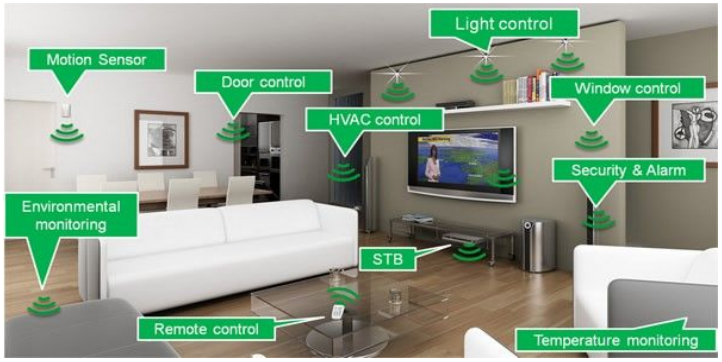
Multi-Agent Reinforcement Learning towards Zero-Shot Communication

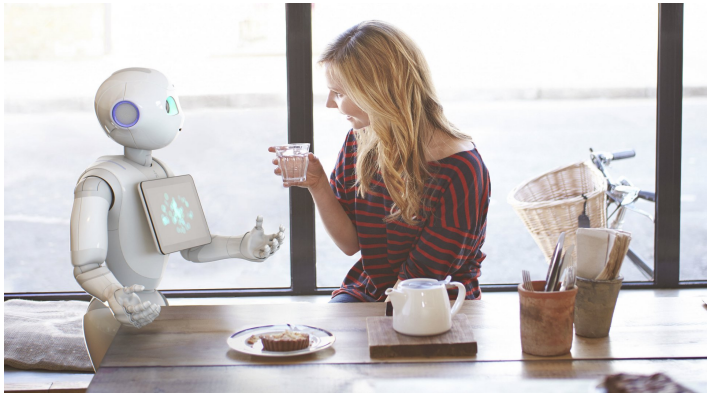
Kalesha Bullard, PhD
DeepMind

Oxford ML Summer School
14 August 2022

* Work done during Postdoc at **Facebook AI Research (**now: **Meta AI)**

WHY Multi-Agent Systems?





WHY **Cooperative** Multi-Agent Systems?

Real World is *Inherently* Multi-Agent

Necessary for Human-AI Coexistence

- Communication and Coordination amongst AI Agents and Humans

Many Relevant Application Domains

- Self-Driving Vehicles
- Robots in Human Environments
- Ecology and Evolutionary Biology
- Climate Change and Sustainability

A Bit of MY Background…

PhD: Georgia Tech (USA)

Clip Credit: Stanford NLP Seminar

Kalesha Bullard

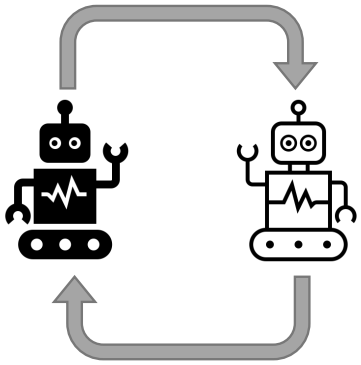
Facebook AI Research

Multi-Agent Reinforcement Learning towards Zero-Shot Emergent Communication

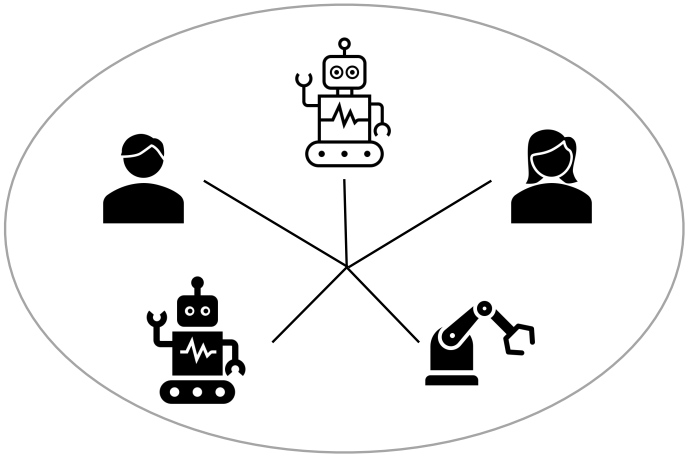
Postdoc: Facebook AI Research [Meta AI] (USA)

