Zero-Shot Coordination and Off-Belief Learning

Jakob Foerster
Associate Professor, Department of Engineering Science

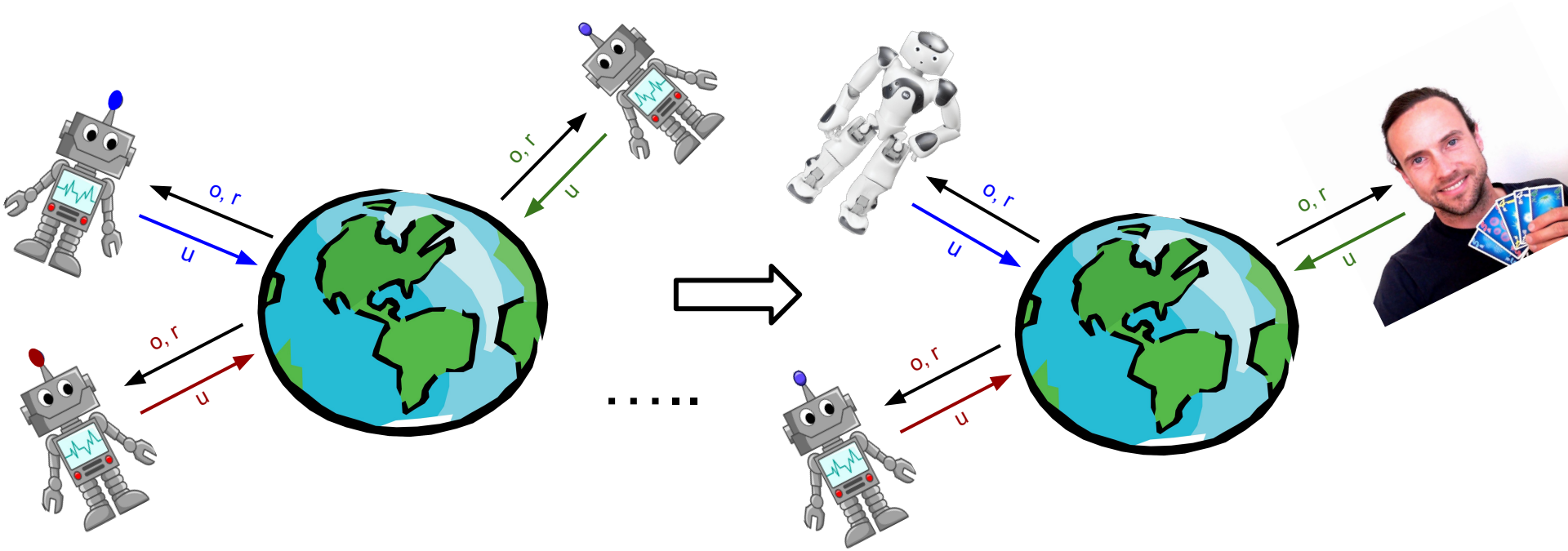
UNIVERSITY OF OXFORD

Outline of the talk

- *Why* do we (need to) care about coordination?
- What's a good *problem setting* and *formulation*?
- Off-belief Learning
- Q&A

Why do we (need to) care about coordination?

Why do we (need to) care about coordination? (Part 1)



Part 2: We want AI agents that can *help* and *support* humans

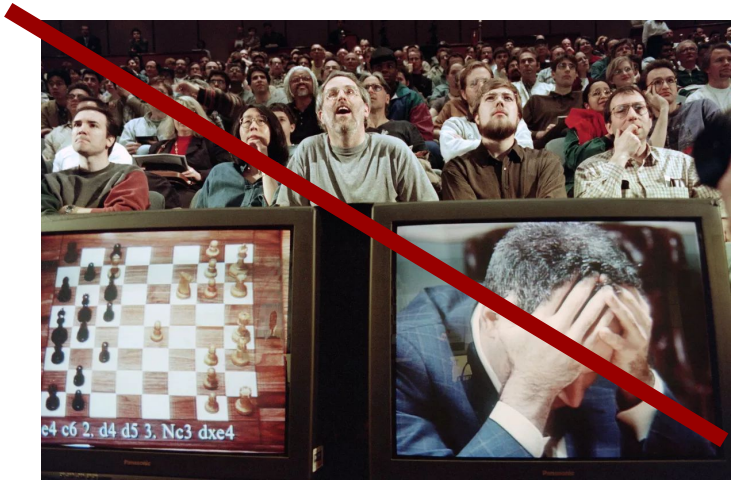
- By definition, a *multi-agent problem* (human & AI agent)
- By definition *fully cooperative* (goal is to help human)
- Commonly the reward function of the human (amongst other things) will be unknown
  - ➢ *Partially observable*

Human-AI Coordination is a Dec-POMDP

- Can't pre-agree on action in each state
  - ➢ *Coordination Problem*

What’s a good problem setting and formulation?

Issue 1: Standard Benchmarks (Poker, Go, chess..) are competitive



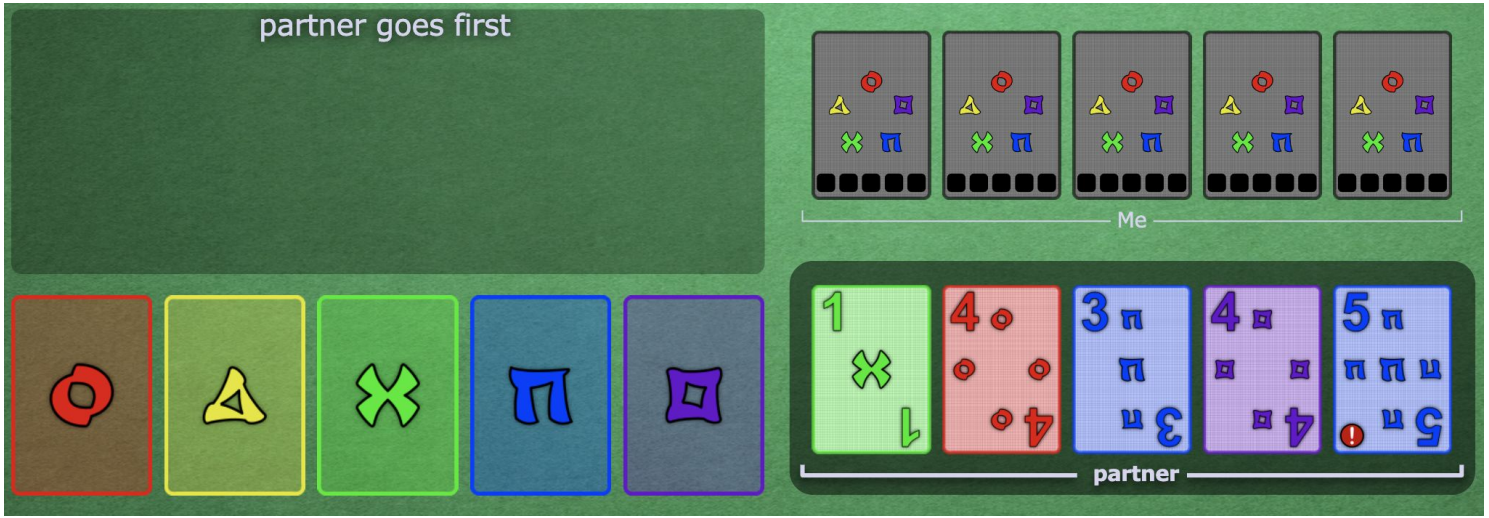
We need *partially observable*, *fully cooperative* benchmarks

We present: Hanabi!

