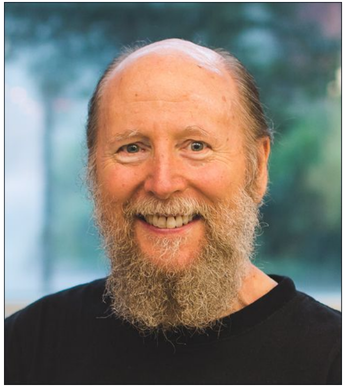
Reinforcement Learning

Reward: The Reward Hypothesis, Inverse RL, Reward Shaping, RLHF

David Abel, Michael Herrmann

7 March, 2025

RL in the News!



ACM A.M. Turing Award Honors Two Researchers Who Led the Development of Cornerstone AI Technology

Andrew Barto and Richard Sutton Recognized as Pioneers of Reinforcement Learning

https://awards.acm.org/about/2024-turing

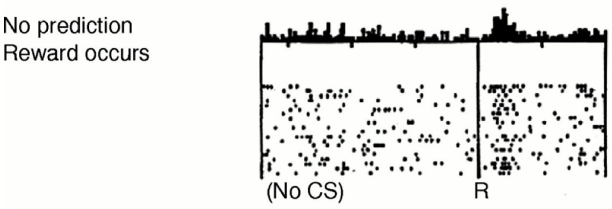
Lecture Overview

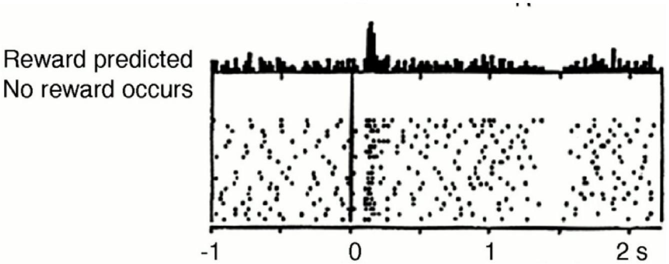
1. Brief note: RL and the Brain
2. Reward Hypothesis
3. Inverse RL
4. Reward Shaping

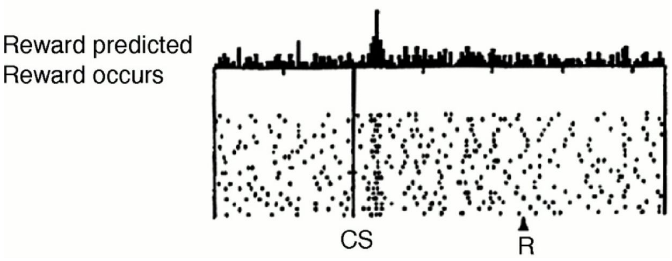
RL and the Brain

A Neural Substrate of Prediction and Reward

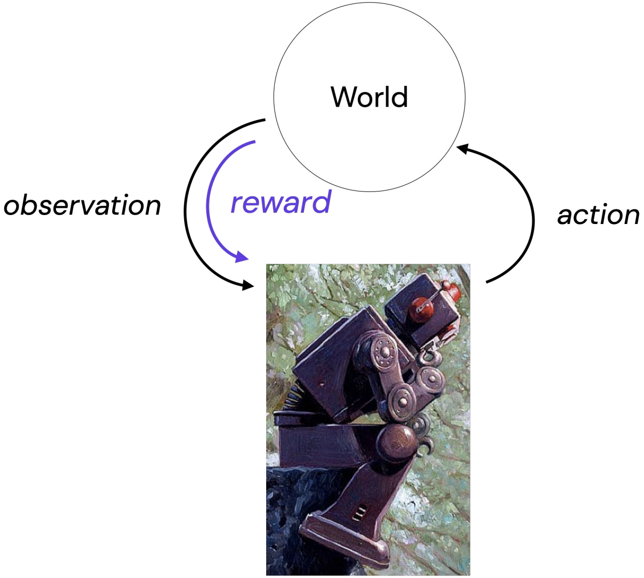
Wolfram Schultz, Peter Dayan, P. Read Montague*







RL and AI



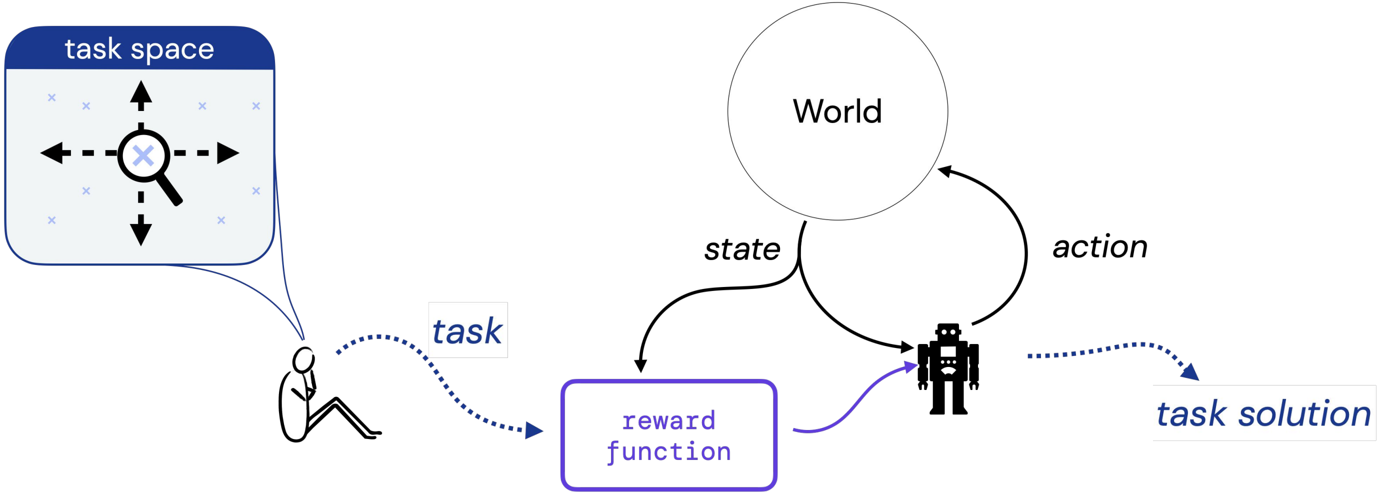
"Part of the appeal of reinforcement learning is that it is in a sense the whole AI problem in a microcosm."
– [Sutton, 1992]