Research Scientist (Current): Google DeepMind (UK)

Research Vision:
Interactive Learning towards Multi Human-AI Cooperation



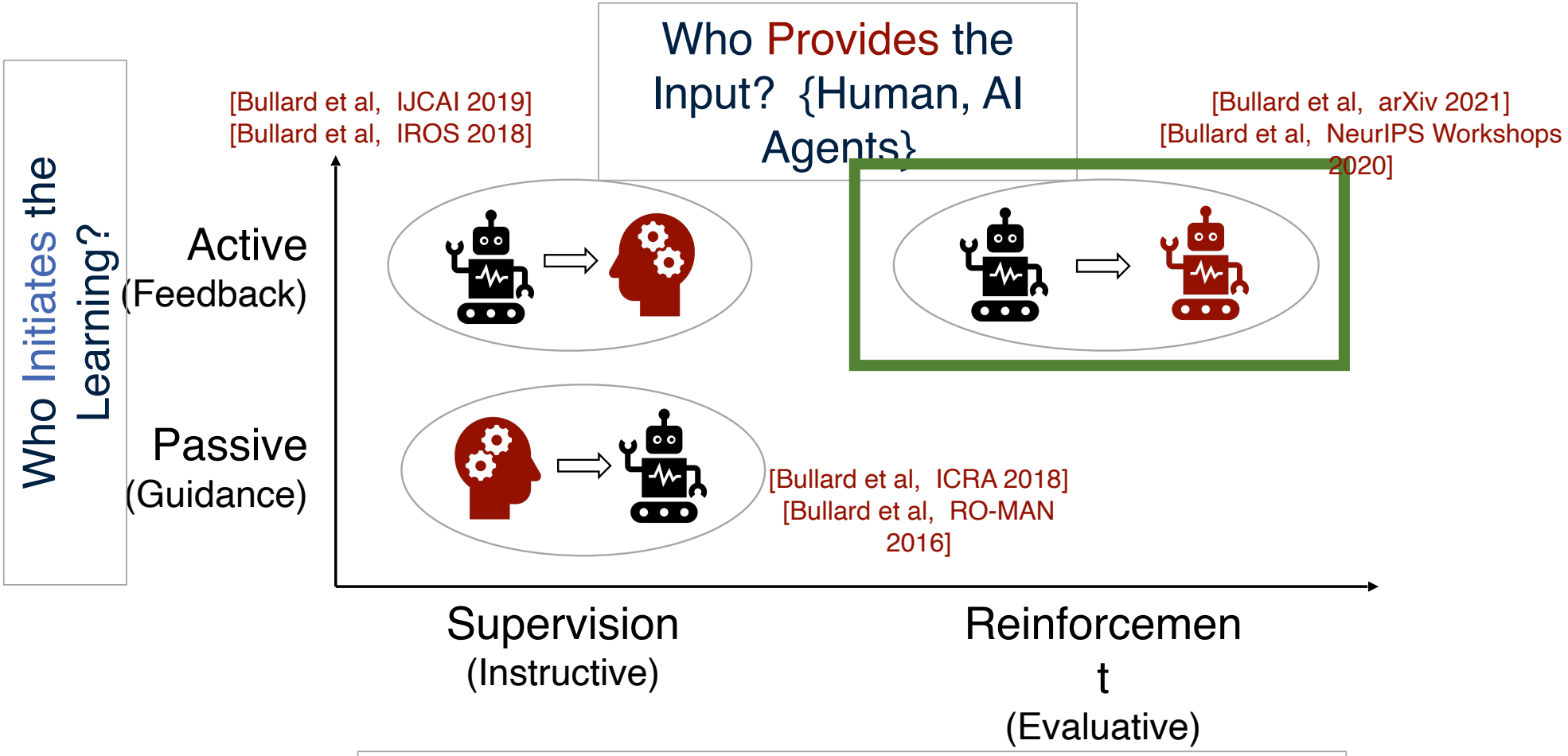
To Enable



Interactive Learning (methods)

Multi Human-AI Cooperation (long-term goal)

Learning through Interaction in Multi-Agent Systems

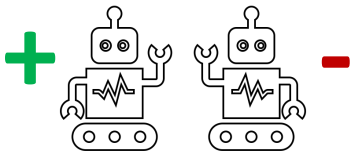


Categorizing Multi-Agent Reinforcement Learning through Reward Structure

Problem is typically *mathematically* formalized as a **GAME** -- between **Players** (Agents)

