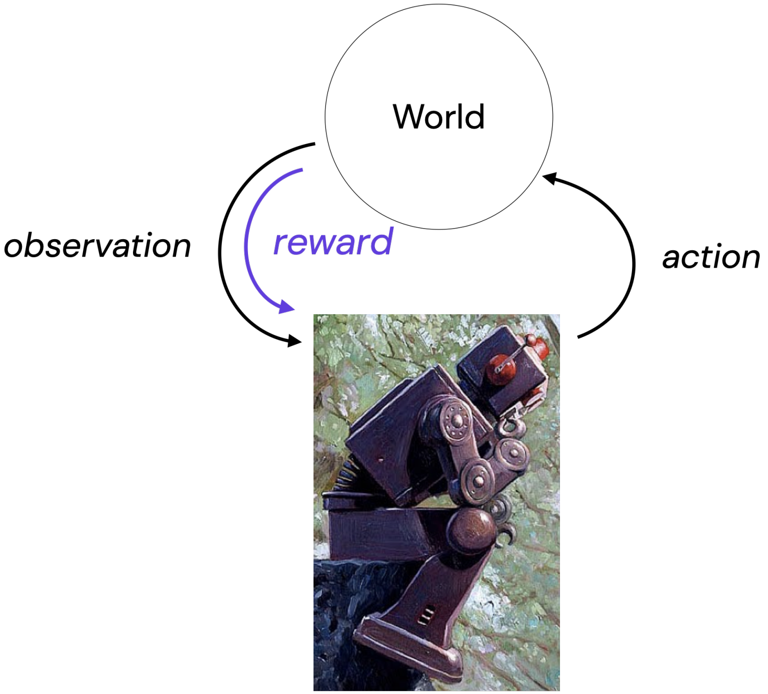
On Tasks and Rewards

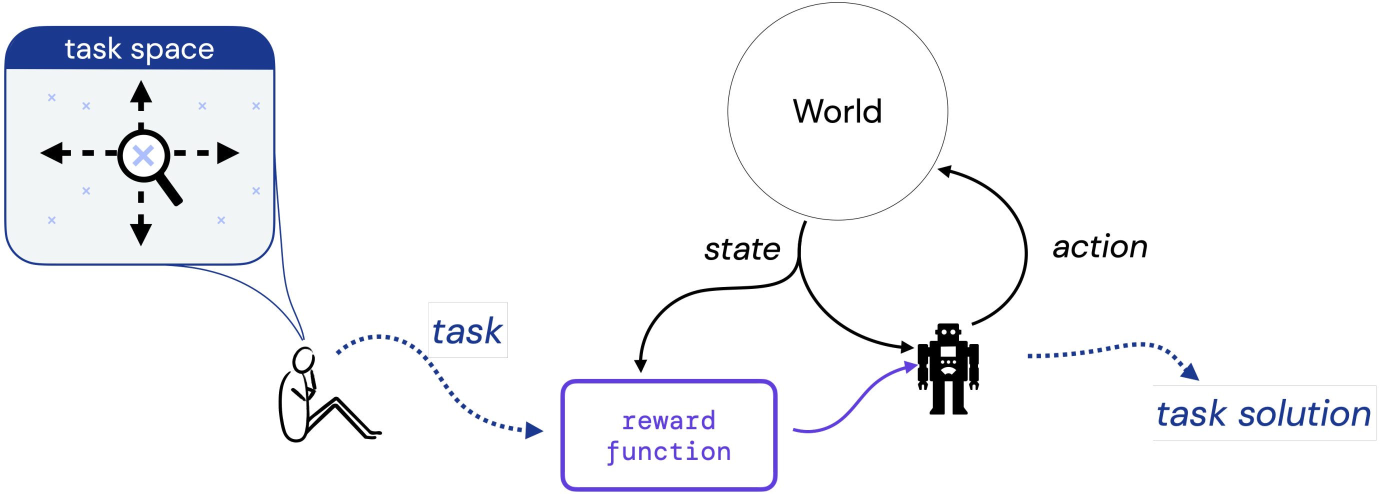
Overview



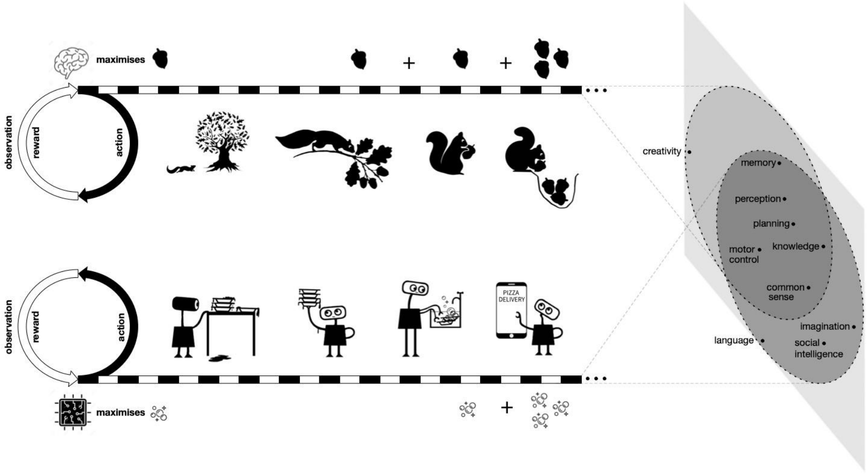
> *"Part of the appeal of reinforcement learning is that it is in a sense the whole AI problem in a microcosm."*
> – [Sutton, 1992]

The Reward Hypothesis

"...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)]()