The Hanabi Challenge: A New Frontier for AI Research
Nolan Bard*, Jakob N. Foerster*, Sarath Chandar, Neil Burch, Marc Lanctot, H. Francis Song, Emilio Parisotto, Vincent Dumoulin, Subhodeep Moitra, Edward Hughes, Iain Dunning, Shibl Mourad, Hugo Larochelle, Marc G. Bellemare, Michael Bowling
Artificial Intelligence, 2020

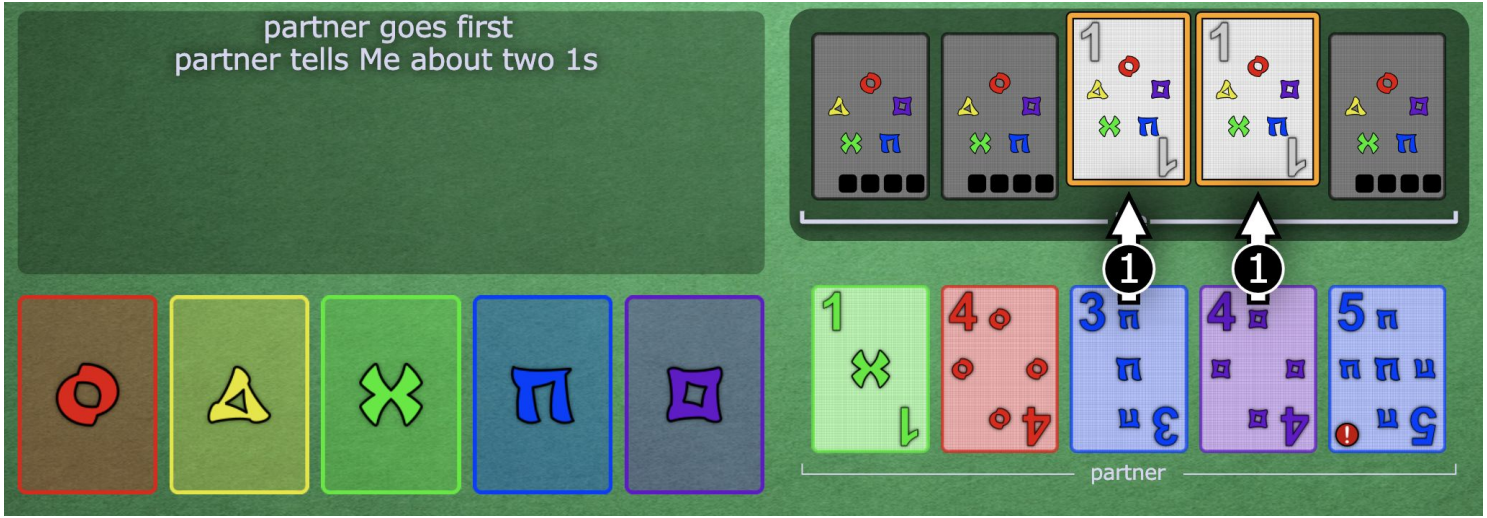
- Dec-POMDP: Fully **cooperative**, **partially observable**, entirely focussed on **Theory of Mind.**
- The perfect *test bed* for **human-AI and AI-AI coordination!**

Introduction to Hanabi: Build Fireworks!



Hint Tokens Remaining: 8, Score: 0

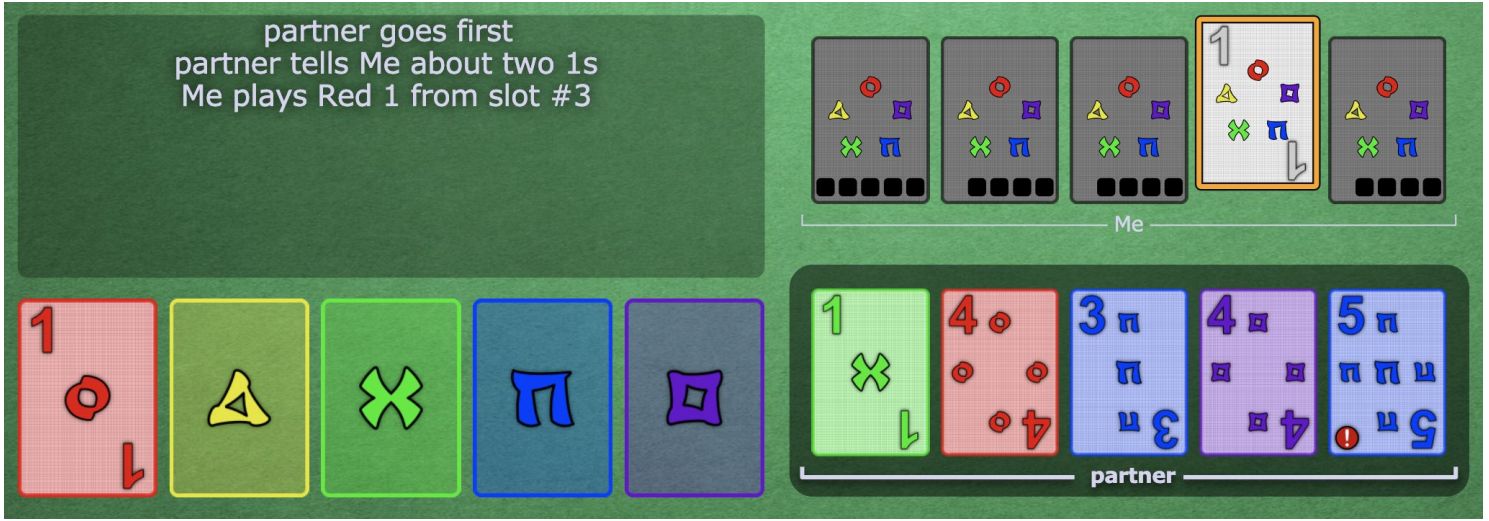
Introduction to Hanabi: Build Fireworks!