Game Problem Setting

Fully **Competitive** (Zero-Sum Games)



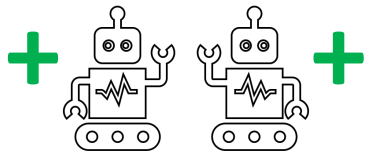
Adversaries

Mixed Motive (General Sum Games)

Examples:

- Social Dilemmas
- Competitive Team Games (e.g. soccer)
- …

Fully **Cooperative** (Coordination Games)



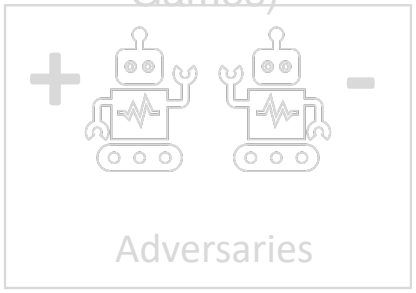
Teammates

Categorizing Multi-Agent Reinforcement Learning through Reward Structure

Problem is typically *mathematically* formalized as a **GAME** -- between **Players** (Agents)

Game Problem Setting

Fully **Competitive** (Zero-Sum Games)

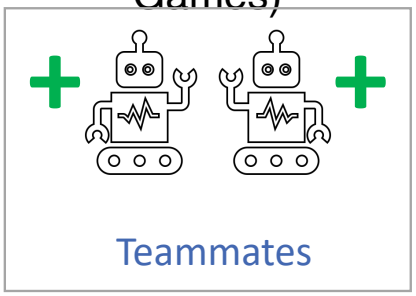


Mixed Motive (General Sum Games)

Examples:

- Social Dilemmas
- Competitive Team Games (e.g. soccer)
- …

Fully **Cooperative** (Coordination Games)



WHY **Cooperative** Multi-Agent Systems?

Real World is *Inherently* Multi-Agent

Necessary for Human-AI Coexistence

- Communication and Coordination amongst AI Agents and Humans

Many Relevant Application Domains

- Self-Driving Vehicles
- Robots in Human Environments
- Ecology and Evolutionary Biology
- Climate Change and Sustainability

Imagine all these **embodied agents** operated in a shared environment…

- How could *learn* to *communicate* -- to aid in **coordination**?
  - Supervised Learning from Datasets?
  - Imitation Learning from Demos?
- Should agents have to train with every potential partner to infer a **general** communication protocol?

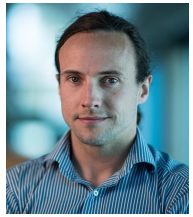
Autonomous Vehicle Ecosystem (in the *future*)

Kalesha Bullard (DeepMind)

Douwe Kiela (Hugging Face)

Franziska Meier (Meta AI)

Joelle Pineau (Meta AI, MILA)

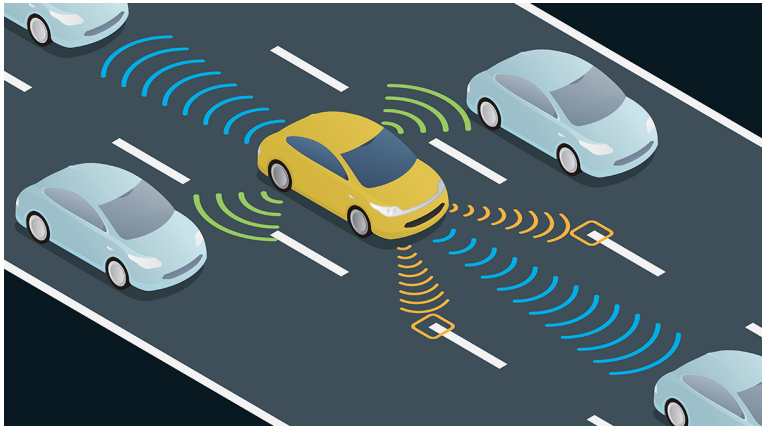


Jakob Foerster (Oxford)