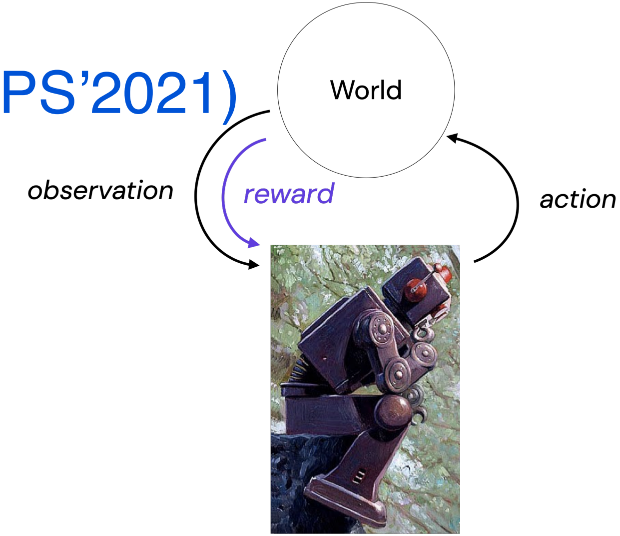


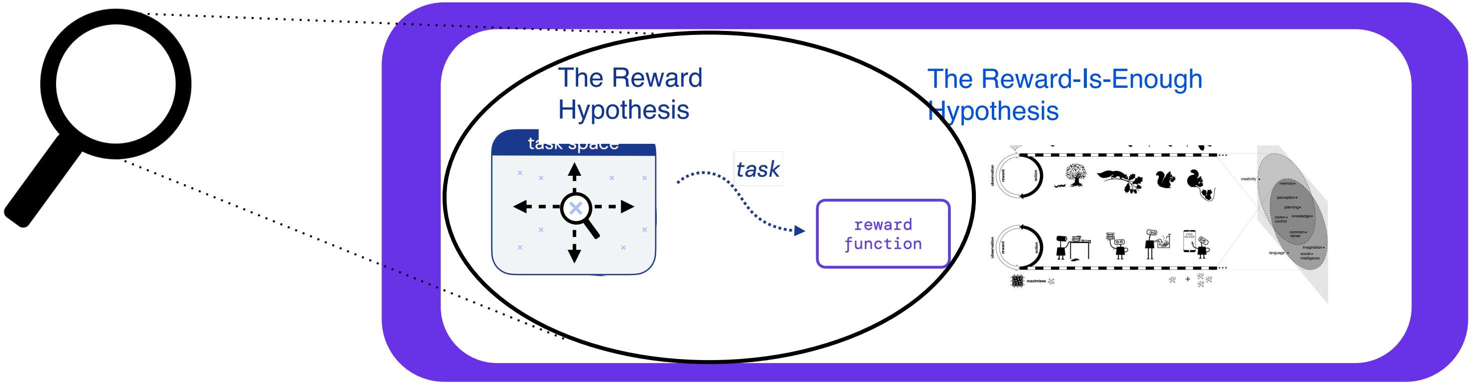
Reward is Enough



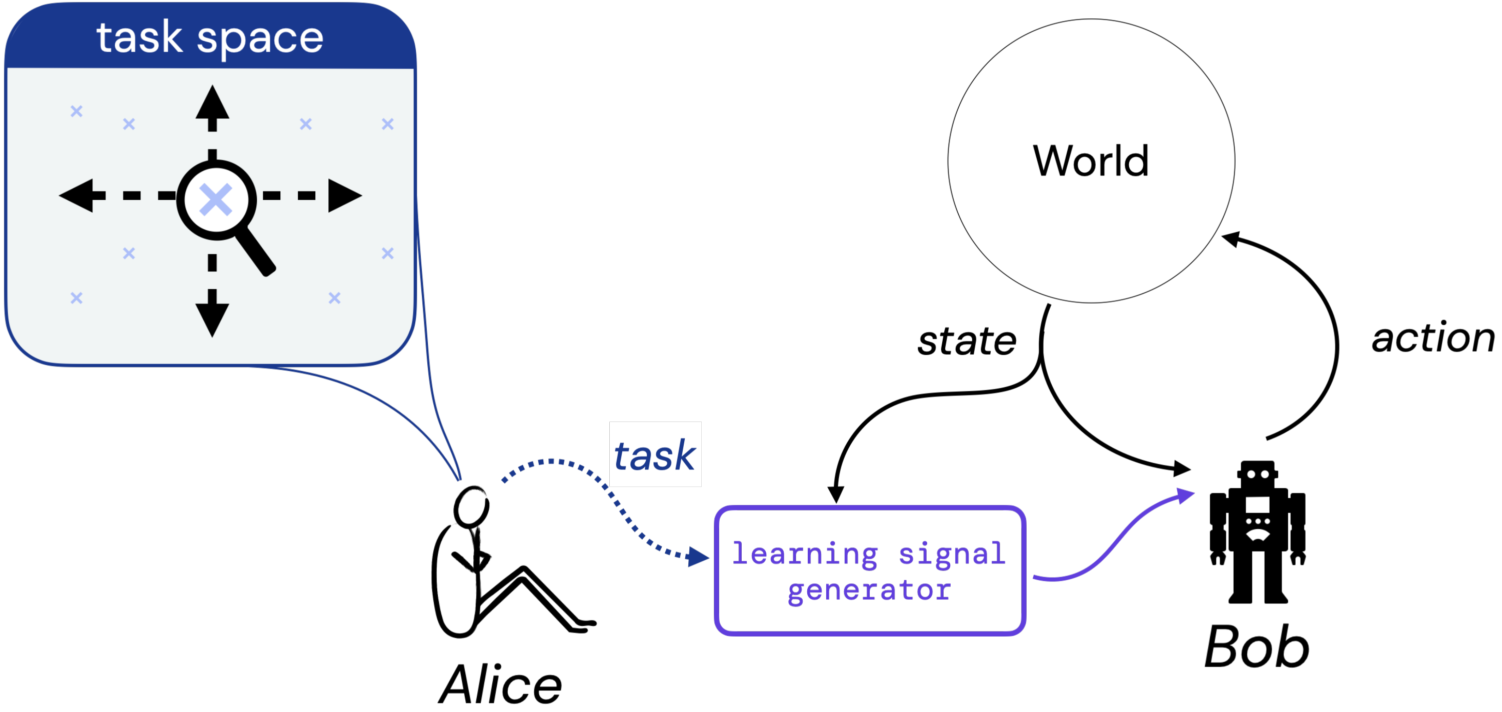
"Intelligence, and its associated abilities, can be understood as subserving the maximisation of reward by an agent acting in its environment"
-- *Silver, Singh, Precup, Sutton (2021)*

On the Expressivity of Markov Reward (NeurIPS'2021)