Hint Tokens Remaining: 7, Score: 0

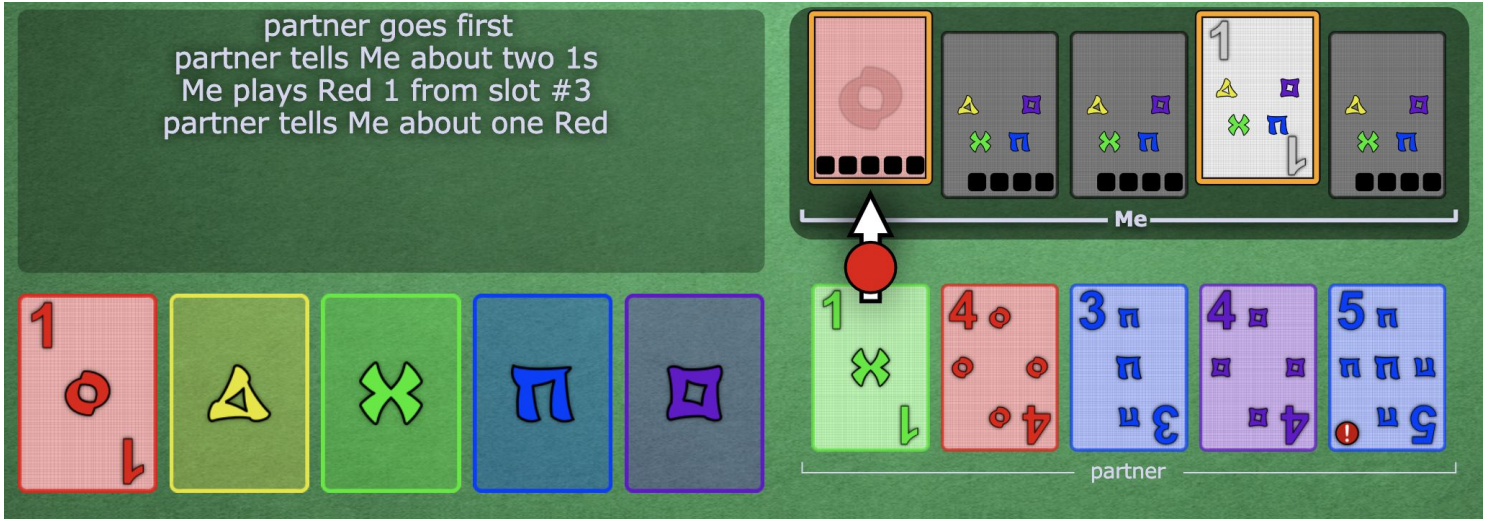
-1

Introduction to Hanabi: Build Fireworks!



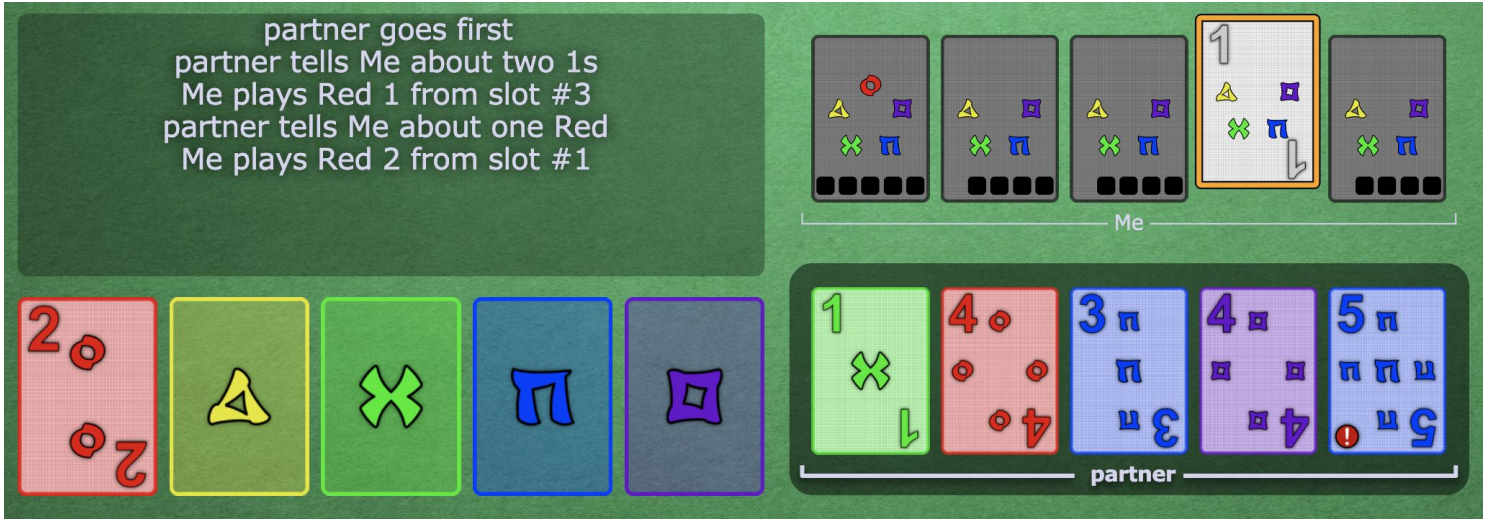
Hint Tokens Remaining: 7 Score: 1

Introduction to Hanabi: Build Fireworks!



Hint Tokens Remaining: 6, Score: 1

Introduction to Hanabi: Build Fireworks!



Hint Tokens Remaining: 6, Score 2

Issue 2: The default problem setting is *Decentralized control ("self-play")*

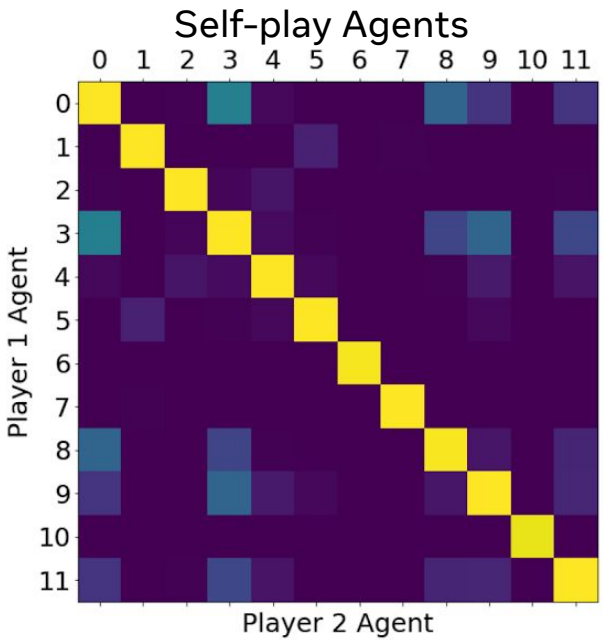
Training

Testing

"Sound" in two-player zero-sum, since *Nash equilibria* are interchangeable.

In cooperative settings, these policies typically only perform well within **the specific team of AI agents** they are trained with.

Near optimal *self-play* policies are terrible at Coordination



Example conventions:

- Hint *Red* or *Yellow*: "Play newest card."
- Hint *White* and *Blue*: "Probably discard last card"

Extremely efficient, but *very different* from human conventions.

Clearly uses, *arbitrary* codes. Why not use *White* to indicate play?

