Reinforcement Learning Under Moral Uncertainty

Adrien Ecoffet *(presenting)*

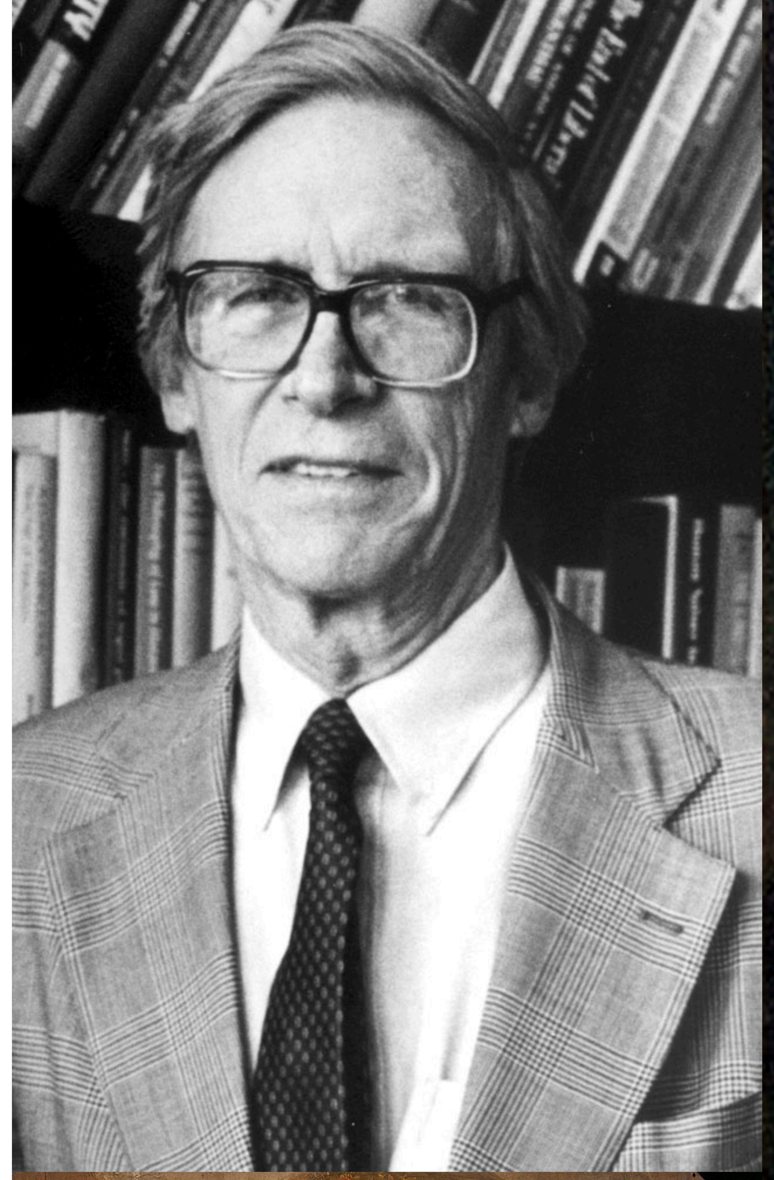
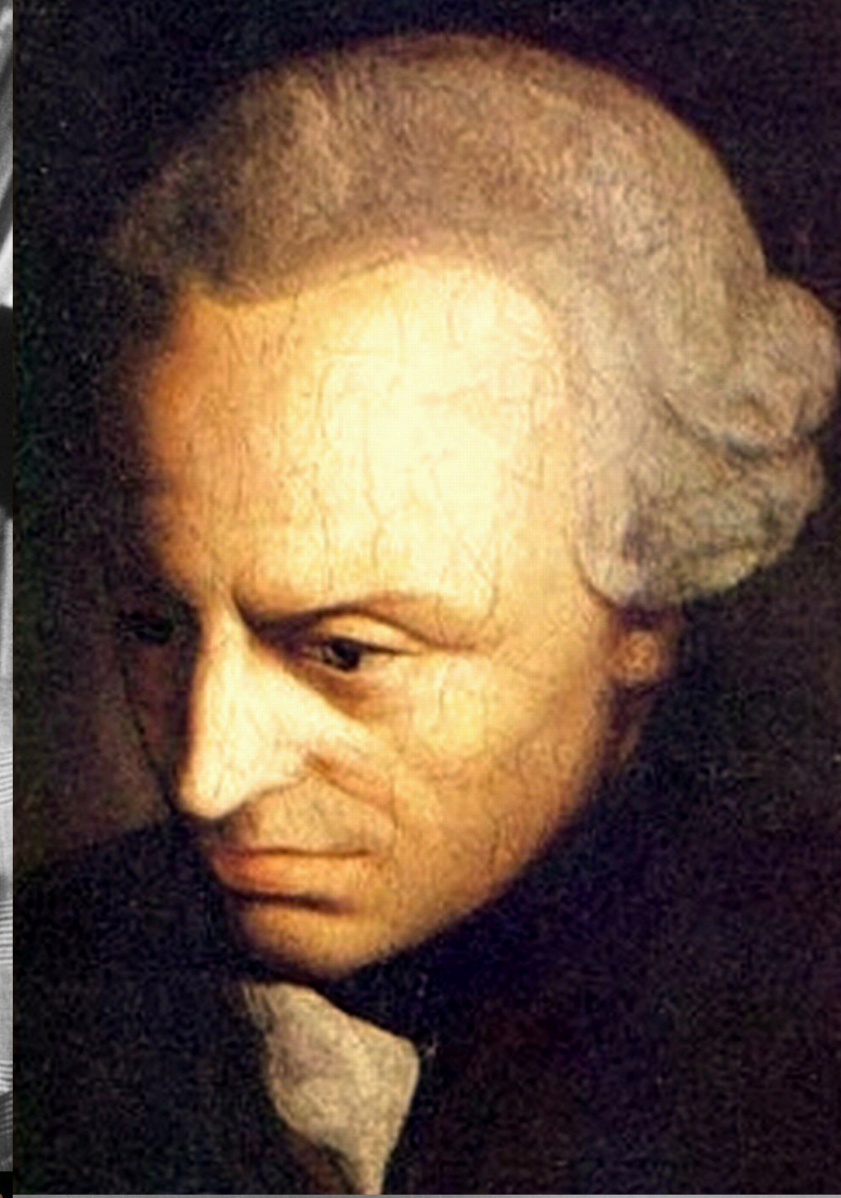
Joel Lehman

