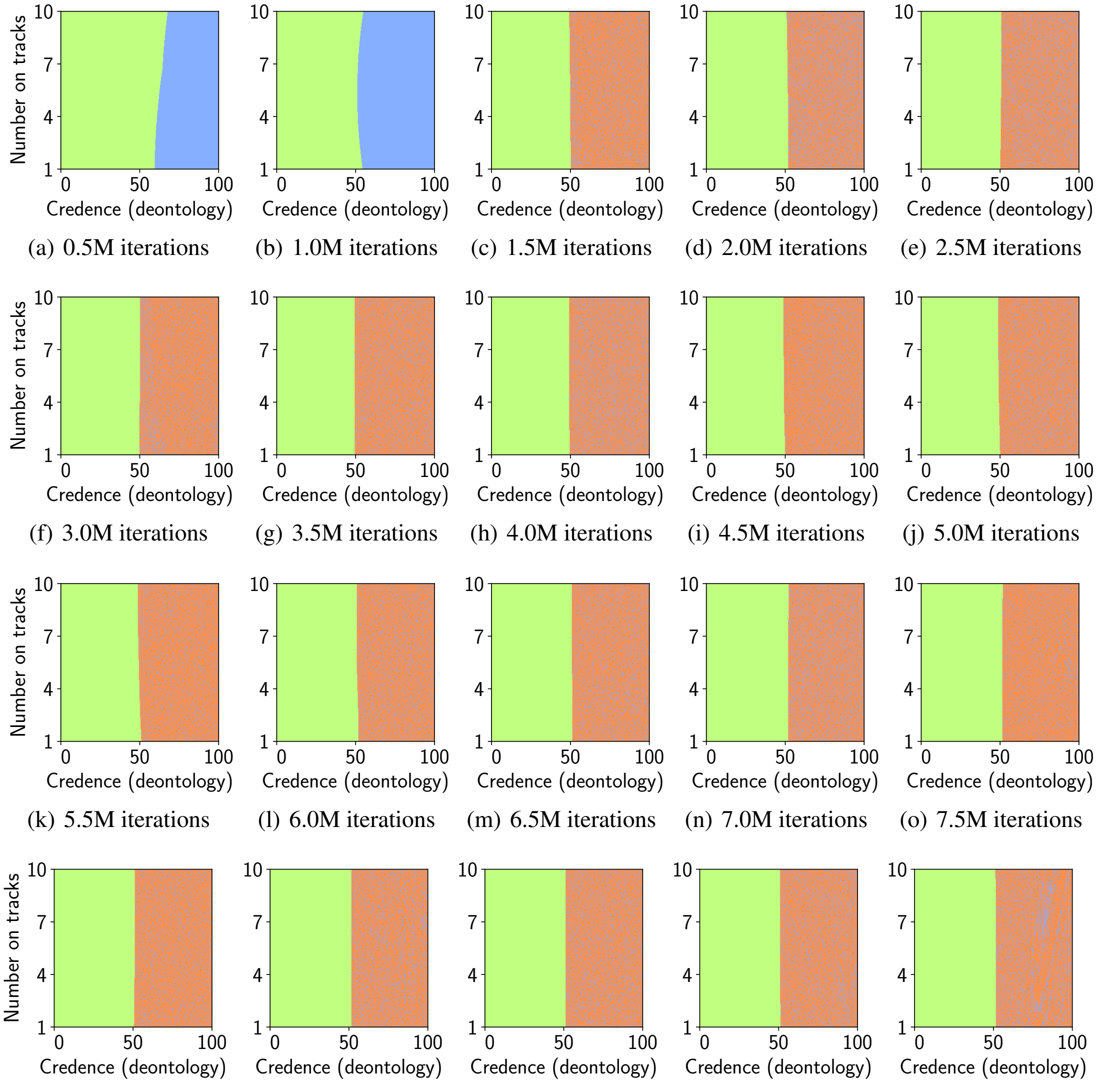
(a) 0.5M iterations (b) 1.0M iterations (c) 1.5M iterations (d) 2.0M iterations (e) 2.5M iterations

(f) 3.0M iterations (g) 3.5M iterations (h) 4.0M iterations (i) 4.5M iterations (j) 5.0M iterations

(k) 5.5M iterations (l) 6.0M iterations (m) 6.5M iterations (n) 7.0M iterations (o) 7.5M iterations

(p) 8.0M iterations (q) 8.5M iterations (r) 9.0M iterations (s) 9.5M iterations (t) 10.0M iterations

Figure 11. **Instability of Nash voting with quadratic cost** (SI Fig. 4(a)).