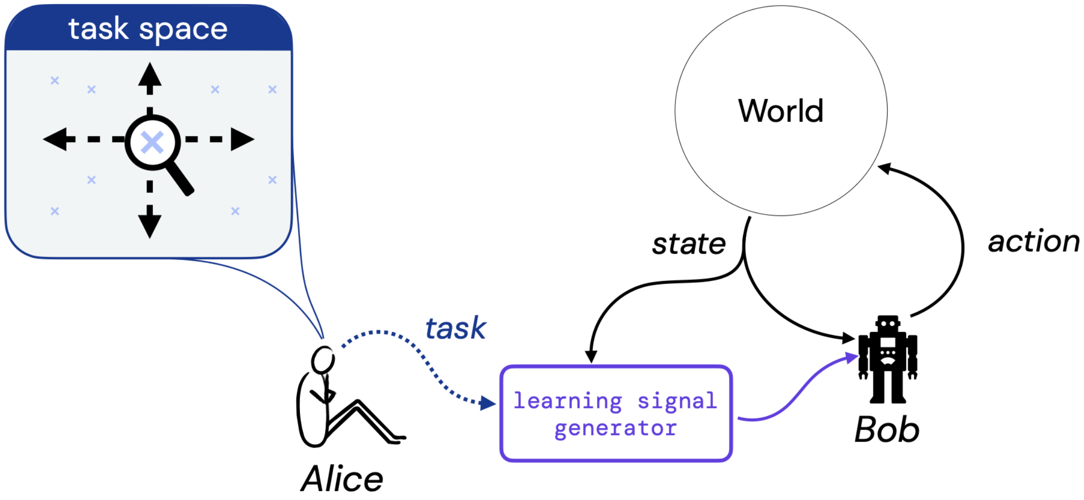


Formalizing the Reward Hypothesis



The Two Question View

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



The Two Question View

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

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

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?

The Two Question View

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

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

The Reward Hypothesis (formalized)

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

Task Question: What is a task?

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

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

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

Related Work: Other Perspectives on Reward

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

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

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

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

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

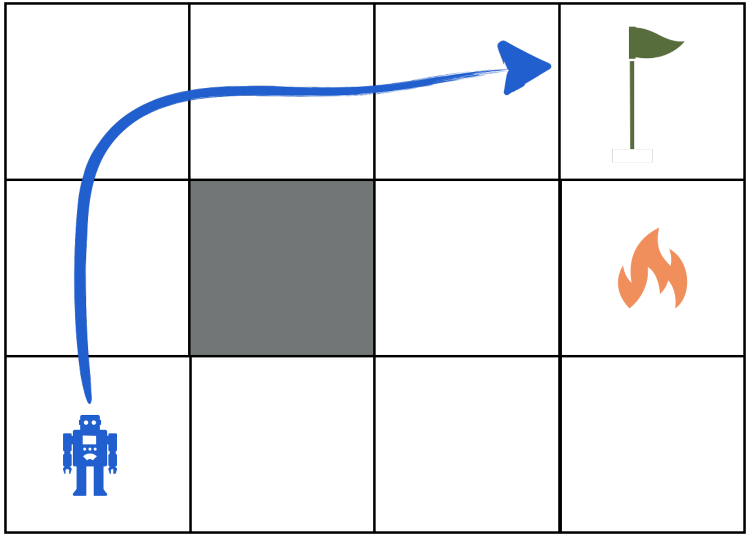
Expectations, Discount, and Rationality
Mitten 1974, Sobel 1975, Weng 2011, Pitis 2019, Gottipati et al. 2020

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

IRL, 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

What is a Task?



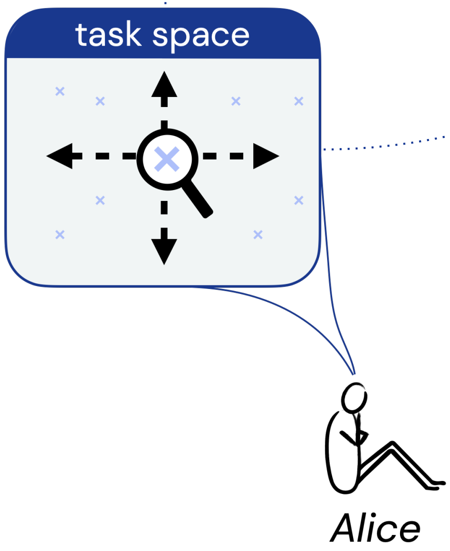
Task Types: SOAPs, POs, TOs

Task Question: What is a task?