The Reward Hypothesis

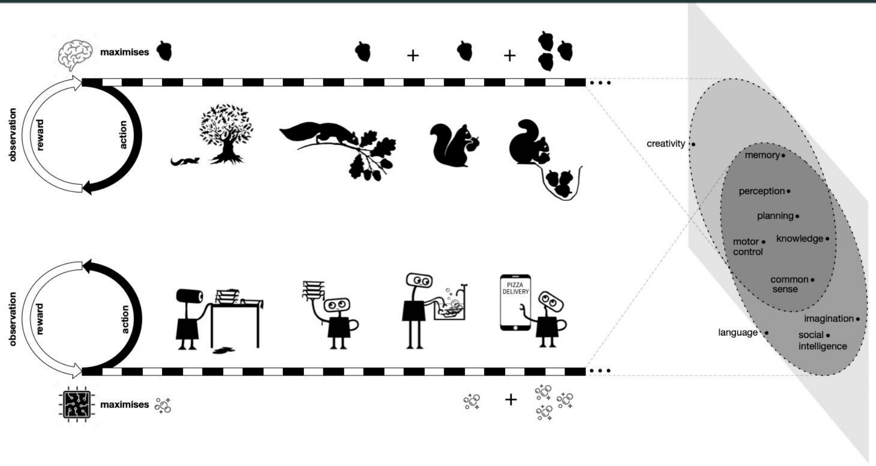
The Reward Hypothesis

"...all of what we mean by goals and purposes can be well thought of as maximization of the expected value of the cumulative sum of a received scalar signal (reward)"