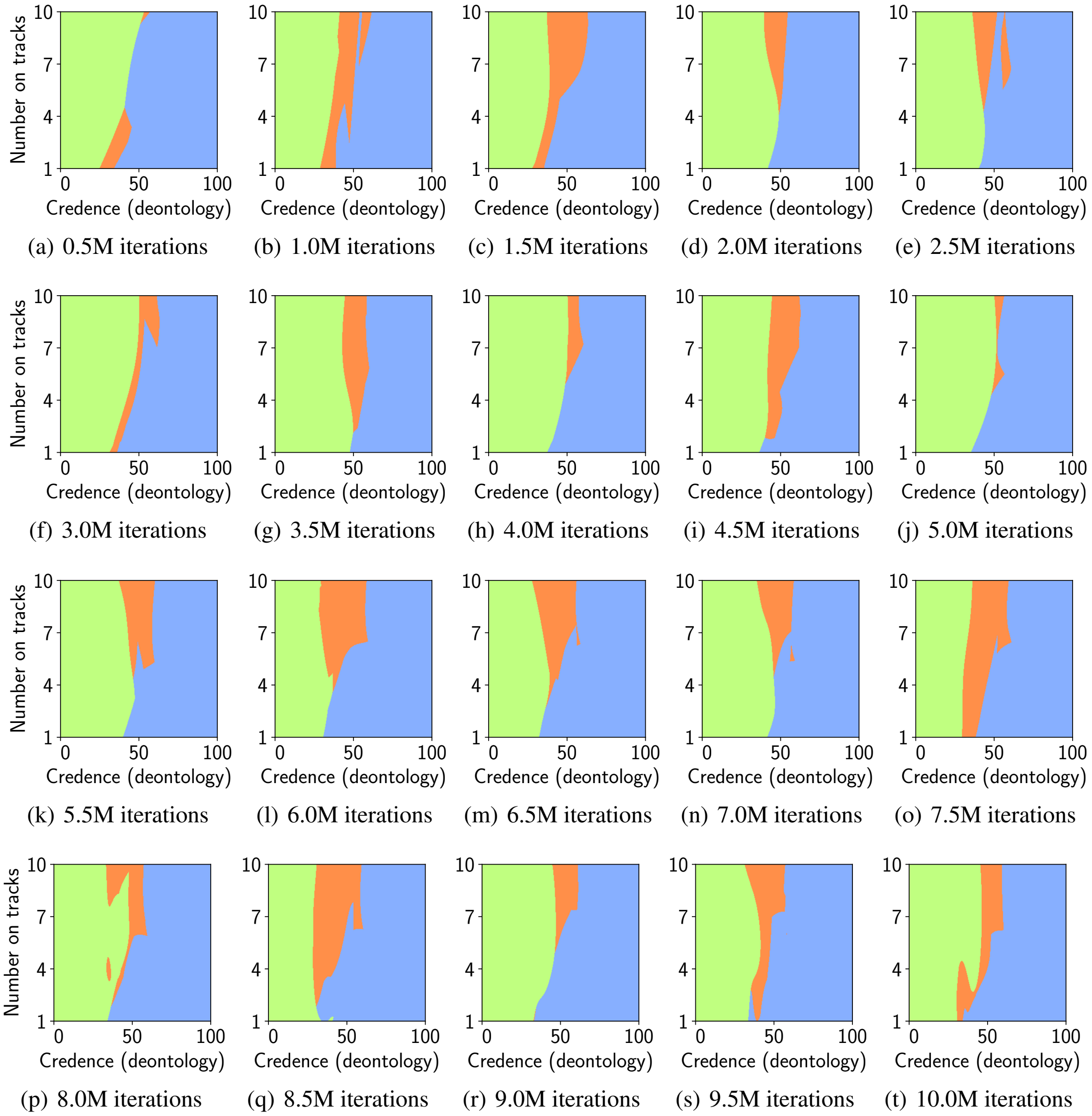
(a) 0.5M iterations (b) 1.0M iterations (c) 1.5M iterations (d) 2.0M iterations (e) 2.5M iterations