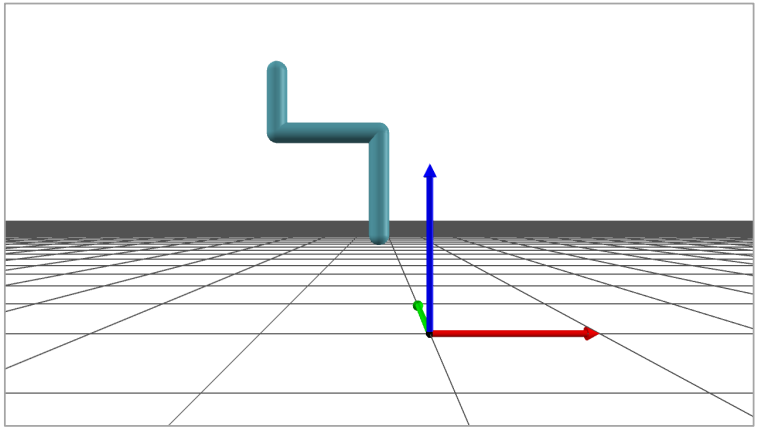
Multi-Agent RL for **Emergence** of Zero-Shot Communication Protocols

"Quasi-Equivalence Discovery for Zero-Shot Emergent Communication" [Bullard et al, arXiv 2021]
"Exploring Zero-Shot Emergent Communication for Embodied Agent Populations" [Bullard et al, NeurIPS Workshops 2020]

Motivating Research Question:
How can general communication skills **emerge** in *Embodied* Agents?
(through *Physical* Action)



Self-Driving Cars



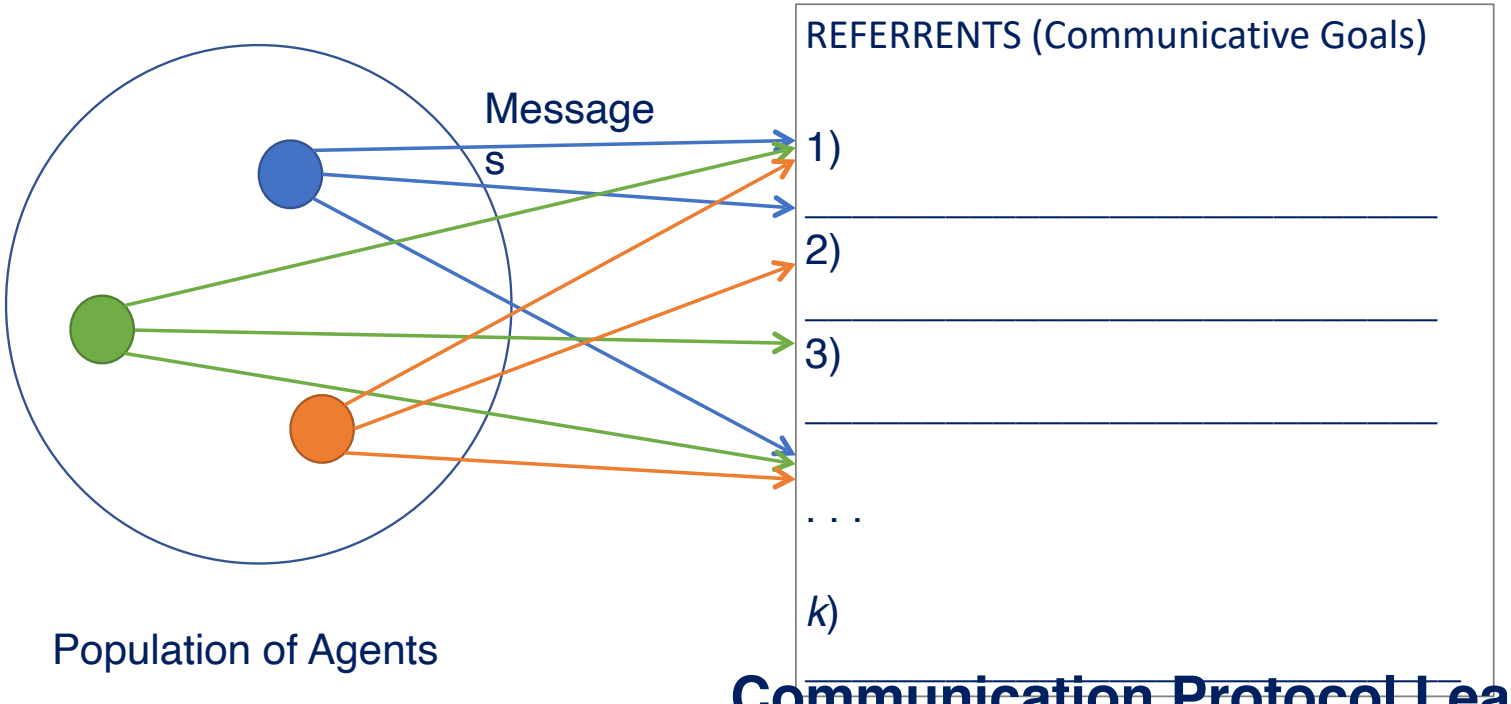
Robot Arms

First --- some Background

Learning Communication Protocols

Multi-Agent Reinforcement Learning Formalism

Background: Learning a Communication Protocol



Notation:

Agent Population P
Set of Comm Goals G
Message Space M

Communication Protocol Learning Objective:

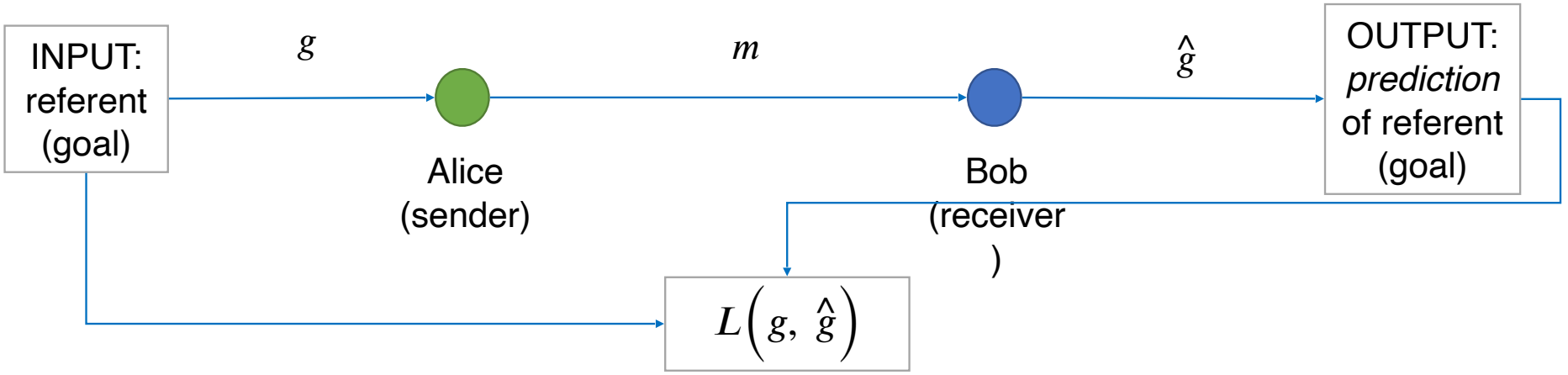
For all agents $i \in P$:
Learn agent policy $\pi^i : G \rightarrow M$

Multi-Agent *Signaling* Game [Lewis, 1969]

Background: **Referential Games** for Protocol Learning

Lazaridou et al, ICLR 2017

Premise of Game: Sender agent given referent (communicative goal) to generate a message for.
Receiver agent must infer what referent (communicative goal) sender is signaling.



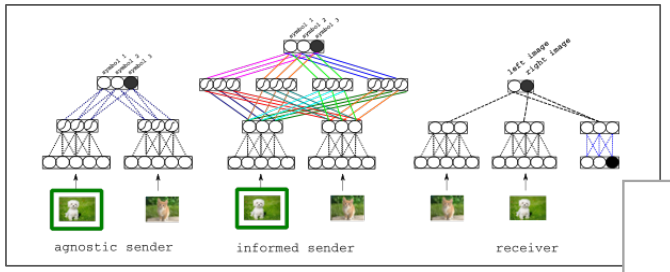
Two Important Assumptions for **Typical** Problem Setting

(1) Discrete Communication Channel:

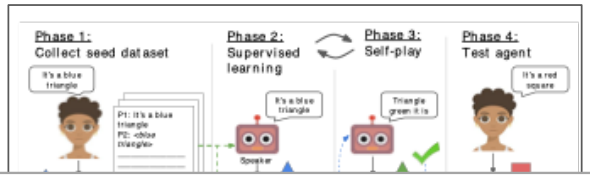
$$\langle m := a \mid a \in A \rangle$$

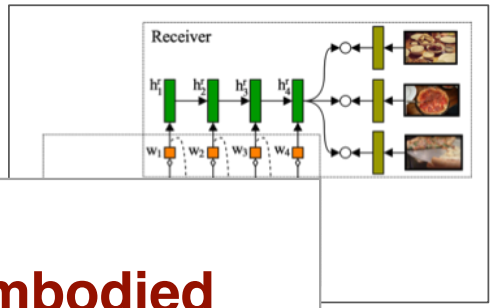
(2) Cheap Talk Setting: $\text{cost}(m) = \text{cost}(g) = 0$