-- *Sutton (2004)*, *Littman (2017)*



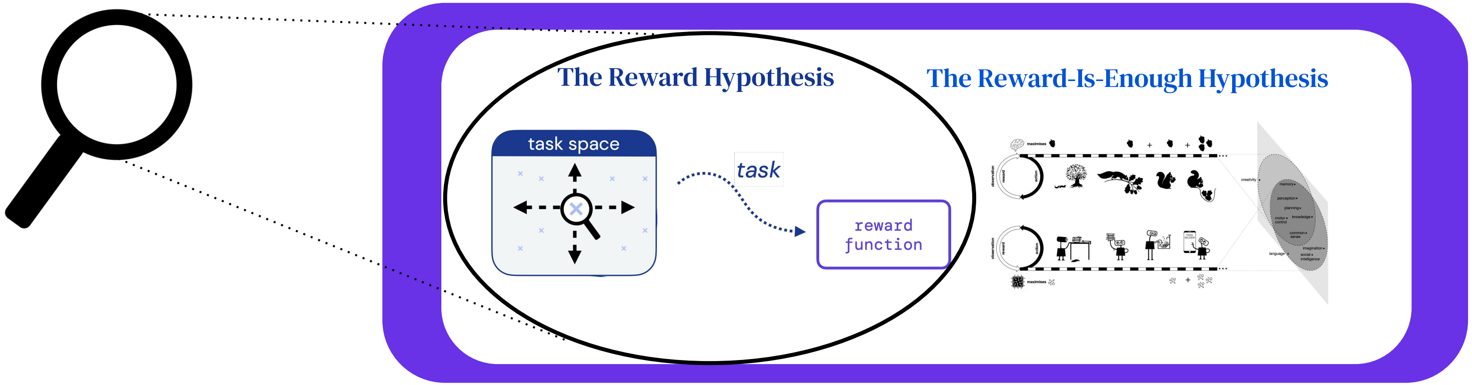
Reward Is Enough



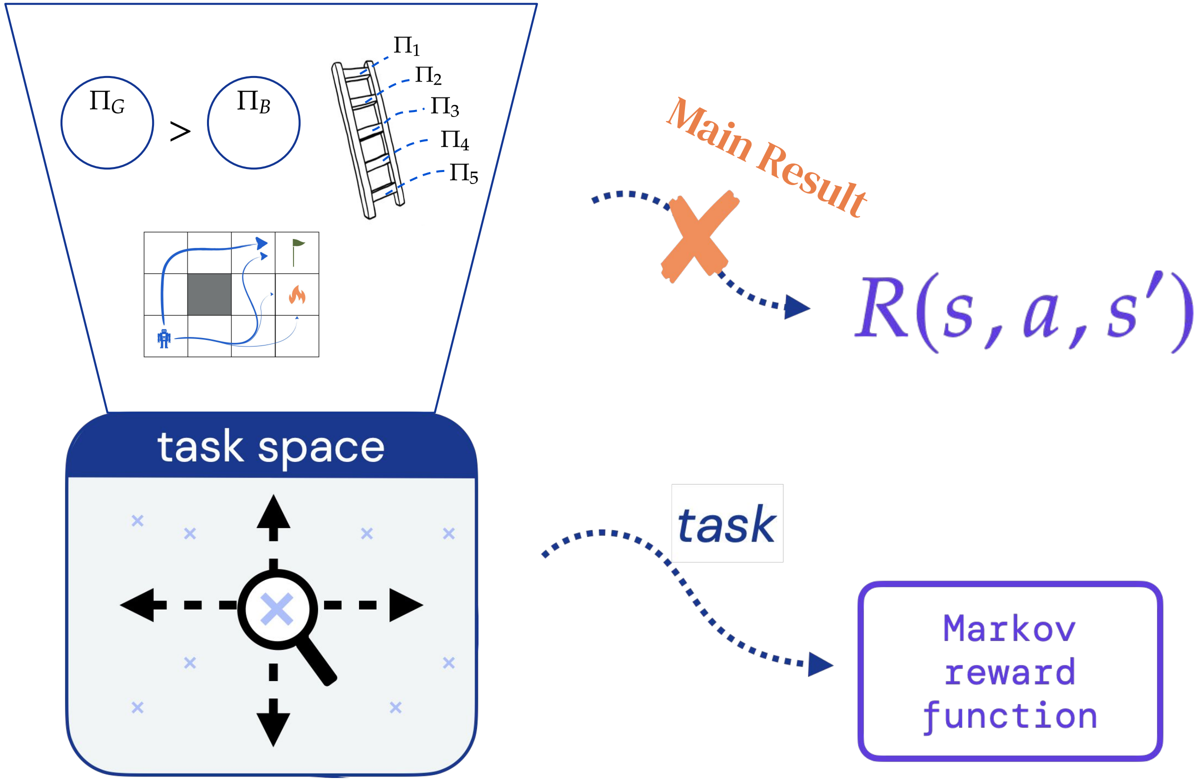
The Reward-Is-Enough Hypothesis

"Intelligence, and its associated abilities, can be understood as subserving the maximisation of reward by an agent acting in its environment"
-- Silver, Singh, Precup, Sutton (2021)