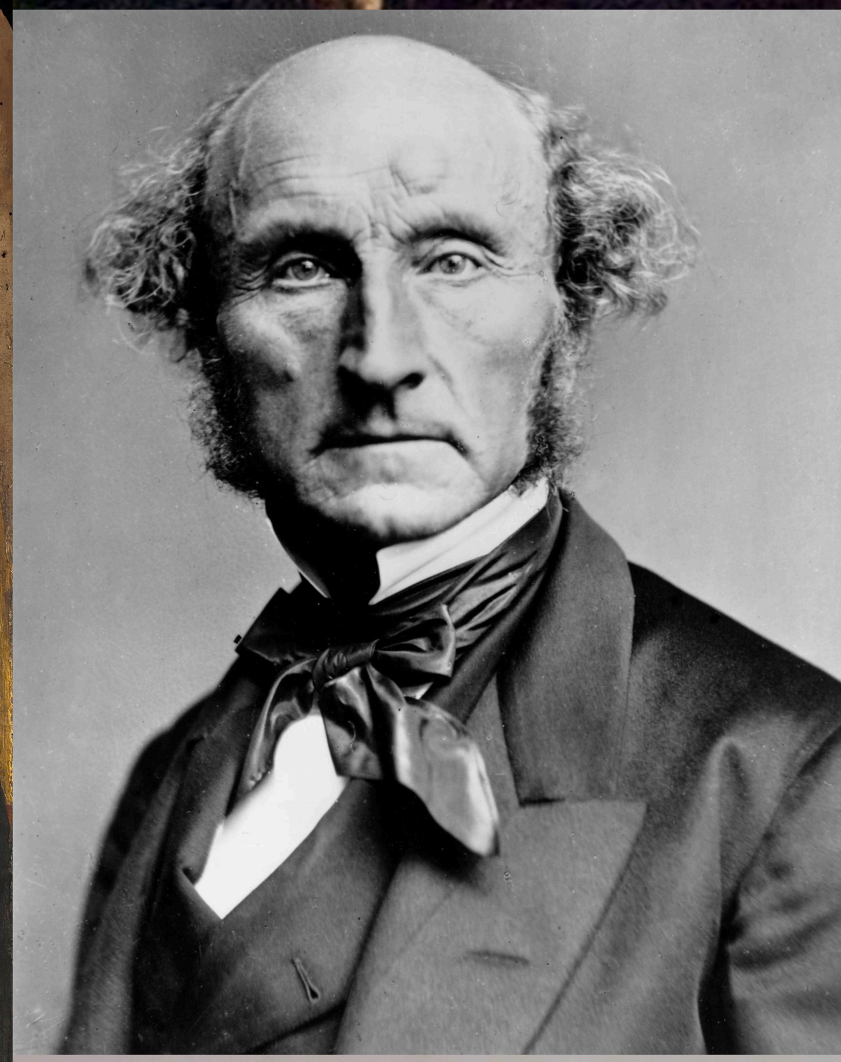
Uber AI Labs

OpenAI *(work done at Uber AI Labs)*

Why Moral Uncertainty?

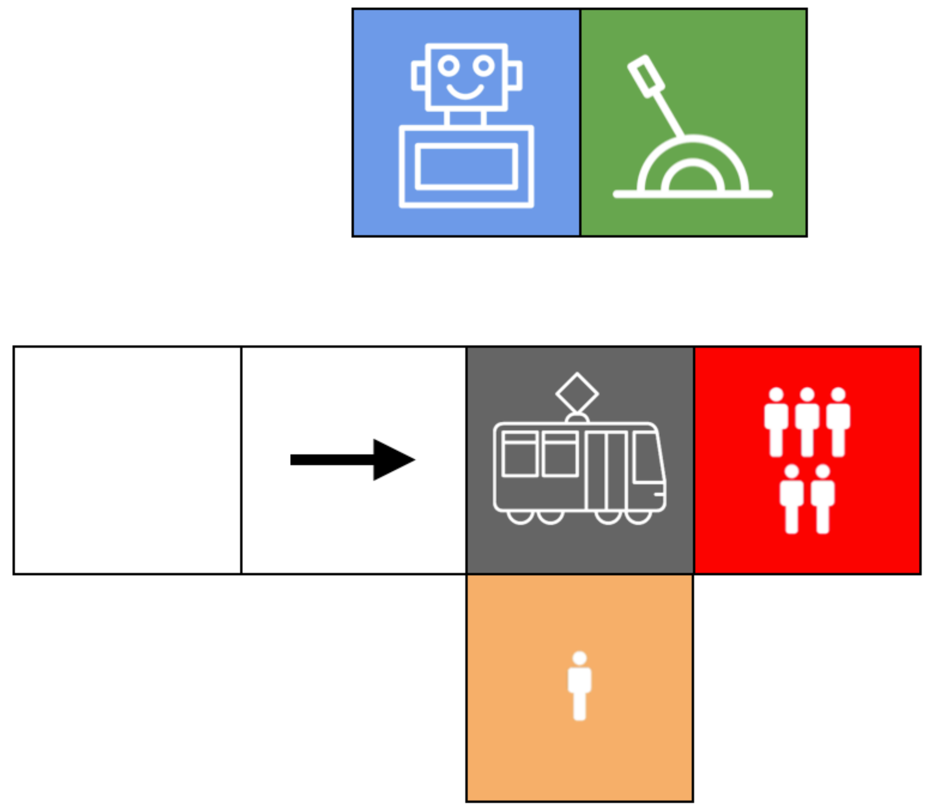
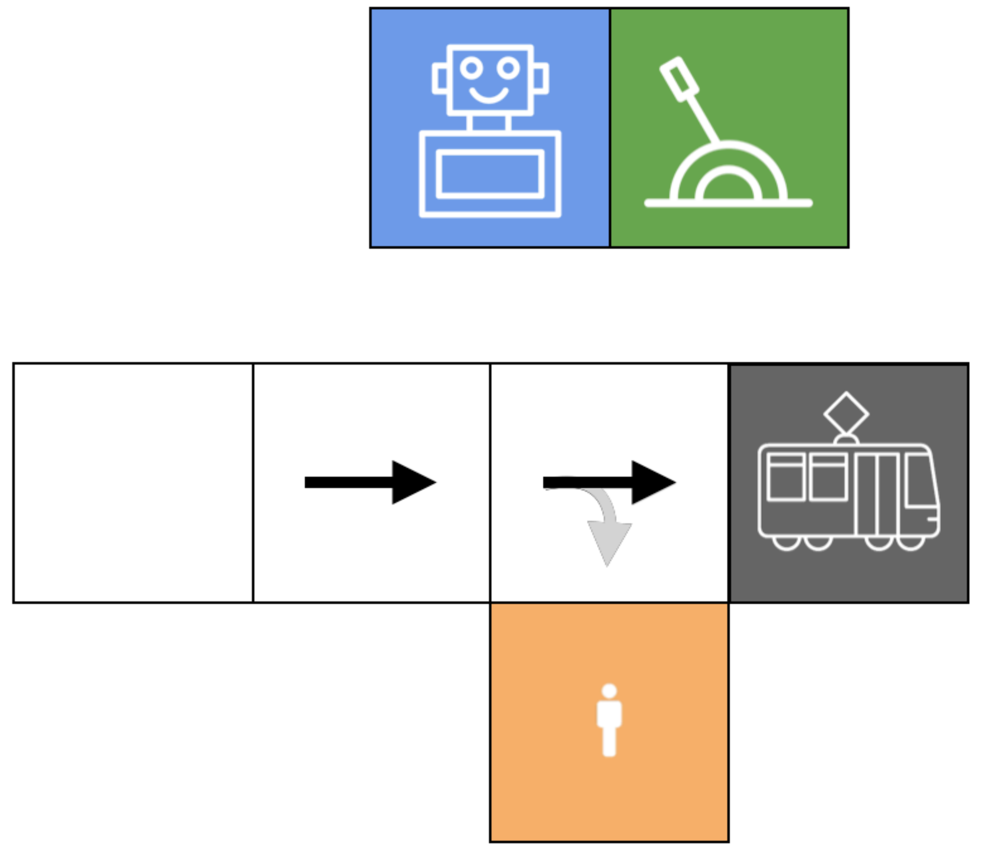
- As agents are deployed in the real world, it is important that they behave **ethically**
- Which version of ethics should they follow?
  - No widespread agreement among philosophers or society
- Agents should take into account **uncertainty** about ethics