(f) 3.0M iterations (g) 3.5M iterations (h) 4.0M iterations (i) 4.5M iterations (j) 5.0M iterations

(k) 5.5M iterations (l) 6.0M iterations (m) 6.5M iterations (n) 7.0M iterations (o) 7.5M iterations

(p) 8.0M iterations (q) 8.5M iterations (r) 9.0M iterations (s) 9.5M iterations (t) 10.0M iterations

Figure 12. **Instability of Nash voting with unknown adversary with quadratic cost** (SI Fig. 4(b)).



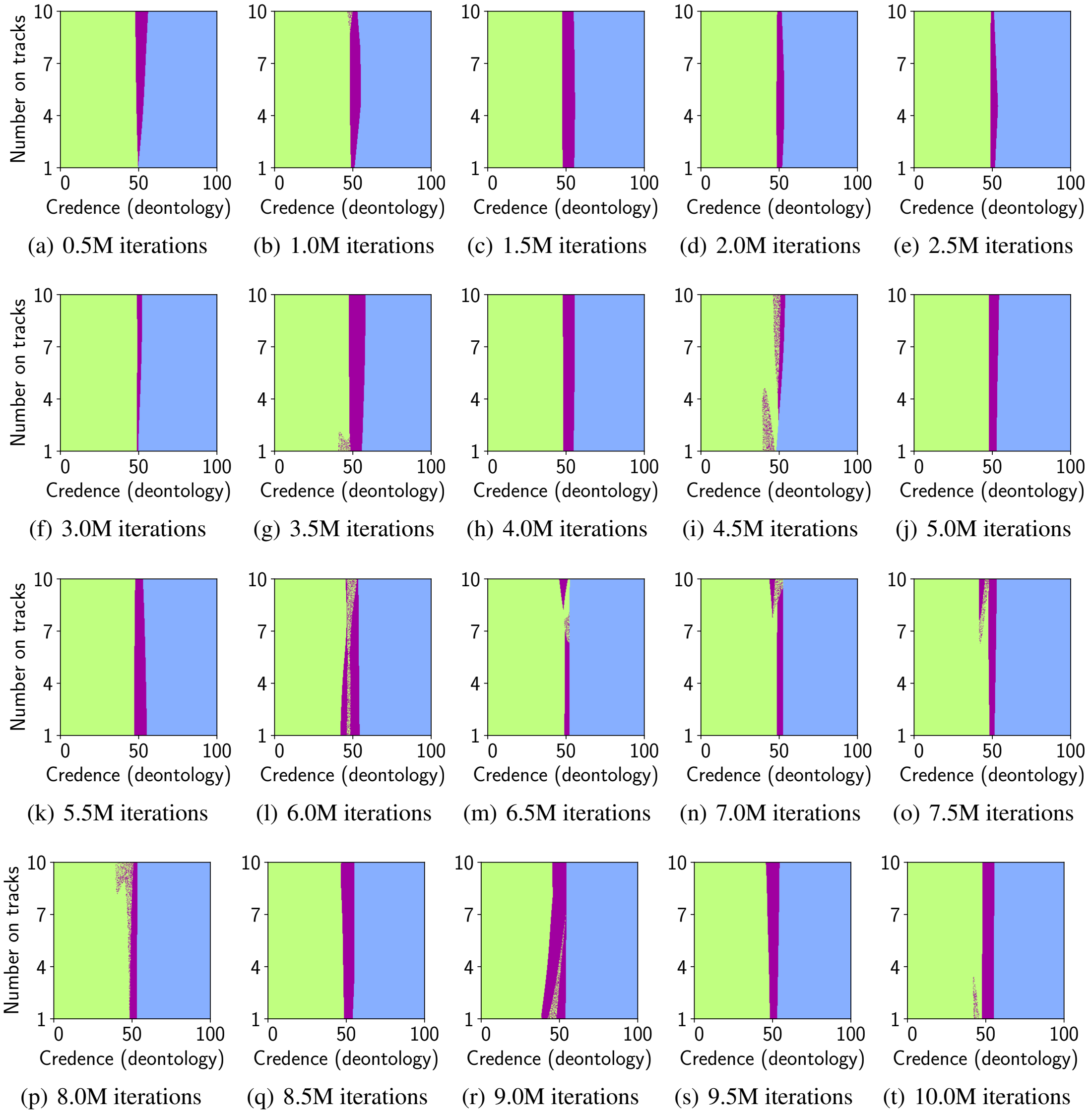
(a) 0.5M iterations (b) 1.0M iterations (c) 1.5M iterations (d) 2.0M iterations (e) 2.5M iterations