Reward Result 1: Markov Reward *Is* Limited

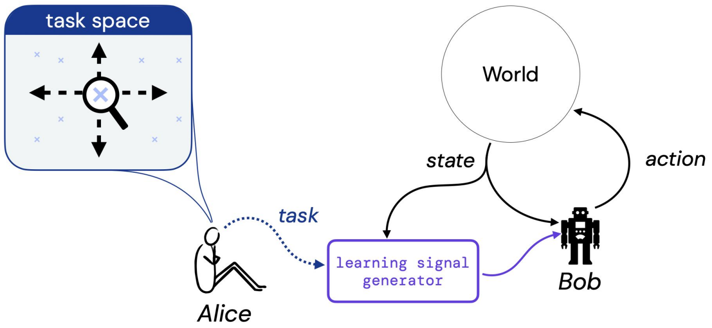


Reward Result 1: Markov Reward *Is* Limited

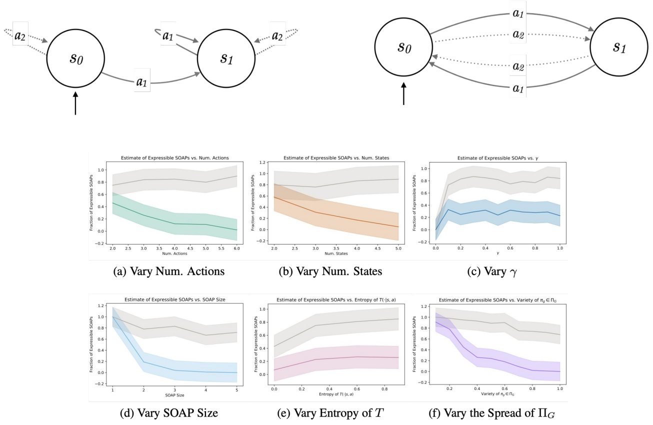


Two Parts

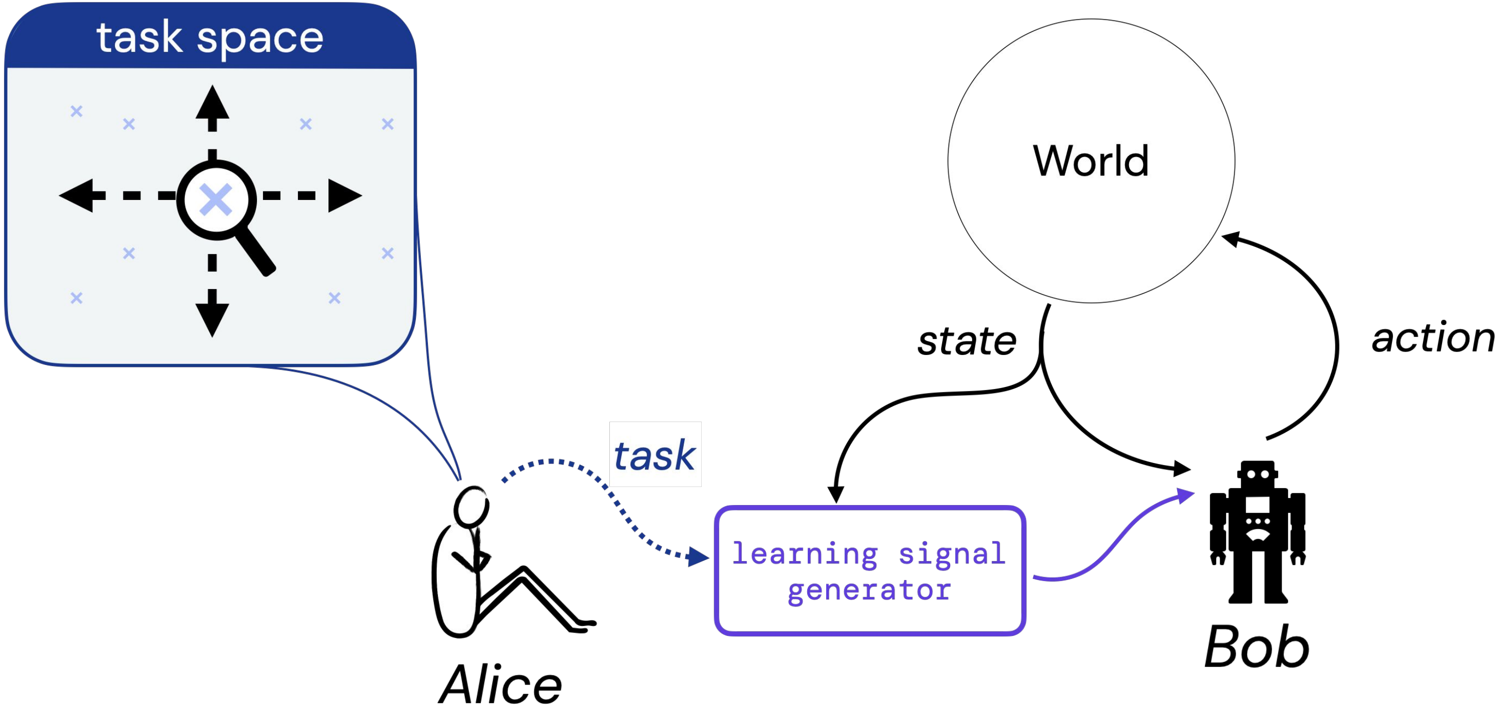
1 **Formalizing the RH in finite MDPs**



2 **Analyzing the RH in finite MDPs**

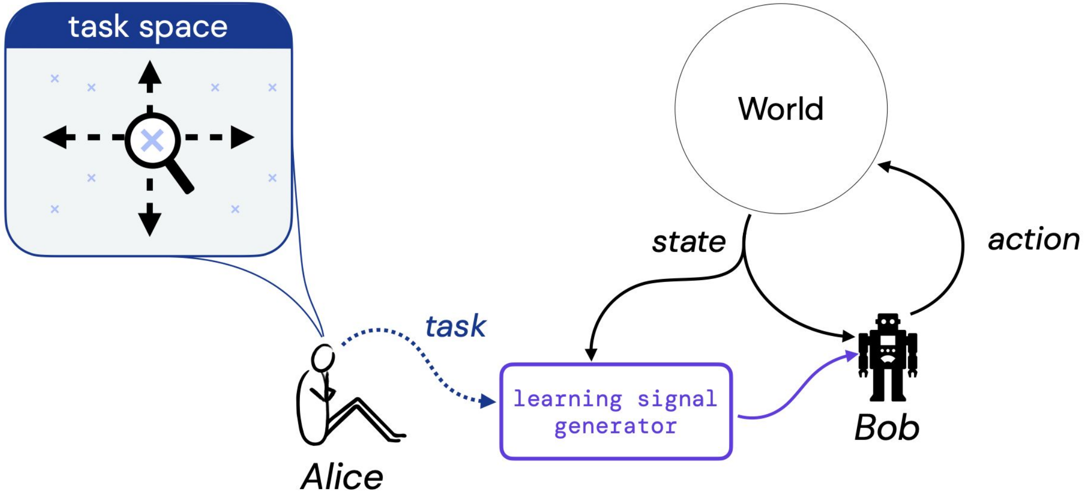


1: Formalising the RH



1: Formalising the RH – Two Questions

Expression Question: Which signal can be used as a mechanism for expressing a given task?



1: Formalising the RH – Two Questions

Expression Question: Which signal can be used as a mechanism for expressing a given task?

$\mathscr{T} = ?$ $\Longrightarrow$

The Reward Hypothesis (formalized)

Given *any* task $\mathscr{T}$ and *any* environment E there is a reward function that realizes $\mathscr{T}$ in E.

Task Question: What is a task?

1: Formalising the RH – Two Questions

Expression Question: Which signal can be used as a mechanism for expressing a given task?

$\mathscr{T} = ?$ $\Longrightarrow$

The Reward Hypothesis (formalized)

Given *any* task $\mathscr{T}$ and *any* environment E there is a reward function that realizes $\mathscr{T}$ in E.

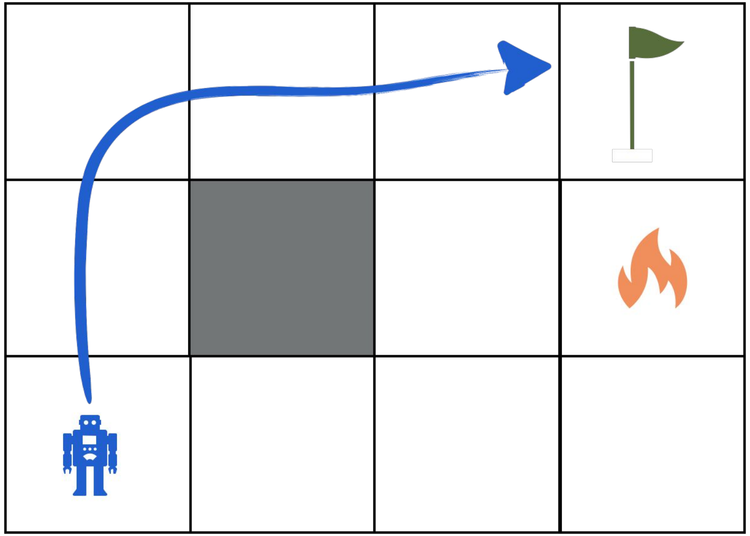
Task Question: What is a task?

Assumption. All environments are finite Controlled Markov Processes (CMPs).

$$E = (\mathcal{S}, \mathcal{A}, T, \gamma, s_0)$$

$R(s), R(s,a), R(s,a,s'), R(s')$

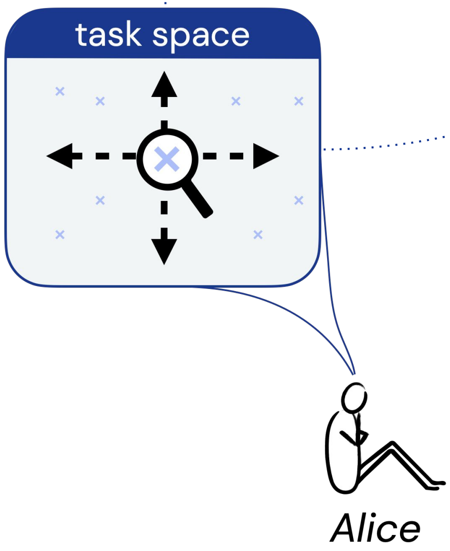
What is a Task?