Result:
Self-play score >24, Cross-Play score ~3, **Human-AI score ~0.8**

Problem Setting Option 1: Human-AI Coordination

Benefits:

- No need for a “proxy” problem setting
- Easy to motivate

Downside:

- Scientist need to always test with humans to evaluate progress
- Distribution of humans chosen may change results (reproducibility?)
- Difficult to generalize to novel problem settings

This problem has been solved in other areas

1) Model Organism

2) Human Study

https://www.pfizer.co.uk/clinical-trials

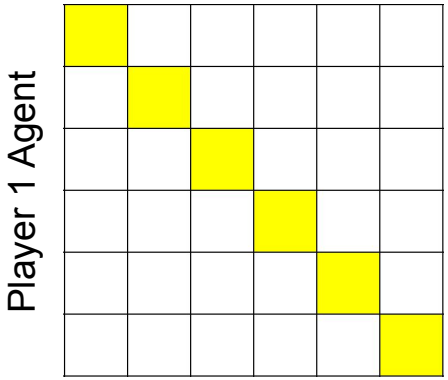
Option 2: Zero-shot coordination [*Hu et al, ICML 2020*]

Testing in Cross-Play

Coaches can agree on a training strategy before training starts.

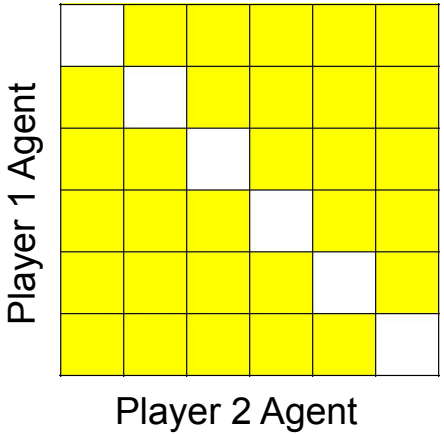
What should the strategy be?

Self-Play Objective



Only performance in the “original” (training time) team matters

ZSC Objective



Algorithm that maximizes *cross-play performance at test time* between:

1) “Independent training runs (‘seeds’) of the same algorithm”
   Original formulation (*Hu et al, ICML 2020*)

2) “Independent training runs of *independent implementations* of the same algorithm, *without sharing code or hyperparameters*
   Updated, current formulation (*Treutleine et al, ICML 2021*)

Comparison w/ Ad-Hoc teamplay [Stone et al, 2010]

Ad-hoc teamplay - what is it?

Quoting from their paper:

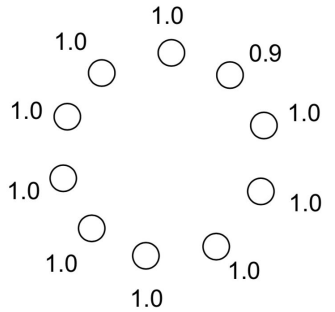
"That is, we challenge the community:

> **To create an autonomous agent that is able to efficiently and robustly collaborate with previously unknown teammates on tasks to which they are all individually capable of contributing as team members."**

And:

"Note further that the agents in A need not be themselves aware that they are acting as teammates;"

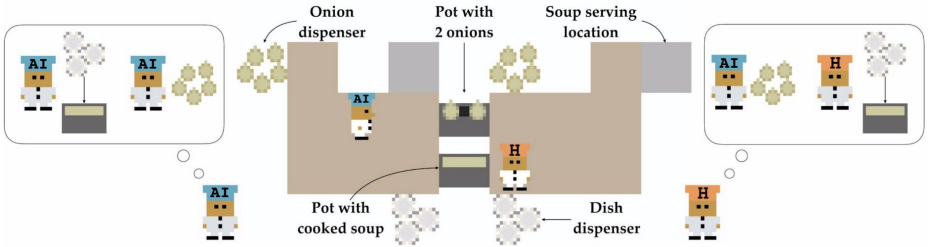
- **Ad-hoc teamplay** aims to play a *best response* to a 'good' or 'near optimal' *pool of policies*
- Since the pool is assumed to be given *a-priori,* ad-hoc teamplay is a *single agent* problem,
- This breaks the *spirit of coordination*, which is about *self-fulfilling prophecies (equilibrium selection)*
- In particular, a *best response* to self-play policies is not in general *self-consistent* or a good coordination strategy
- In our "lever game", ZSC plays 0.9, ad-hoc 1.0:



Other Related work in Machine Learning

- Human-AI coordination as a post-hoc finding:
  - E.g. Avalon (“Finding friend and foe in multi-agent games”, Serrino et al), Dota 3 (Open-AI)

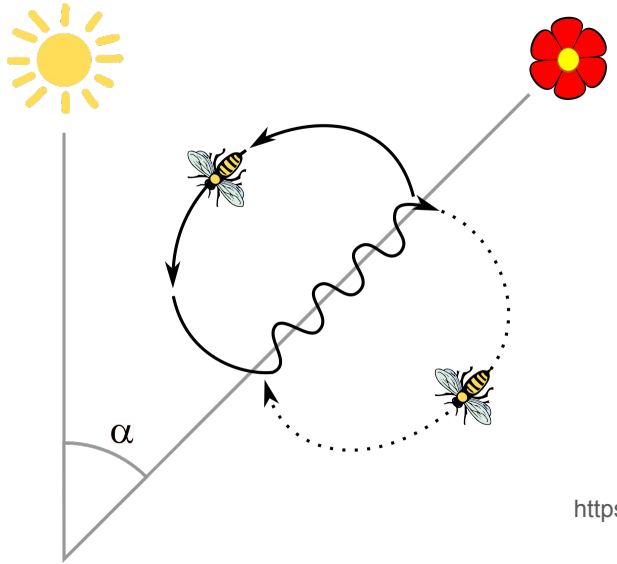
- Human-AI coordination using *human data*:
  - e.g. in Overcooked (”On the Utility of Learning about Humans for Human-AI Coordination”. Carroll et al, NeurIPS 2019.



How can we learn optimal grounded policies?

Fundamental issue of RL in Dec-POMDPs

Rather than using the *grounded information*, agents can learn *arbitrary conventions* for encoding information, like a bee *waggle dance*:



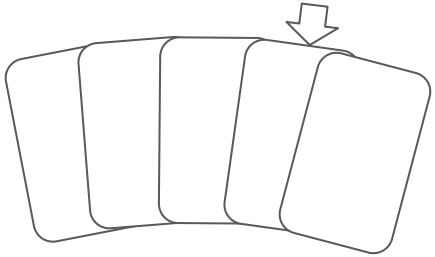
https://en.wikipedia.org/wiki/Waggle_dance

Even if accounting for symmetries there might still be many ways of encoding information..