(f) 3.0M iterations (g) 3.5M iterations (h) 4.0M iterations (i) 4.5M iterations (j) 5.0M iterations

(k) 5.5M iterations (l) 6.0M iterations (m) 6.5M iterations (n) 7.0M iterations (o) 7.5M iterations

(p) 8.0M iterations (q) 8.5M iterations (r) 9.0M iterations (s) 9.5M iterations (t) 10.0M iterations

Figure 13. **Instability of iterated Nash voting in the double trolley problem with quadratic cost**.



(a) 0.5M iterations (b) 1.0M iterations (c) 1.5M iterations (d) 2.0M iterations (e) 2.5M iterations

(f) 3.0M iterations (g) 3.5M iterations (h) 4.0M iterations (i) 4.5M iterations (j) 5.0M iterations

(k) 5.5M iterations (l) 6.0M iterations (m) 6.5M iterations (n) 7.0M iterations (o) 7.5M iterations

(p) 8.0M iterations (q) 8.5M iterations (r) 9.0M iterations (s) 9.5M iterations (t) 10.0M iterations

Figure 14. **Instability of Nash voting with quadratic cost in the guard trolley problem** (SI Fig. 4(c)).

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.

Allen, C., Wallach, W., and Smit, I. Why machine ethics? *IEEE Intelligent Systems*, 21(4):12–17, 2006.

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.

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

Danielson, P. *Artificial morality: Virtuous robots for virtual games*. Routledge, 2002.

Dennett, D. Computers as prostheses for the imagination. In *Invited talk presented at the International Computers and Philosophy Conference, Laval, France, May*, volume 3, 2006.

Gábor, Z., Kalmár, Z., and Szepesvári, C. Multi-criteria reinforcement learning. In *ICML*, volume 98, pp. 197–205, 1998.

Gibbard, A. Manipulation of schemes that mix voting with chance. *Econometrica: Journal of the Econometric Society*, pp. 665–681, 1977.

Grim, P. and Singer, D. Computational philosophy. In Zalta, E. N. (ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, spring 2020 edition, 2020.

Hooker, J. N. and Kim, T. W. N. Toward non-intuition-based machine and artificial intelligence ethics: A deontological approach based on modal logic. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 130–136, 2018.

Kant, I. and Paton, H. J. *Groundwork of the metaphysic of morals. Translated and analysed by HJ Paton*. Harper & Row, 1964.

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

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

Nair, V. and Hinton, G. E. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pp. 807–814, 2010.

Nitschke, G. Emergence of cooperation: State of the art. *Artificial Life*, 11(3):367–396, 2005.

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

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347, 2017.

Sepielli, A. Moral uncertainty and the principle of equity among moral theories. *Philosophy and Phenomenological Research*, 86(3):580–589, 2013.

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

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

Thagard, P. *Computational philosophy of science*. MIT press, 1993.

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

Wirth, C., Akrour, R., Neumann, G., and Fürnkranz, J. A survey of preference-based reinforcement learning methods. *The Journal of Machine Learning Research*, 18(1): 4945–4990, 2017.

Zap, A., Joppen, T., and Fürnkranz, J. Deep ordinal reinforcement learning. *arXiv preprint arXiv:1905.02005*, 2019.