Tasks: SOAPs, POs, TOs

Task Question: What is a task?

Set of Acceptable Policies (SOAP) | Policy Ordering (PO) | Trajectory Ordering (TO)



Tasks: SOAPs, POs, TOs

Task Question: What is a task?

Set of Acceptable Policies (SOAP)

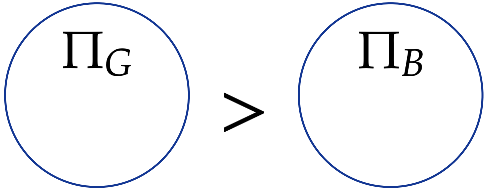
$$\Pi_G \subseteq \Pi$$

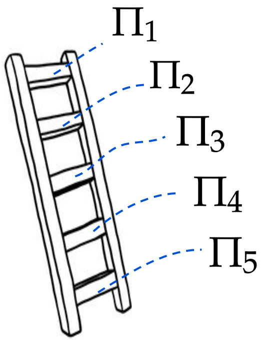
Policy Ordering (PO)

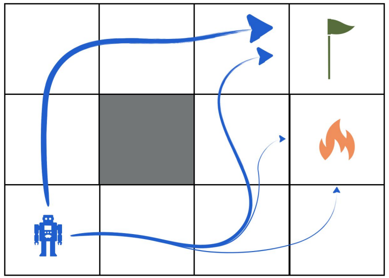
$$L_\Pi$$

Trajectory Ordering (TO)

$$L_{\tau,N}$$







Example:
"Reach the goal in less than 10 steps in expectation."

Example:
"I prefer you reach the goal in 5 steps, else within 10, else don't bother."

Example:
I prefer safely reaching the goal and avoid lava at all costs.

Task Realization

$$\mathcal{T} \in \{\Pi_G, L_\Pi, L_{\tau,N}\}$$

SOAP *PO* *TO*

$\Longrightarrow$

The RH: Given *any* task $\mathcal{T}$ and *any* environment E there is a reward function that <u>realizes</u> $\mathcal{T}$ in E.

MDP

$$E = (\mathcal{S}, \mathcal{A}, T, \gamma, s_0)$$

R

Markov

Task

$\mathcal{T}$

$$V : \Pi \times \mathcal{S} \rightarrow \mathbb{R}$$

Task Realization

$$\mathscr{T} \in \{\Pi_G, L_\Pi, L_{\tau,N}\}$$

SOAP *PO* *TO*

$$\implies$$

The RH: Given *any* task $\mathscr{T}$ and *any* environment E there is a reward function that <u>realizes</u> $\mathscr{T}$ in E.

Set of Acceptable Policies (SOAP)

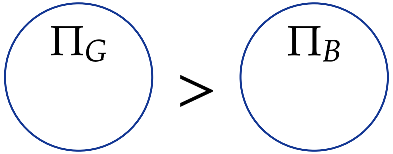
$$\Pi_G \subseteq \Pi$$

Policy Ordering (PO)

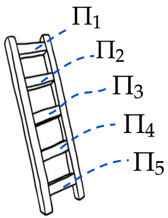
$$L_\Pi$$

Trajectory Ordering (TO)

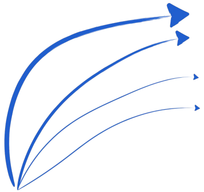
$$L_{\tau,N}$$



$$V^{\pi_g}(s_0) > V^{\pi_b}(s_0)$$

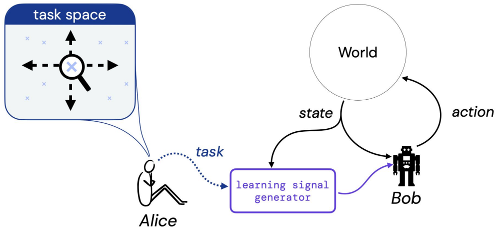


$$V^{\pi_1}(s_0) > V^{\pi_2}(s_0) \ldots$$



$$G(\tau_1; s_0) > G(\tau_2; s_0) \ldots$$

Recap



Which signal can be used to express any task? → **The RH:** Reward

What is a task? → $\mathscr{T} \in \{\Pi_G, L_\Pi, L_{\tau,N}\}$

SOAP *PO* *TO*

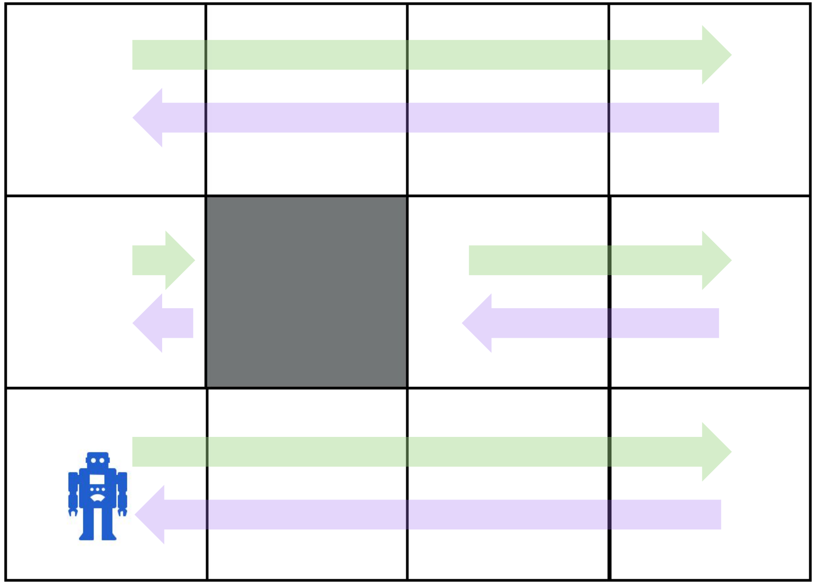
MAIN QUESTION

Given *any* task $\mathscr{T}$ and *any* environment $E = (\mathcal{S}, \mathcal{A}, T, \gamma, s_0)$, is there a Markov reward function that <u>realizes</u> $\mathscr{T}$ in E?