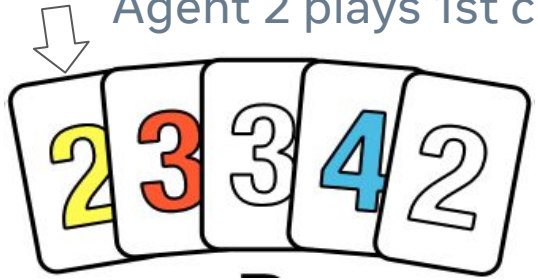
Self-Play Example

"Fireworks"

Agent 1 discards 4th card:



Agent 2 plays 1st card:

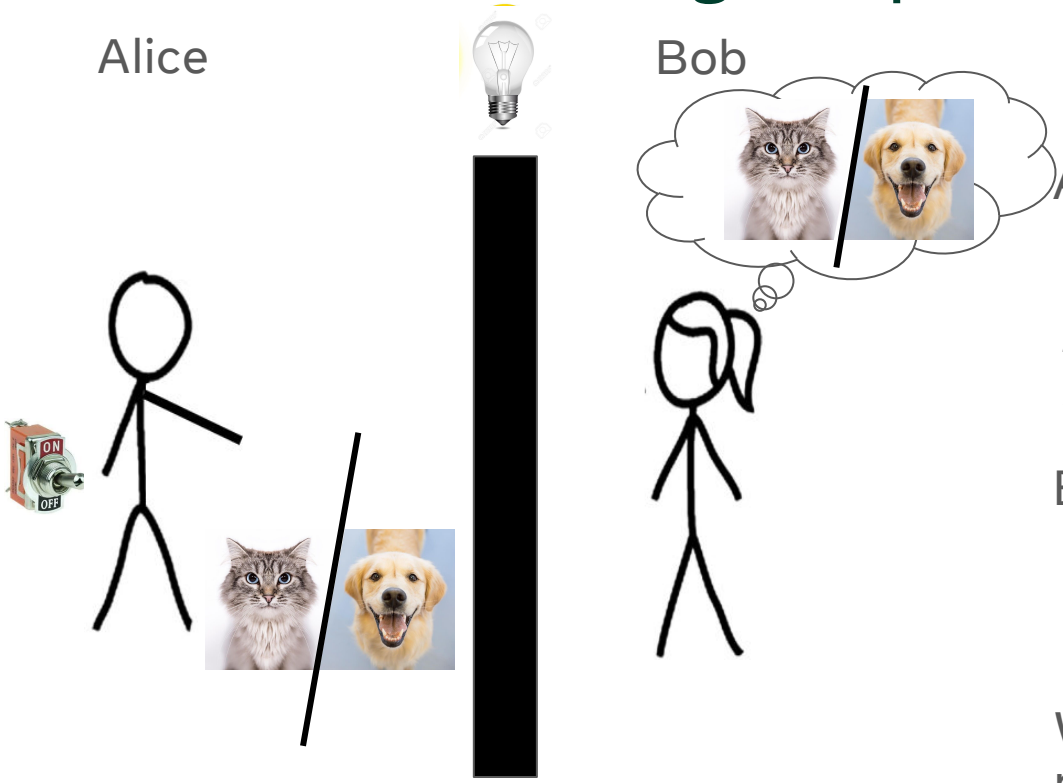


Agent 1 happens to **discard 4th card** when **1st card** of agent 2 is **playable**

Agent 2 learns to play **1st card** whenever **4th card** is discarded by agent 1

A highly *arbitrary* convention is formed.

Thought experiment



Alice can:

1. Turn on the light bulb.
2. Bail out, get 1 dollar.
3. Pay $5 to remove the barrier.

Bob can:

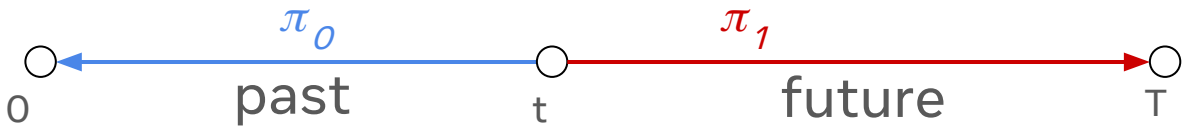
1. Guess the pet, get 10 dollar if correct, -10 dollar if wrong.

What is the optimal self-play policy? What about OBL?

We introduce: Off-Belief Learning

Off-Belief-Learning (OBL) solves this:

- We train an OBL policy, π_1, that *interprets* all *past actions* as if they were played by policy π_0, but assumes that *future actions* are played by π_1



- Rather than allowing the agents to agree upon *arbitrary protocols*, π_0 specifies the *meaning* of each action.
- Example:
  - Random π_0 restricts the meaning to be only *grounded* information.
  - In Hanabi “1st card is red” *only* means this card is red (not: “play 2nd card”).
  - Discarding a card means nothing :)

Optimal Grounded Policy

1. *Grounded* belief

If π_0 is constant it cancels out

$$P(\tau|\tau^i,\pi_0)=\frac{P(\tau)\prod_t P(o_t^i|\tau)\pi_0(a_t|\tau_t^{-i})}{\sum_{\tau'}P(\tau')\prod_t P(o_t^i|\tau')\pi_0(a_t|\tau_t'^{-i})} \quad \Rightarrow \quad P(\tau|\tau^i)=\frac{P(\tau)\prod_t P(o_t^i|\tau)}{\sum_{\tau'}P(\tau')\prod_t P(o_t^i|\tau')}.$$

2. Optimal grounded policy

   a. Assumes that given a grounded belief, both players **act optimally in the rest of the game**.

   b. Best responding to a random policy is not an **optimal** grounded policy.

   c. Training a feedforward policy w/ grounded belief as input does not result in grounded policy.

OBL Value Functions / Fictitious TD Learning

- The value function is defined as:

$$V^{\pi_0 \to \pi_1}(\tau^i) = \mathbb{E}_{\tau \sim \mathcal{B}_{\pi_0}(\tau^i)}\left[V^{\pi_1}(\tau)\right]$$

What *would be* the distribution over *world states* if we had reached τ^i while playing with π^0?

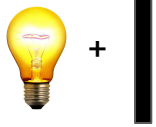
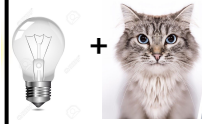
- Q-function is defined as:

$$Q^{\pi_0 \to \pi_1}(a|\tau_t^i) = \sum_{\tau_t, \tau_{t+1}} \mathcal{B}_{\pi_0}(\tau_t|\tau_t^i)\left(R(s_t, a) + \mathcal{T}(\tau_{t+1}|\tau_t)V^{\pi_1}(\tau_{t+1})\right)$$

Off-Belief Learning vs Self-Play

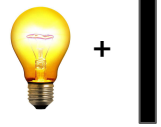
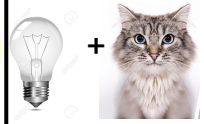
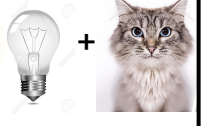
- Assume worst case, where Bob's initial policy *only* turns on the light for "cat".
- What do Alice and Bob learn under OBL, assuming π_0 is random?
- Since π_0 is *random, OBL state* is *independent* of the light bulb.