Related Work – Learning Multi-Agent Communication

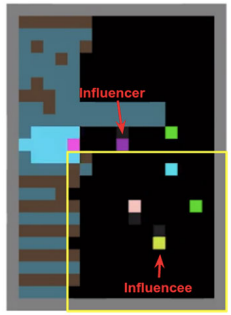


Lazaridou et al, ICLR 2017





, NIPS



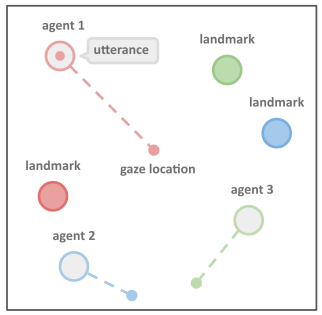
Jaques et al, ICML 2019

Key Limitation [for Embodied Agents]:

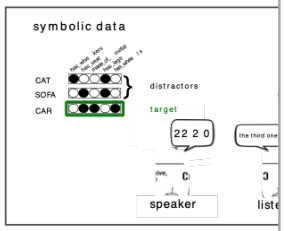
-- Focus on Symbolic (Discrete) Communication Channels

-- Assumes Cheap Talk Setting

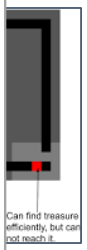
$$cost(message) = 0$$



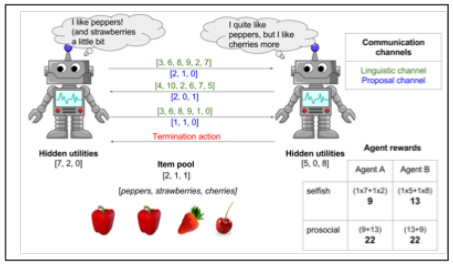
Mordatch & Abbeel, AAAI 2018



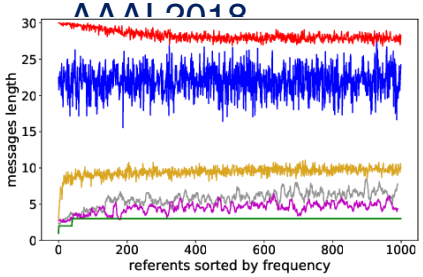
Lazaridou et a 2018



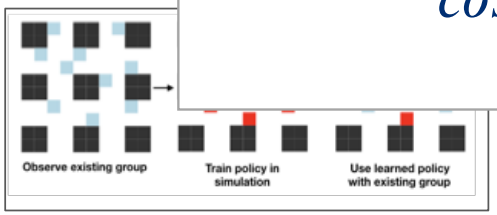
urIPS



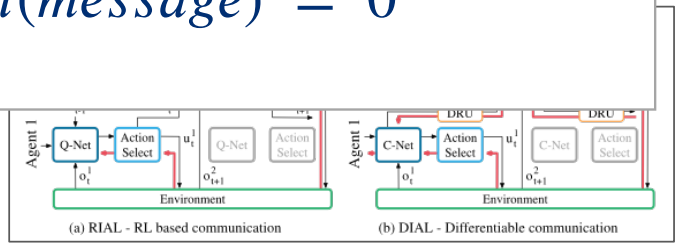
Cao et al, ICLR 2018



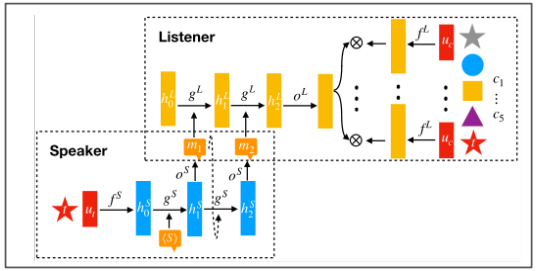
Chaabouni et al, NeurIPS 2019



Lerer & Peysakhovich, AIES 2019



Foerster et al, NIPS 2016



Li & Bowling, NeurIPS 2019

Problem Formalism (*Fully* Cooperative)