Reward Result 1: Markov Reward is Limited

Theorem 1. *For each of SOAP, PO, and TO, there exist* $(E, \mathcal{T})$ *pairs for which no reward function realizes* $\mathcal{T}$ *in* E.

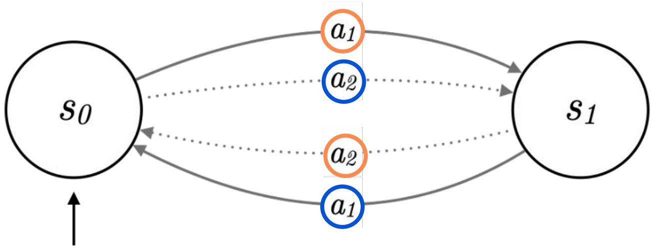
Example 1: What Kind of SOAPs Cannot Be Expressed?



$$\Pi_G = \{\pi_{\leftarrow}, \pi_{\rightarrow}, \ldots\}$$

SOAP = “Always go in the same direction”

Example 2: What Kind of SOAPs Cannot Be Expressed?



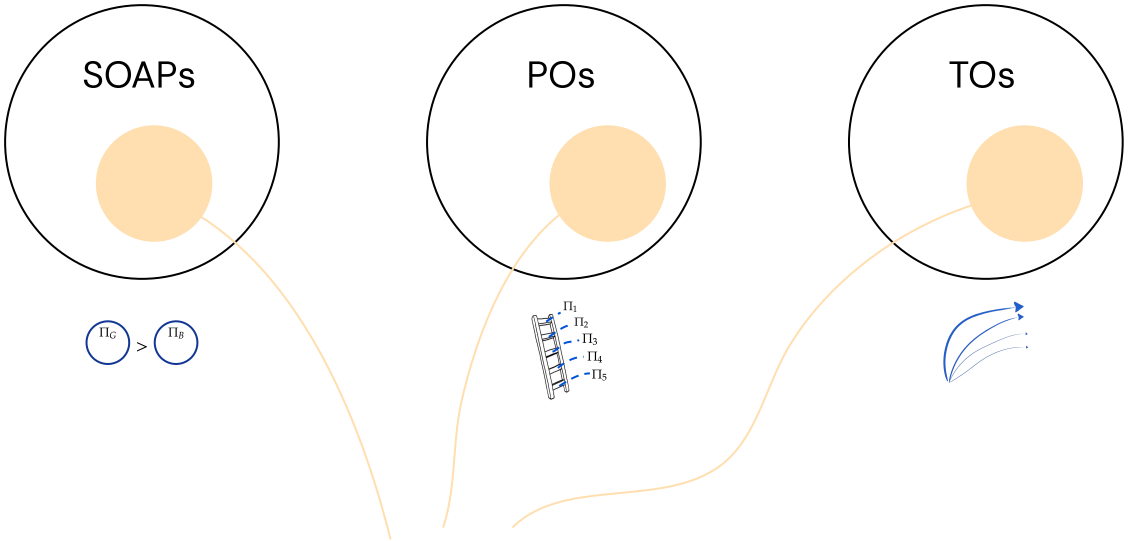
$$\Pi_G = \{\boxed{\pi_{21}}, \boxed{\pi_{12}}\}$$

XOR Problem

...Other types?

Question: Can We Find The Realizing Rewards?

Definition 1. *The* RewardDesign *problem is:* ***Given*** $E = (\mathcal{S}, \mathcal{A}, T, \gamma, s_0)$, *and a* $\mathscr{T}$, ***output*** *a reward function* R_{alice} *that ensures* $\mathscr{T}$ *is realized in* $M = (E, R_{alice})$.



Reward can express

Answer: Yes! PolyTime Reward Design

Definition 1. *The* REWARDDESIGN *problem is:* ***Given*** $E = (\mathcal{S}, \mathcal{A}, T, \gamma, s_0)$*, and a* $\mathcal{T}$*,* ***output*** *a reward function* R_{alice} *that ensures* $\mathcal{T}$ *is realized in* $M = (E, R_{alice})$*.*

Given:

Environment

$$E = (\mathcal{S}, \mathcal{A}, T, \gamma, s_0)$$

Task

$$\mathcal{T}$$

$$\Longrightarrow$$

Output:

$$R$$

Or, $\perp$, if no R exists.

Theorem 2. *The* REWARDDESIGN *problem can be solved in polynomial time, for any finite* E*, and any* $\mathcal{T}$*.*

Corollary 1. *Given* $\mathcal{T}$ *and* E*, deciding whether* $\mathcal{T}$ *is expressible in* E *is solvable in polynomial time for any finite* E*.*

Answer: Yes! PolyTime Reward Design

Algorithm 1 SOAP Reward Design

INPUT: $E = (\mathcal{S}, \mathcal{A}, T, \gamma, s_0), \Pi_G$.
OUTPUT: R, or $\perp$.