Set of Acceptable Policies (SOAP)

Policy Ordering (PO)

Trajectory Ordering (TO)



Task Types: SOAPs, POs, TOs

Task Question: What is a task?

Set of Acceptable Policies (SOAP)

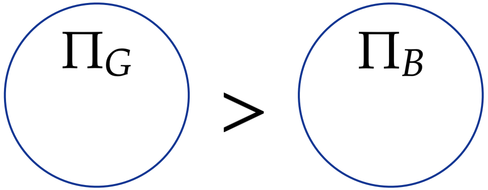
$$\Pi_G \subseteq \Pi$$

Policy Ordering (PO)

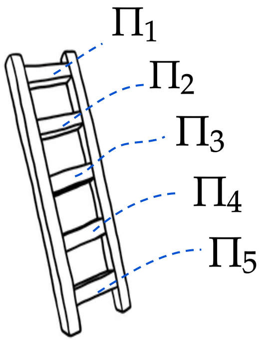
$$L_\Pi$$

Trajectory Ordering (TO)

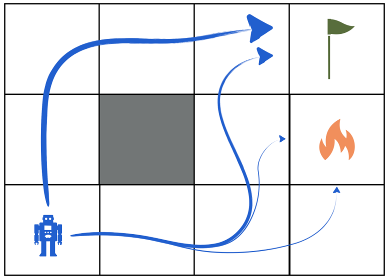
$$L_{\tau,N}$$



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 realizes $\mathcal{T}$ in E.

MD

Task

$\mathcal{T}$

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

R

Markov

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

Task Realization

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

SOAP *PO* *TO*

$\Longrightarrow$

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

Set of Acceptable Policies (SOAP)

$$\Pi_G \subseteq \Pi$$

Policy Ordering (PO)

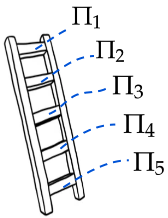
$$L_\Pi$$

Trajectory Ordering (TO)

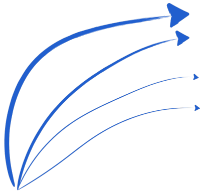
$$L_{\tau,N}$$

$$\Pi_G > \Pi_B$$

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

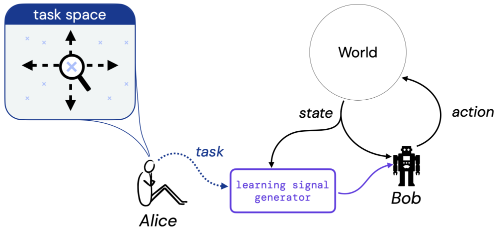


$$V^{\pi_1}(s_0) > V^{\pi_2}(s_0) \ldots$$



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

Recap



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

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

SOAP *PO* *TO*

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 <u>realizes</u> $\mathcal{T}$ in E?

Question 1: What Can Reward Express?

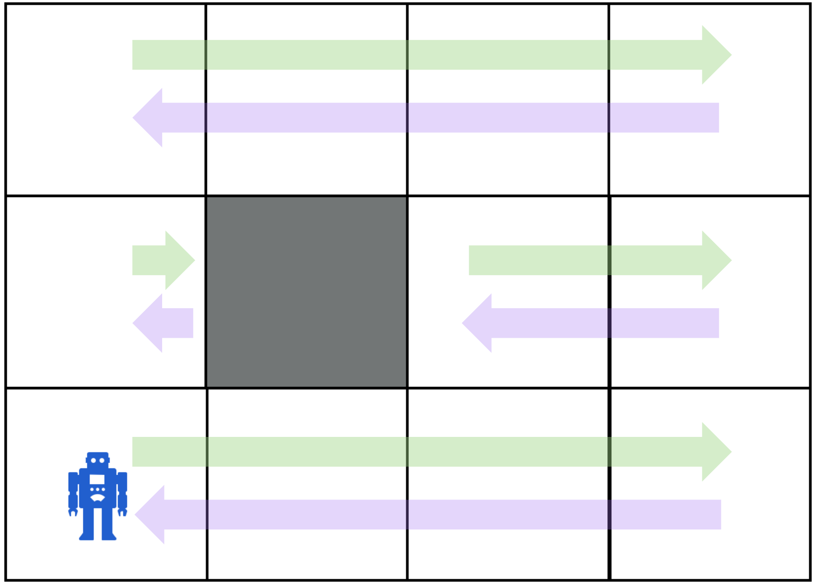
Theorem 1. *For each of SOAP, PO, and TO, there exist $(E, \mathcal{T})$ pairs for which no reward function realizes $\mathcal{T}$ in E.*