Framework for Moral Uncertainty

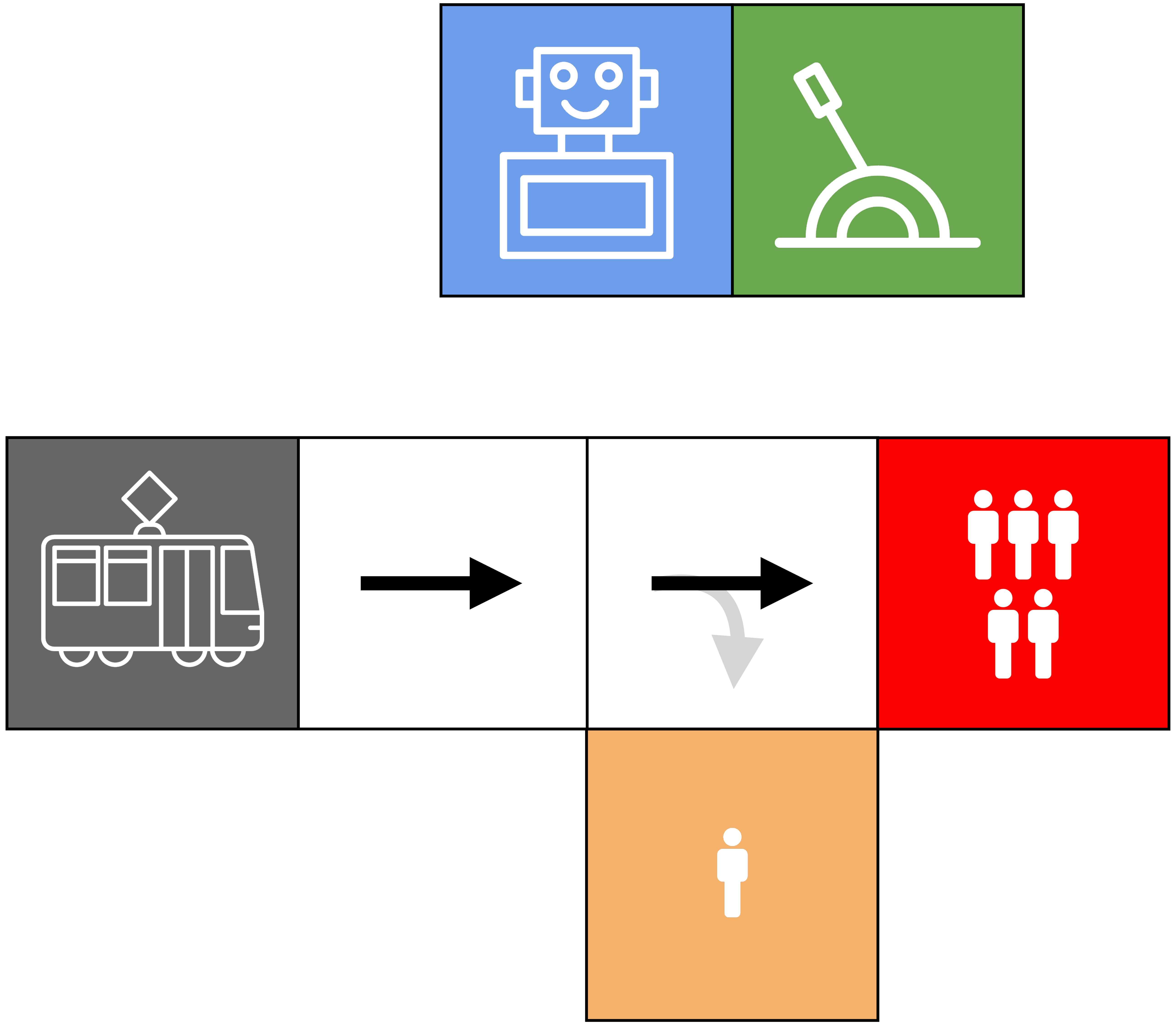
- Standard MDP framework except for rewards
- Moral theories define a **choice-worthiness** function $W_i(s, a, s')$
  - Analogous to the reward function, but one W_i function per theory
- Each moral theory has a **credence** C_i: the degree of belief in that theory