Self-Play

State	Bob action	Alice Observation	Alice Opt. action	$
				**10**
				5

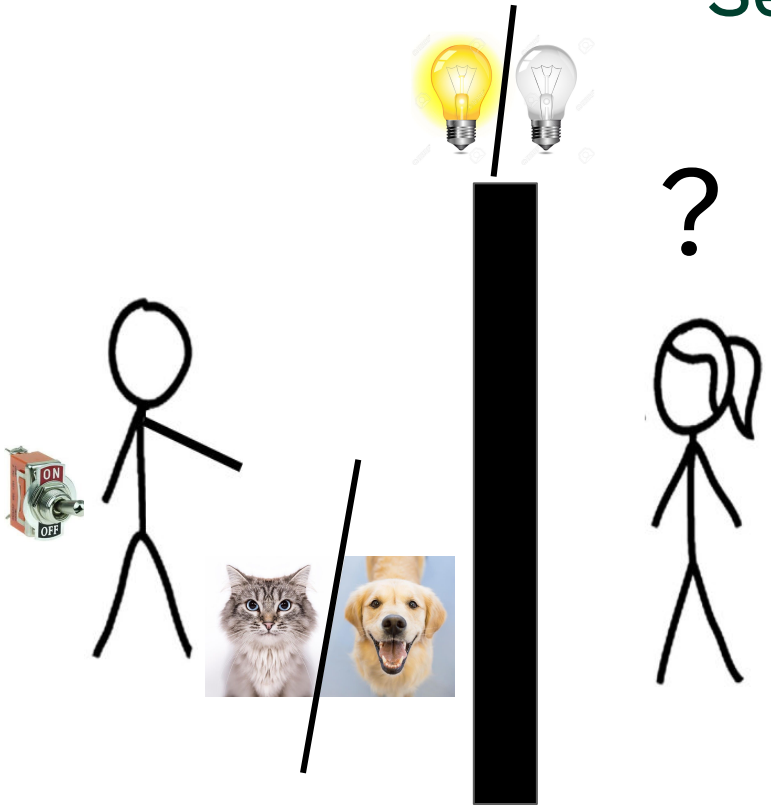
Under SP, Bob's optimal policy is to signal w/ light.

Off-Belief Learning

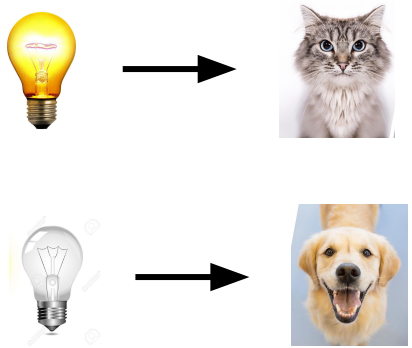
State	Bob action	Alice Observation	OBL state	Alice Opt. action	$
					0
					**5**

Under OBL, Bob's optimal policy is to *remove the barrier.*

Self-Play



- Optimal self-play policy learns arbitrary convention:

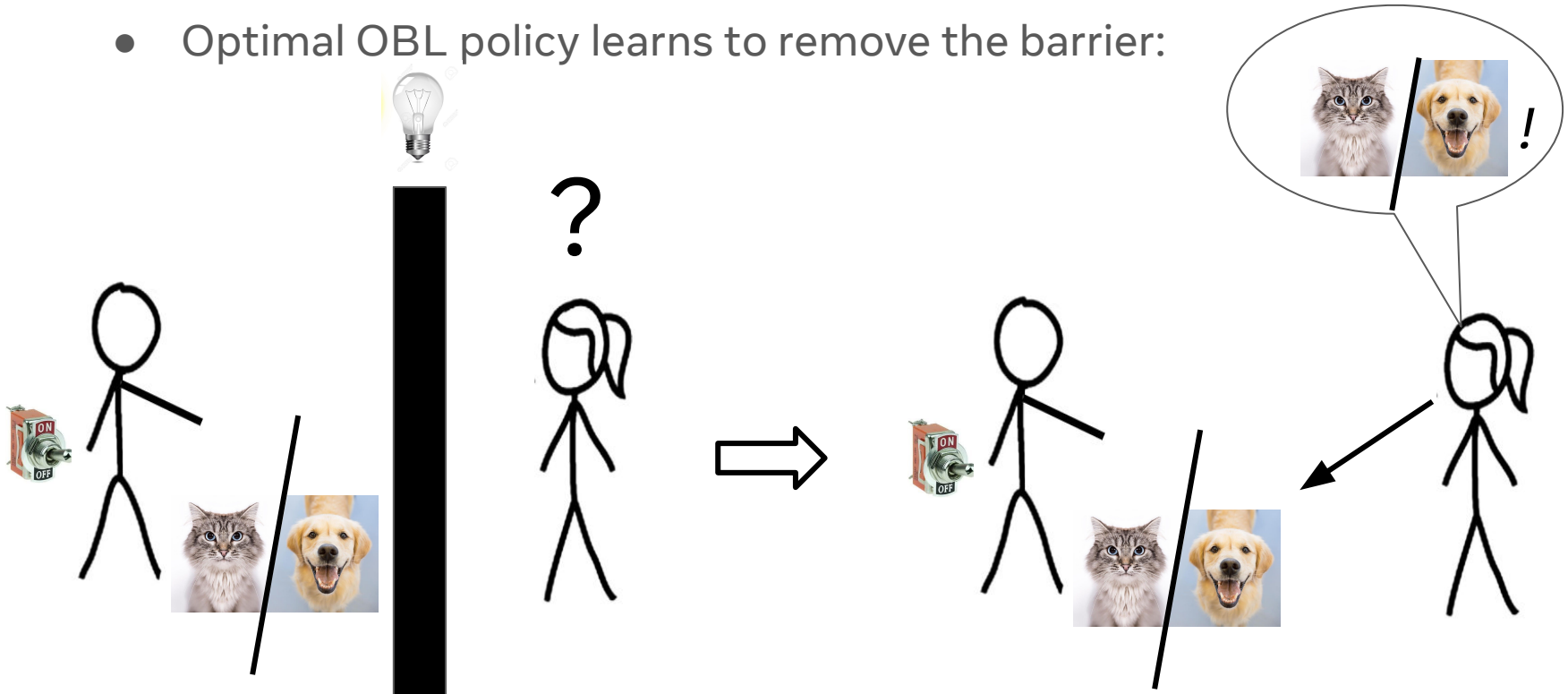


(or vice versa!)

- Average self-play reward is +10
- Average cross-play reward is 0!
- Will *fail* at test time (e.g. with human)

OBL

- Optimal OBL policy learns to remove the barrier:

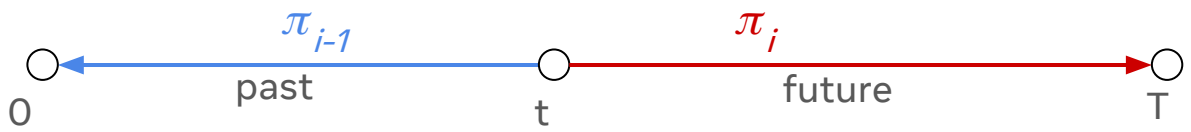


- Average self-play reward is +5
- Average cross-play reward is +5
- Human compatible!

OBL-Hierarchy

- This is great, but what if we do need some (finite) amount of *counterfactual reasoning*?