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 <u>realizes</u> $\mathcal{T}$ in E?

Expressivity Example 1

What kinds of SOAPs are not expressible?

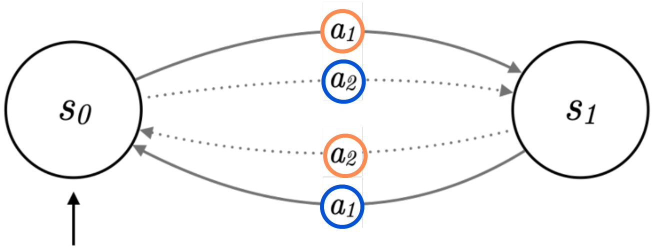


$$\Pi_G = \{\pi_{\leftarrow}, \pi_{\rightarrow}, \ldots\}$$

SOAP = “Always go in the same direction”

Expressivity Example 2

What kinds of SOAPs are not expressible?



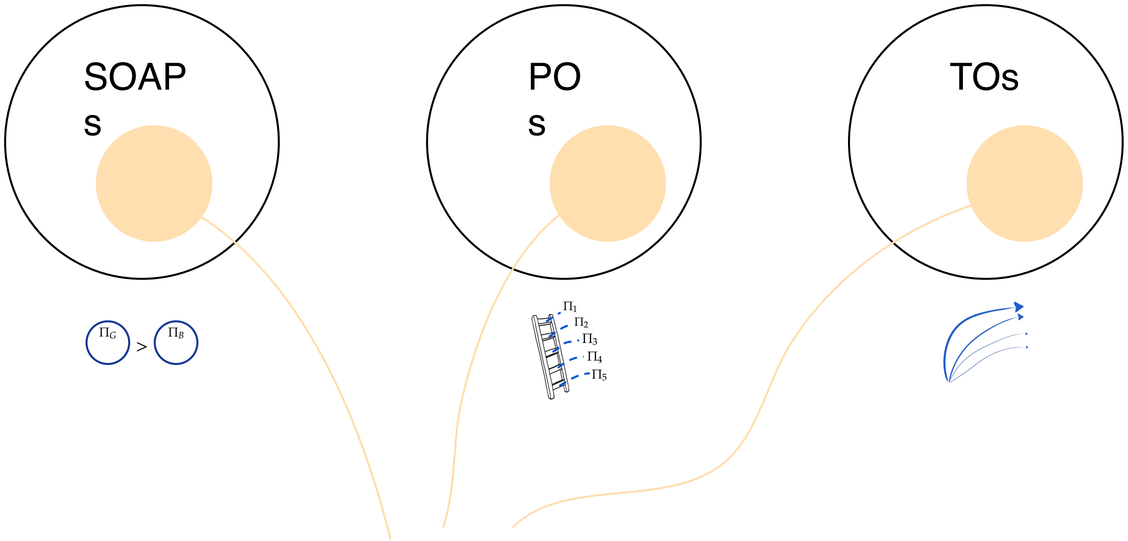
$$\Pi_G = \{\boxed{\pi_{21}}, \boxed{\pi_{12}}\}$$

XOR Problem

...Other types?

Question 2: Can We Find the Realizing Rewards?

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* $M = (E, R_{alice})$.



Reward can express

Main Result 2: 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.