Assumptions

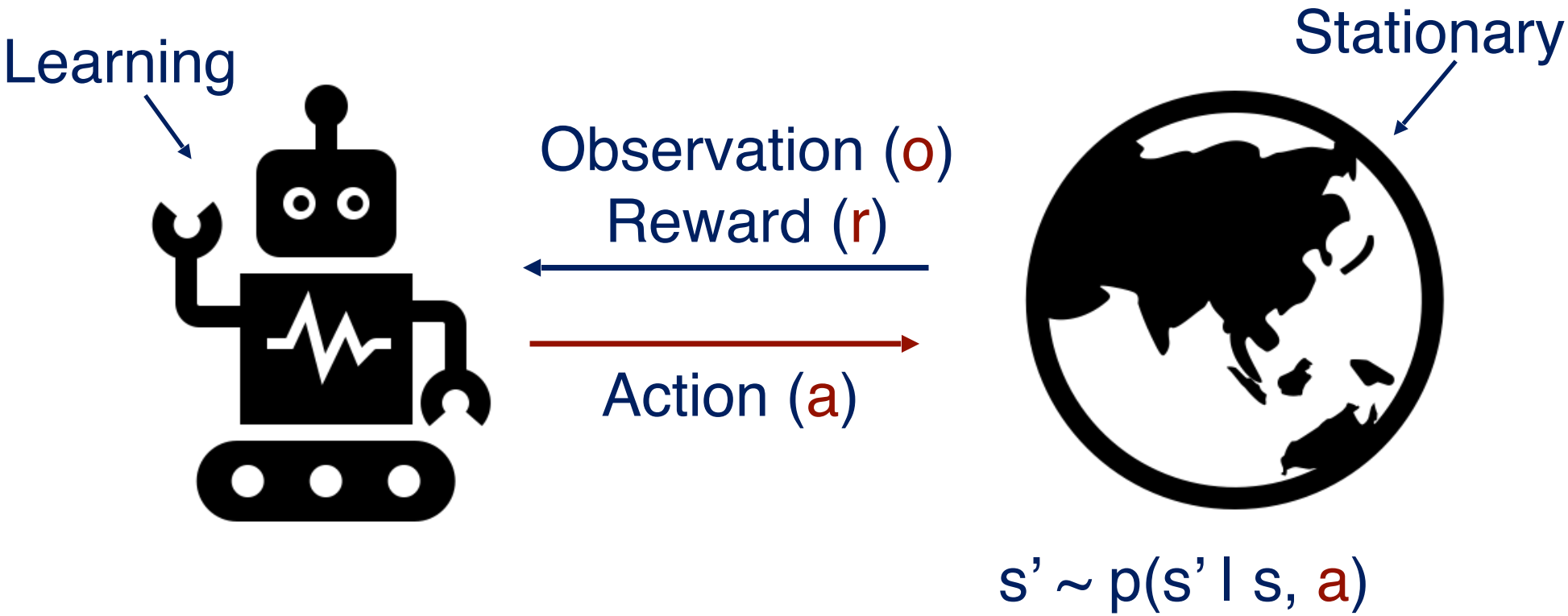
- Communicative Goals given a priori
- Protocol *emerges solely* through agent-agent interactions
  - *No* data or supervisory signal

Formalized as Decentralized Partially Observable Markov Decision Process (Dec-POMDP)