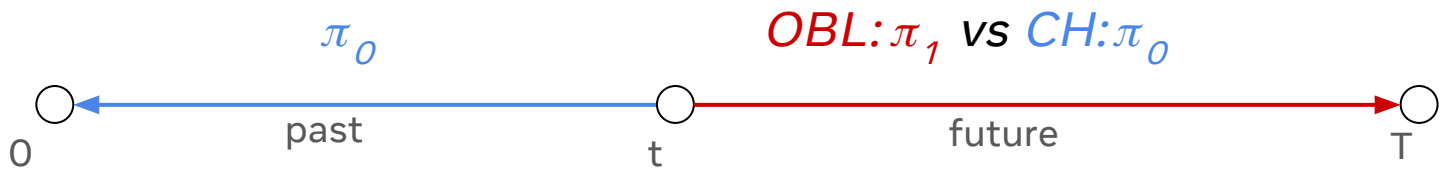
- OBL can be extended into a *hierarchy*, where each level i takes as starting point the optimal policy of the previous level, *i-1*.



- I.e assume past actions were played by π_{i-1}
- Near *optimal* play style from *one level* induces conventions for the *next one*

OBL vs Cognitive Hierarchies

- At the first level, both OBL and Cognitive hierarchies assume that *past actions* are played by a fixed given policy π_0
- In contrast to OBL, CH simply plays a best-response to π_0 i.e. assumes that future actions are also taken by π_0.



- CH will not learn to use *grounded channels* to signal at level 1.

Properties of OBL

Theorem 1. *For any $T > 0$ and starting policy π_0, OBL computes a unique policy π_1.*

Theorem 2. *For every policy π_1 generated by OBL from π_0, $J(\pi_1) \geq J(\pi_0) - eTt_{max}$, i.e. OBL is a policy improvement operator except for a term that vanishes as $T \to 0$.*

Properties of OBL

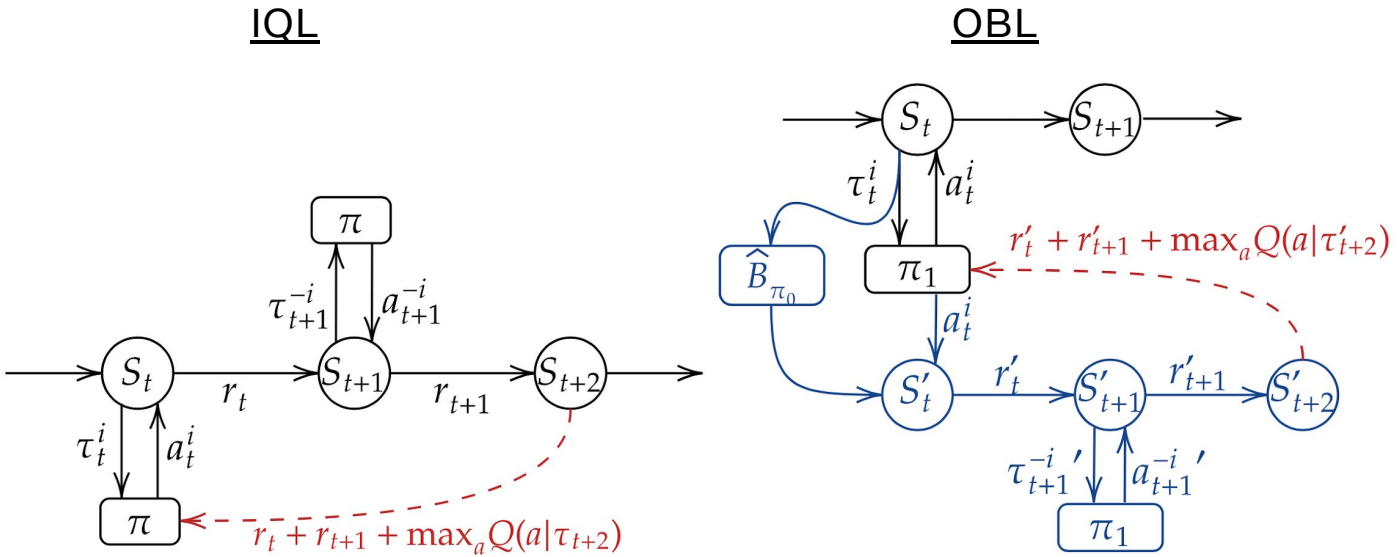
Theorem 3. *If repeated application of the OBL policy improvement operator converges to a fixed point policy π, then π is an ϵ-subgame perfect equilibrium of the Dec-POMDP, where $\epsilon = et_{max}T$.*

Theorem 4. *Application of OBL with temperature 0 to any constant policy $\pi_0(a|\tau^i) = f(a)$ - or in fact any policy that only conditions on public state - yields an optimal grounded policy.*

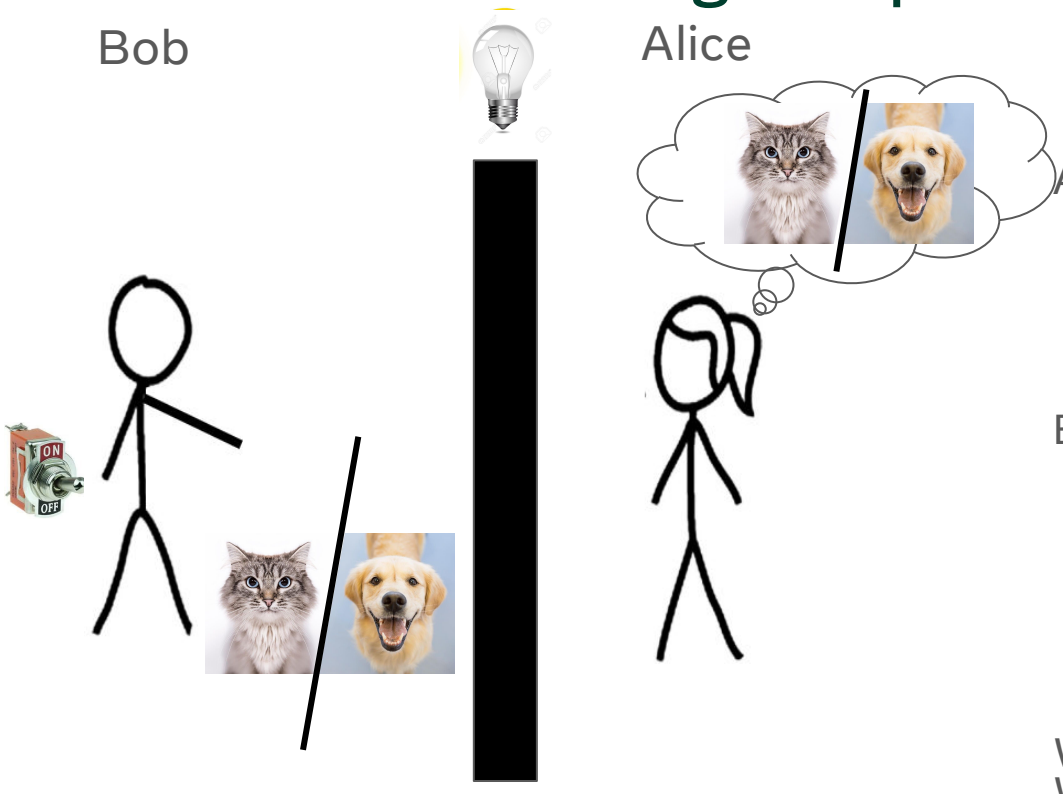
Scalable Fictitious Transition Mechanism