Algorithm: SOAP 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

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 $\mathcal{T}$.*

Other Analysis: Two Kinds of SOAP

“range” SOAP

$$\Pi_G > \Pi_B$$

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

“equal” SOAP

$$\Pi_G > \Pi_B$$

$$V^{\pi_g}(s_0) = V^{\pi'_g}(s_0)$$

AND

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

Proposition 1. *The “range” realization of SOAP is strictly more general than the “equal” realization.*

Other Analysis

Extensions of Main Results

Theorem 3. *There exist choices of* $E_\neg = (\mathcal{S}, \mathcal{A}, s_0)$ *and* $\mathscr{T}$, *such that there is no* (T, R, γ) *that realizes* $\mathscr{T}$ *in* $E_\neg$.

Theorem 4. *The* Finite-RewardDecision *problem is NP-hard.*

Multi-Environment

Theorem 5. *Given a task* $\mathscr{T}$ *and a finite set of CMPs,* $\mathcal{E} = \{E_1, \ldots, E_n\}$, *with shared state–action space, there exists a polynomial time algorithm that outputs one reward function that realizes the task (when possible) in all CMPs in* $\mathcal{E}$.

Theorem 6. *Task realization is not closed under sets of CMPs. That is, there exist choices of* $\mathscr{T}$ *and* $\mathcal{E} = \{E_1, \ldots, E_n\}$ *such that* $\mathscr{T}$ *is realizable in each* $E_i \in \mathcal{E}$ *independently, but there is not a single reward function that realizes* $\mathscr{T}$ *in all* $E_i \in \mathcal{E}$ *simultaneously.*

Limitations & Assumptions

Environment.

> Finite CMPs.
> γ is part of E.

Task Realization.

> Start-state value determines task realization.
> Ignore learning dynamics.

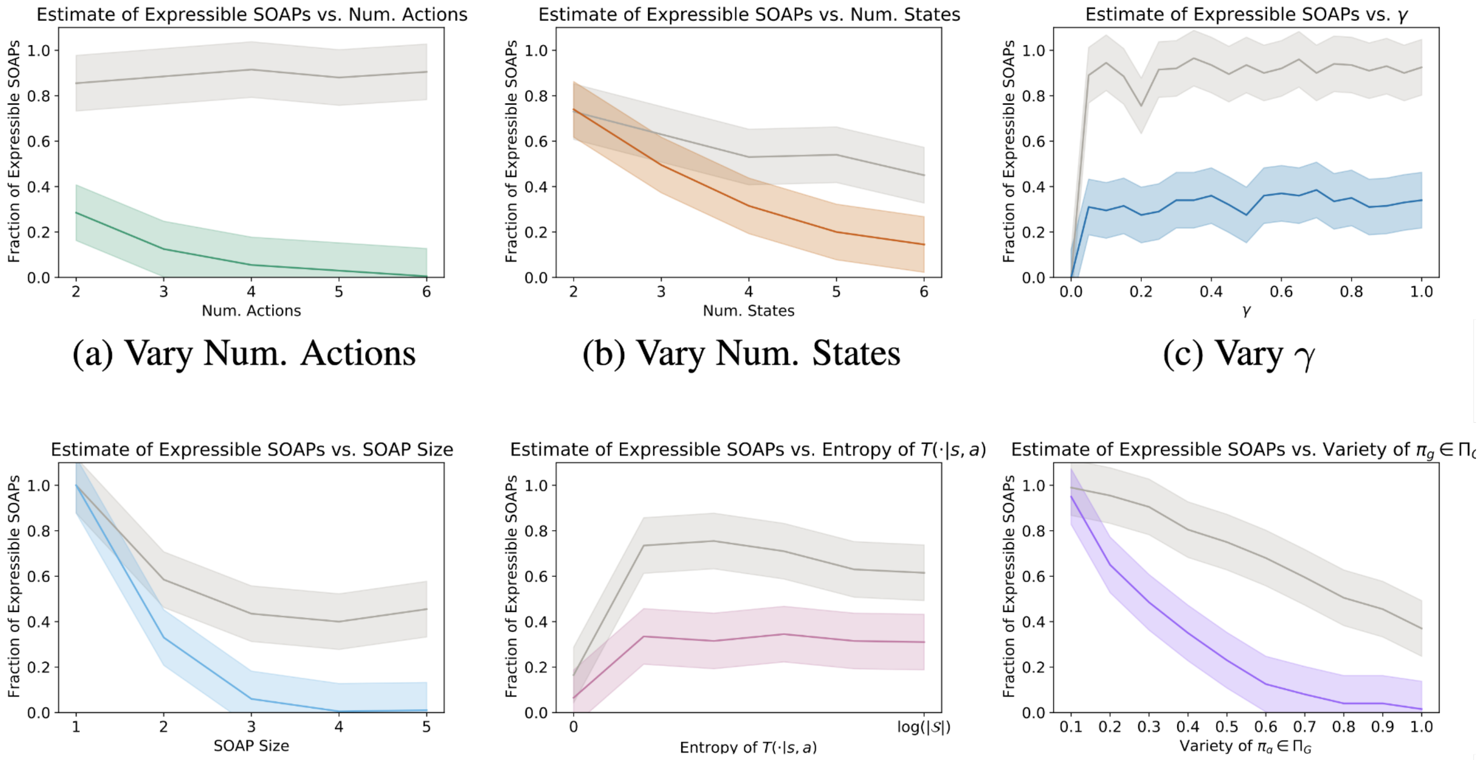
Task.