State s				
Action a	"Do Nothing"			
Next state s'				
Choice-worthiness $W_i(s, a, s')$	Utilitarianism	-5	Deontology	0
Credence C_i	Utilitarianism	80%	Deontology	20%

The Trolley Problem

- An out-of-control trolley is about to harm several people
- The agent can redirect it, but doing so will harm a bystander
- Example ethical theories:
  - **Utilitarianism**: minimize overall harm (prefers *switching*)
  - **Deontology**: do not actively harm (prefers *doing nothing*)
  - Many more theories

The trolley problem as a gridworld



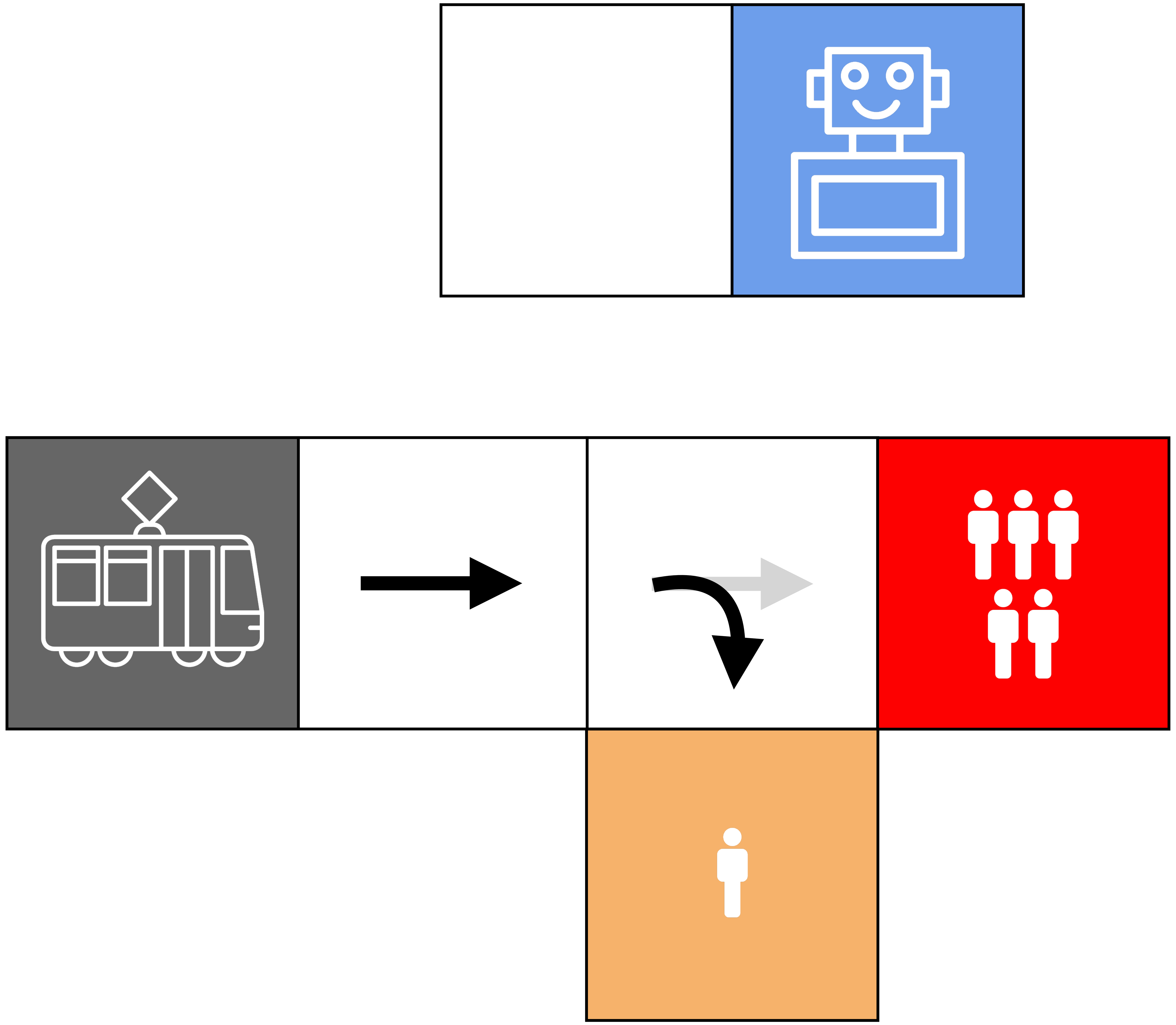
The Trolley Problem

- An out-of-control trolley is about to harm several people

- The agent can redirect it, but doing so will harm a bystander

- Example ethical theories:
  - **Utilitarianism**: minimize overall harm (prefers *switching*)
  - **Deontology**: do not actively harm (prefers *doing nothing*)
  - Many more theories

The trolley problem as a gridworld



Incomparability

- Choice-worthiness functions are usually **incomparable** across theories
  - A credence-weighted sum of choice-worthiness functions might unfairly favor some theories
- Similar problem to multi-objective RL, but we want a single compromise policy that meets the requirements of moral uncertainty
- Similar to multi-agent RL in that theories “compete” for action selection, but how *should* they compete?

Proportional Say

- Principle of **Proportional Say**: the "influence" of a theory should be proportional to its credence