$\Pi_{\text{fringe}} = \texttt{compute\_fringe}(\Pi_G)$
for $\pi_{g,i} \in \Pi_G$ **do** ▷ Compute state-visitation distributions.
    $\rho_{g,i} = \texttt{compute\_exp\_visit}(\pi_{g,i}, E)$

for $\pi_{f,i} \in \Pi_{\texttt{fringe}}$ **do**
    $\rho_{f,i} = \texttt{compute\_exp\_visit}(\pi_{f,i}, E)$

$C_{\text{eq}} = \{\}$ ▷ Make Equality Constraints.
for $\pi_{g,i} \in \Pi_G$ **do**
    $C_{\text{eq}}.\texttt{add}(\rho_{g,0}(s_0) \cdot X = \rho_{g,i}(s_0) \cdot X)$

$C_{\text{ineq}} = \{\}$ ▷ Make Inequality Constraints.
for $\pi_{f,j} \in \Pi_{\text{fringe}}$ **do**
    $C_{\text{ineq}}.\texttt{add}(\rho_{f,j}(s_0) \cdot X + \epsilon \leq \rho_{g,0}(s_0) \cdot X)$

$R_{\text{out}}, \epsilon_{\text{out}} = \texttt{linear\_programming}(\text{obj.} = \max \epsilon, \text{constraints} = C_{\text{ineq}}, C_{\text{eq}})$ ▷ Solve LP.

if $\epsilon_{\text{out}} > 0$ **then** ▷ Check if successful.
    return R_{out}
else
    return $\perp$

Recap: Expressivity of Markov Reward

MAIN QUESTION

Given *any* task $\mathcal{T}$ and *any* environment $E = (\mathcal{S}, \mathcal{A}, T, \gamma, s_0)$, is there a Markov reward function that realizes $\mathcal{T}$ in E?

Theorem 1. *For each of SOAP, PO, and TO, there exist $(E, \mathcal{T})$ pairs for which no reward function realizes $\mathcal{T}$ in E.*

Theorem 2. *The RewardDesign problem can be solved in polynomial time, for any finite E, and any $\mathcal{T}$.*

Recap: Expressivity of Markov Reward

On the Expressivity of Markov Reward







David Abel | Will Dabney | Anna Harutyunyan | Mark K. Ho | Michael L. Littman | Doina Precup | Satinder Singh

GDM

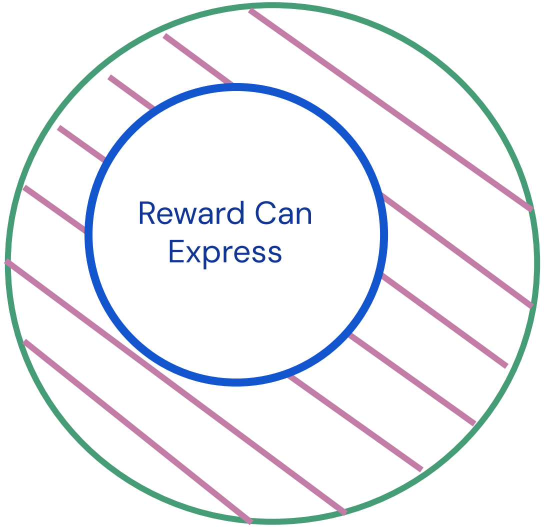
New York University

Brown University

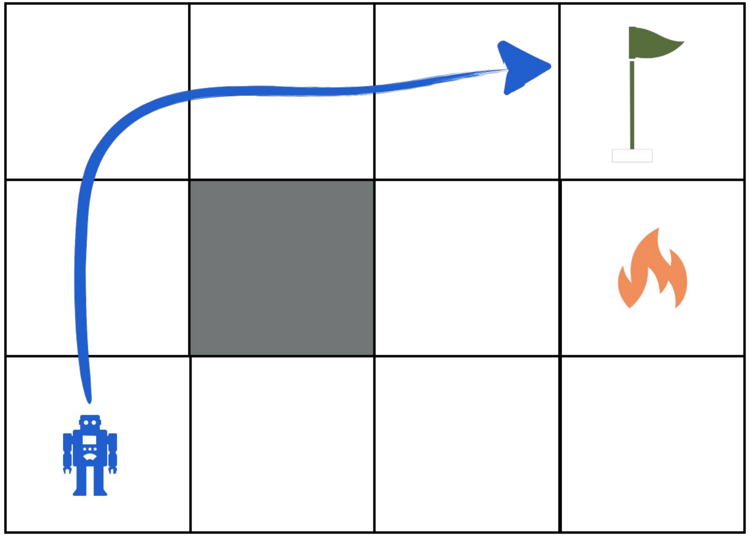
Curious?

Settling the Reward Hypothesis by Bowling et al.

Axiom 1	Completeness
Axiom 2	Transitivity
Axiom 3	Independence
Axiom 4	Continuity
Axiom 5	**Temporal γ-Indifference**



Inverse Reinforcement Learning



Main Q: If we can observe what an agent *does*, can we infer what reward function it maximizes?

Inverse Reinforcement Learning

Definition: Inverse RL Problem.

Given: An environment and behavior

*Output: A reward function that **explains** the behavior*

Main Q: If we can observe what an agent *does*, can we infer what reward function it maximizes?

Inverse Reinforcement Learning

Definition: Inverse RL Problem.

Given: A controlled Markov process, $(\mathcal{S}, \mathcal{A}, p)$, *and behavior*

Output: A reward function that ***explains*** *the behavior*

Main Q: If we can observe what an agent *does*, can we infer what reward function it maximizes?

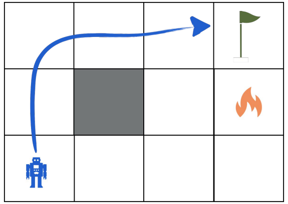
Inverse Reinforcement Learning

Definition: Inverse RL Problem.