> Tasks of interest are SOAPs, POs, and TOs.

Reward Functions.

> Deterministic.
> Markov.

Experiment 1: SOAP Expressivity



(a) Vary Num. Actions

(b) Vary Num. States

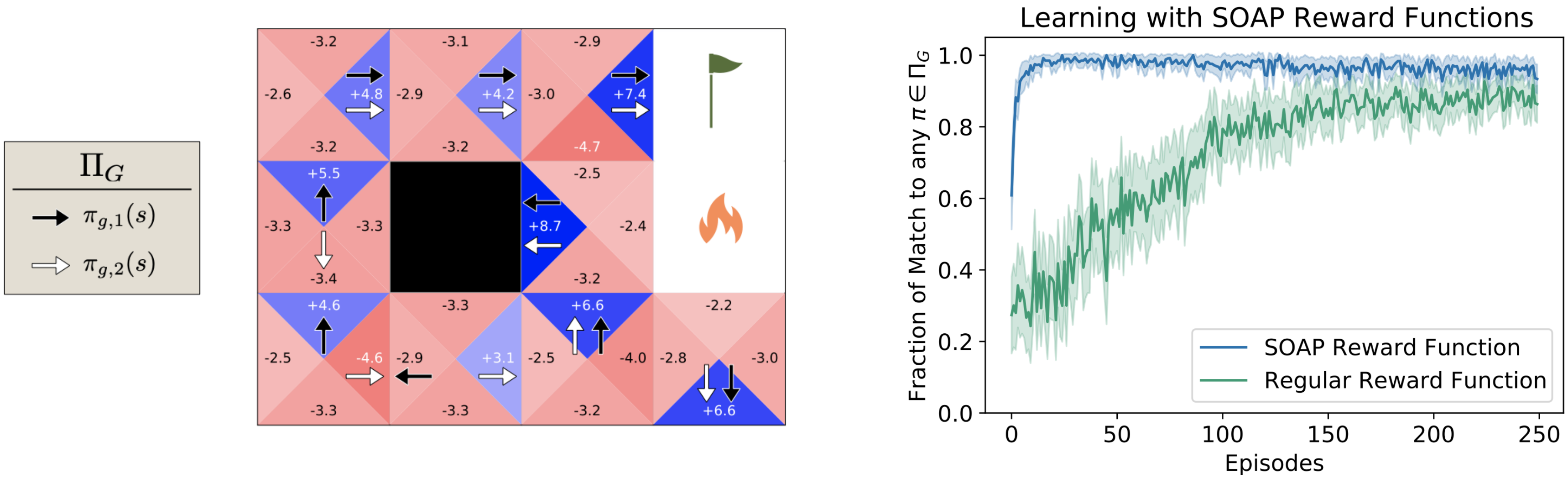
(c) Vary γ

(d) Vary SOAP Size

(e) Vary Entropy of T

(f) Vary the Spread of Π_G

Experiment 3: Learning with SOAP Rewards (Grid)



Main Result Overview

“Always go in the same direction”