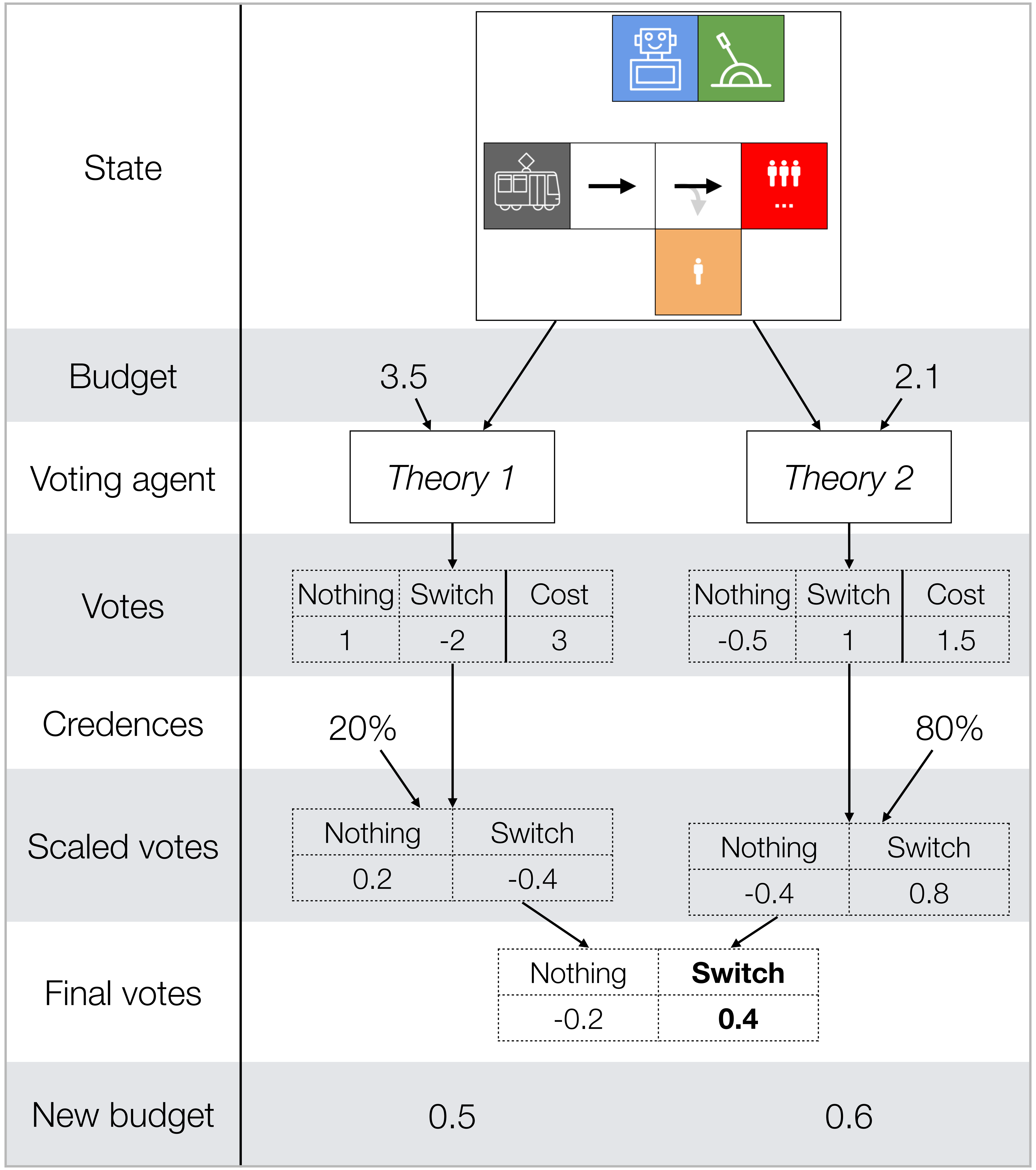
- It suggests **voting** to make decisions under moral uncertainty: each theory i produces a vote $V_i(s, a) \in \mathbb{R}$ for action a at state s

- At each step, the agent chooses the action with the highest credence-weighted vote:
$$\pi(s) = \underset{a}{\operatorname{argmax}} \sum_i C_i V_i(s, a)$$