Given: A controlled Markov process, $(\mathcal{S}, \mathcal{A}, p)$, *and policy* π

Output: A ***reward function*** *that* ***explains*** *the policy:*

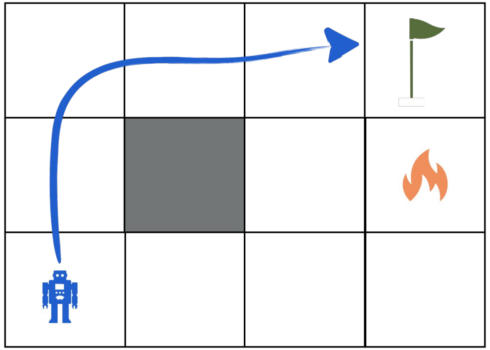
$$\pi = \arg\max_{\pi' \in \Pi} v_r^{\pi'}(s_0)$$



Discussion (2 minutes)

There is a fundamental limitation to Inverse RL. Can you spot it?

Unidentifiability in Inverse Reinforcement Learning



$$\pi = \arg\max_{\pi' \in \Pi} v_r^{\pi'}(s_0)$$

Every policy is optimal w.r.t. the zero reward function! (and constant…)

Discussion (2 minutes)

There is a fundamental limitation to Inverse RL. Can you spot it?

Solution 1 to Unidentifiability

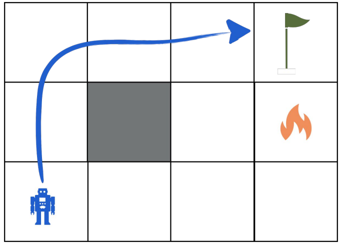
$$\pi = \arg\max_{\pi' \in \Pi} v_{\pi}^{r}(s_0) + \boxed{\omega(r)}$$

Add a regulariser:

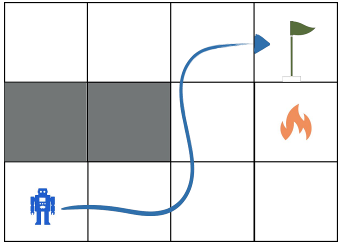
Simple rewards
Complex rewards
Interesting rewards
...

Popular approach:
MaxEnt by Ziebert and Bagnell (2008)

Solution 2 to Unidentifiability



$$\pi = \arg\max_{\pi' \in \Pi} v_\pi^r(s_0)$$



Intervention!

Repeated Inverse RL by Amin et al. (2017)

Reward Shaping

$r(s)$

+1 Win

-1 Lose

0 otherwise

Breadcrumbs…

+0.1 capture pawn
+0.2 capture bishop
+0.5 capture queen
…

Reward Shaping

$r(s)$

+1 Win

-1 Lose

0 otherwise

~~Breadcrumbs…~~

~~+0.1 capture pawn~~
~~+0.2 capture bishop~~
~~+0.5 capture queen~~
~~…~~

Problem!
Can change optimal behavior

Solution: Potential-Based Shaping

$$r_{\text{new}}(s, a, s') = \boxed{r(s)} + \boxed{f(s, a, s')}$$

Original **Breadcrumbs**

Policy Invariance Under Reward Transformations: Theory and Application to Reward Shaping
by Ng, Harada and Russell (1999)

Solution: Potential-Based Shaping

$$r_{\text{new}}(s, a, s') = \underbrace{\boxed{r(s)}}_{\text{Original}} + \underbrace{\boxed{f(s, a, s')}}_{\text{Breadcrumbs}}$$

$$\boxed{f(s, a, s') = \gamma\phi(s') - \phi(s)}$$

Potential-Based Shaping Function

Policy Invariance Under Reward Transformations: Theory and Application to Reward Shaping by Ng, Harada and Russell (1999)

Solution: Potential-Based Shaping

Theorem. A shaping function preserves the optimal policy if and only if it is a potential-based shaping function

$$\boxed{f(s, a, s') = \gamma\phi(s') - \phi(s)}$$

Potential-Based Shaping Function

Policy Invariance Under Reward Transformations: Theory and Application to Reward Shaping by Ng, Harada and Russell (1999)

More:

Reward and Safety
Everitt et al. 2017, Ortega et al. 2018, Kumar et al. 2020 Uesato et al. 2020

Reward and Preferences
MacGlashan et al. 2016, Wirth et al. 2017, Christiano et al. 2017, Xu et al. 2020

Reward Learning & Design
Ackley & Littman 1992, Singh et al. 2010, Sorg 2011, Zheng et al. 2020, Jeon et al. 2020

Reward and Constrained MDPs
Mannor & Shimkin 2004, Szepesvári 2020, Roijers et al. 2020, Zahavy et al. 2021

Reward and Teaching
Goldman & Kearns 1995, Simard et al. 2017, Ho et al. 2019

Reward and Logical tasks in RL
Littman et al. 2017, Li et al. 2017, Jothimurugan et al. 2020, Tasse et al. 2020

Expectations, Discount, and Rationality
Mitten 1974, Sobel 1975, Weng 2011, Pitis 2019, Gottipati et al. 2020

Reward and Target Distribution
Akshay et al. 2013, Hafner et al. 2020

CIRL, Assistive Learning
Syed et al. 2008, Hadfield-Menell et al. 2016, Amin et al. 2017, Shah et al. 2020

Natural Language
MacGlashan et al. 2015, Williams et al. 2017

Optional Reading

Reward Hypothesis:

- *On the Expressivity of Markov Reward,* Abel et al. (2021)
- *Settling the Reward Hypothesis,* Bowling et al. (2023)

Inverse RL:

- *Algorithms for inverse reinforcement learning* by Ng and Russell (1999)
- *Maximum Entropy Inverse Reinforcement Learning* by Ziebert and Bagnell (2008)
- *Repeated Inverse Reinforcement Learning* by Amin et al. (2017)

Reward Shaping:

- *Potential-Based Shaping and Q-value Initialization are Equivalent* by Wiewora (2003)
- *Policy Invariance Under Reward Transformations: Theory and Application to Reward Shaping* by Ng, Harada, Russell (1999)