We can implement OBL via Q-learning or other Deep RL methods:

$$Q^{\pi_0 \to \pi_1}(a_t|\tau_t^i) = \mathbb{E}_{\tau_t \sim \mathcal{B}_{\pi_0}(\tau_t^i), \tau_{t+k} \sim (\mathcal{T}, \pi_1)} \Big[\sum_{t'=t}^{t+k-1} R(\tau_{t'}, a_{t'}) + \sum_{a_{t+k}} \pi_1(a_{t+k}|\tau_{t+k}^i) Q^{\pi_0 \to \pi_1}(a_{t+k}|\tau_{t+k}^i) \Big]$$



Thought experiment



Alice can:

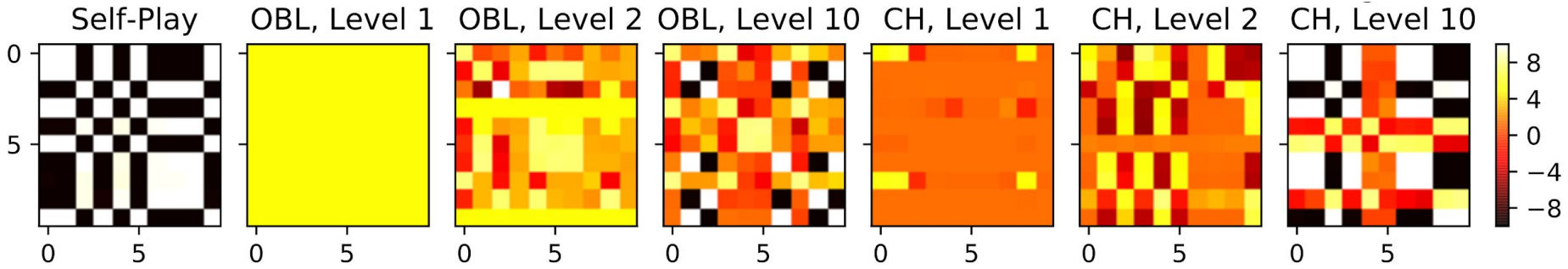
1. Turn on the light bulb.
2. Bail out, get 1 dollar.
3. Pay $5 to remove the barrier.

Bob can:

1. Guess the pet, get 10 dollar if correct, -10 dollar if wrong.
2. Bail out, receive 0.5 dollar.

What is the optimal self-play policy?
What about OBL?

Results in the *Toy Game*

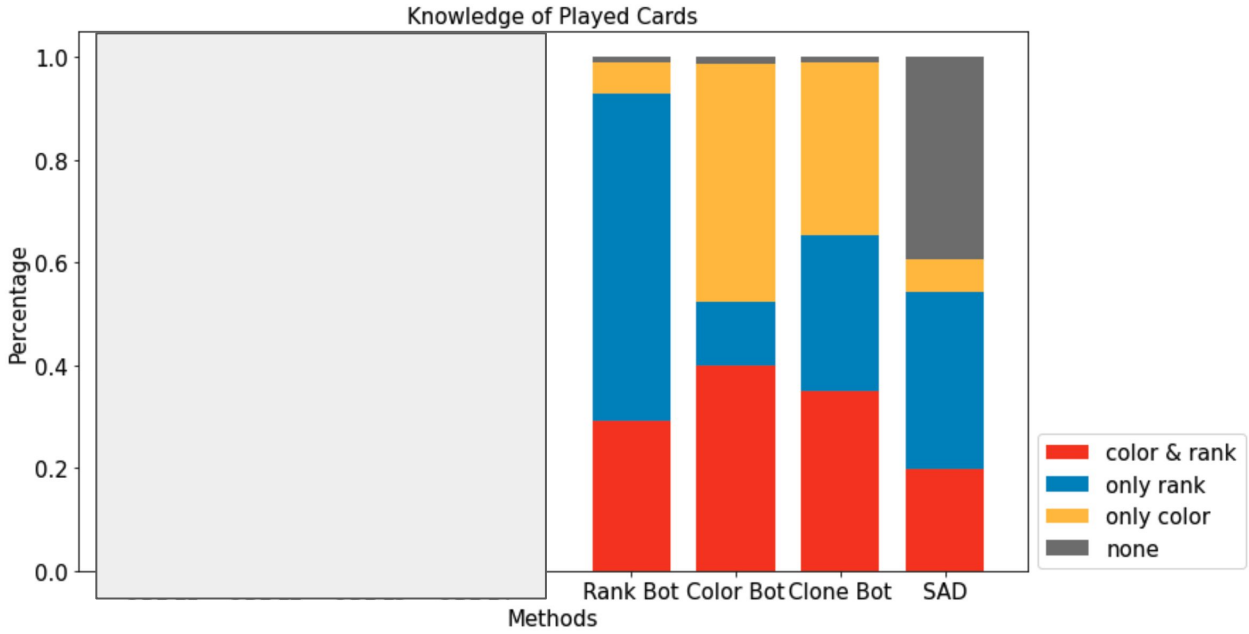


Optimal grounded policy (5 points)

Hanabi Evaluation

1. Self-play score: is it a good policy?
2. Cross-play score: does it produce consistent policy across different runs?
    a. a necessary condition for good **zero-shot coordination** score
3. Ad-hoc teamplay with some distinct policies unseen during training.
    a. Other-Play(Rank Bot)
    b. Other-Play(Color Bot)
4. Zero-Shot human-AI:
    a. Clone Bot from huma data

Analysis of Play-Style in Hanabi



Results in Hanabi

Method	Self-Play	Cross-Play	w/ Other-Play (Rank Bot)	w/ Other-Play (Color Bot)	w/ Clone Bot
SAD(*)	23.97 ± 0.04	2.52 ± 0.34	3.81 ± 0.99	0.06 ± 0.01	0.26 ± 0.12
Other-Play	24.14 ± 0.03	21.77 ± 0.68	22.81 ± 0.87	4.05 ± 0.37	8.55 ± 0.48
K-Level	16.97 ± 1.19	17.17 ± 0.98	14.80 ± 1.77	12.36 ± 1.44	13.03 ± 1.91

Summary

I presented:

- Zero-shot coordination, a *proxy setting* for human-AI
- Off-Belief Learning, a novel method that allows agents to learn optimal grounded policies
- OBL trains π_1 which assumes that past actions were played by a fixed, known policy π_0, but future actions will be played by π_1
- Conventions can be derived by applying OBL in a hierarchy
- OBL obtains SOTA human-AI results in Hanabi