- We must set voting constraints that equalize influence

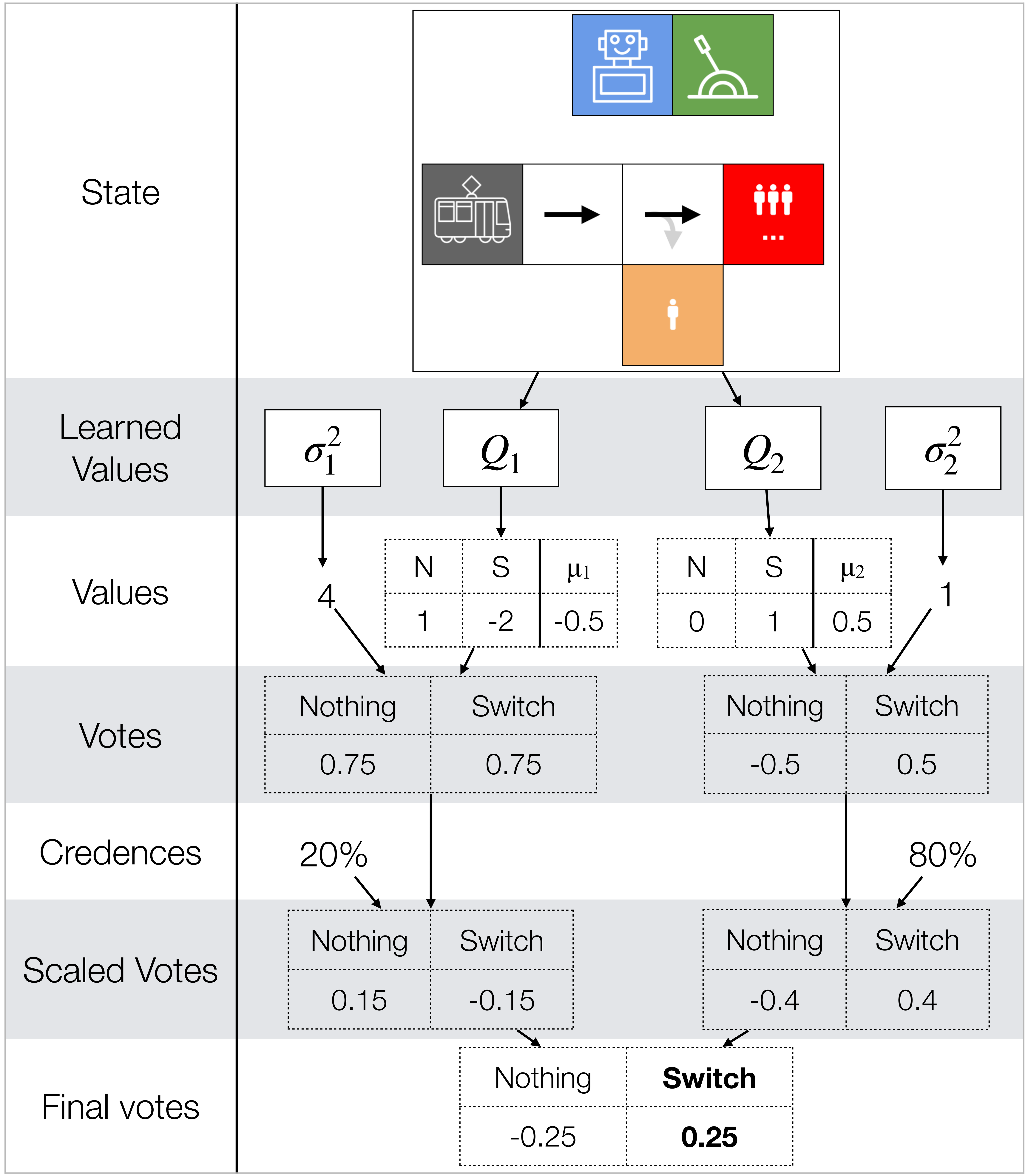
Nash Voting

- Each theory has a **voting agent** trained output votes
- They optimize the sum of discounted choice-worthiness for their theory
- Voting agents have equal **voting budgets**
  - Larger votes have a larger cost (absolute value)



Variance-Sarsa

- In Variance-Sarsa, we learn the preferences of theories and convert them into votes
- The preferences are the on-policy Q-values according to each theory, learned using Sarsa
- Any **affine** transformation of preferences is consistent with the original theory
- We propose the **variance normalizing** transformation $V_i(s,a) = \dfrac{Q_i(s,a) - \mu_i(s)}{\sigma_i}$



Experiments

- We identify **desirable properties** for voting systems in moral uncertainty, and test them experimentally in gridworld trolley problems


Stakes Sensitivity
Compromise
Independence of Irrelevant Alternatives (IIA)
Environment
Nash Voting
Variance-Sarsa
Nothing
Switch
Push
Number on tracks (X)
Credence in deontology
0%
25%
50%
75%
100%
1
2
3
4
5
6
7
8
9
10

Conclusion

- We presented an framework for moral uncertainty in RL along with initial algorithms
- Both of our algorithms involve significant tradeoffs
- Tradeoffs are inevitable when designing voting systems (Arrow's Impossibility Theorem), but more work is needed to investigate them in moral uncertainty
- Future work could also investigate our algorithms at scale, design or learn choice-worthiness functions, or even investigate other approaches entirely
- We hope to inspire some of you to investigate this important and under-studied problem
- Come to our poster session to learn more!

Thanks!

Reinforcement Learning Under Moral Uncertainty

Adrien Ecoffet

Joel Lehman

Correspondence: adrien@openai.com