Berstein et al, MOOR 2002

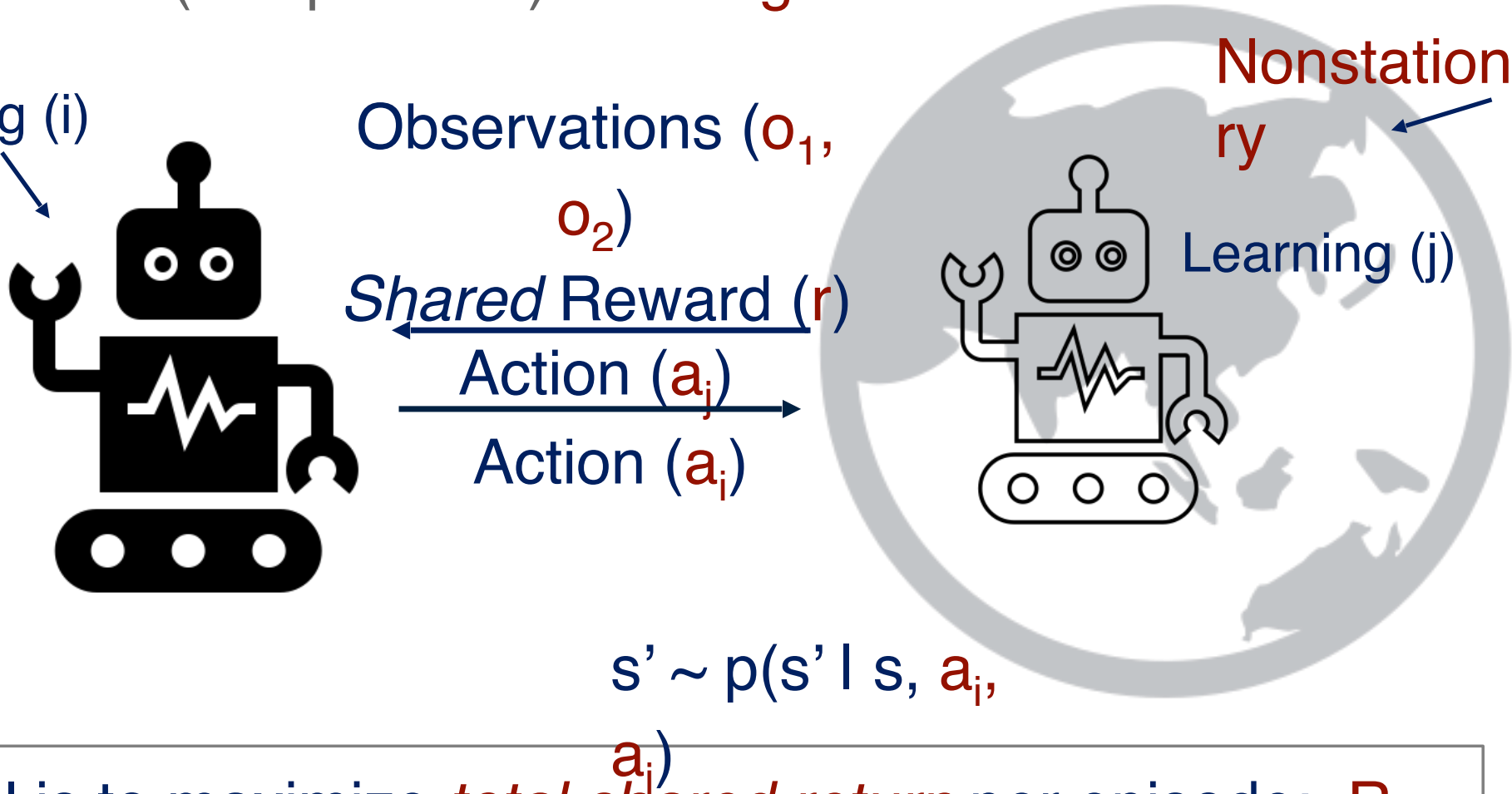
- $N = \{1, \ldots, n\}$ set of n agents (here $n = 2$)
- $S =$ (finite) set of states
- $\boldsymbol{O} =$ set of joint observations (*private* individual observations)
- $\boldsymbol{A} =$ set of joint (communication) actions
- T: $S \times \boldsymbol{A} \times S \rightarrow \mathbb{R}$:= transition probability function
- $R : S \times \boldsymbol{A} \rightarrow \mathbb{R}$:= *shared* reward for all agents
- $G :=$ set of (communicative) goals

Background: Reinforcement Learning



Goal is to maximize *total return* per episode: $R = \sum_t \gamma^t r_t$

Background: (Cooperative) Multi-Agent Reinforcement Learning



Goal is to maximize *total shared return* per episode: $R = \sum_t \gamma^t r_t$

Background: Multiagent Training Regimes

- Centralized Training
  - Agents *can* share information and provide feedback to each other
- Decentralized Training
  - Agents *cannot* share information nor provide feedback to each other
- Decentralized Execution (Inference Time)
  - Agents can use only *own* sensory input and learned policy to make decisions

Problem Setting

Prior: Zero-Shot Coordination

Introduce: Zero-Shot Communication

Problem Setting: Zero-Shot Coordination

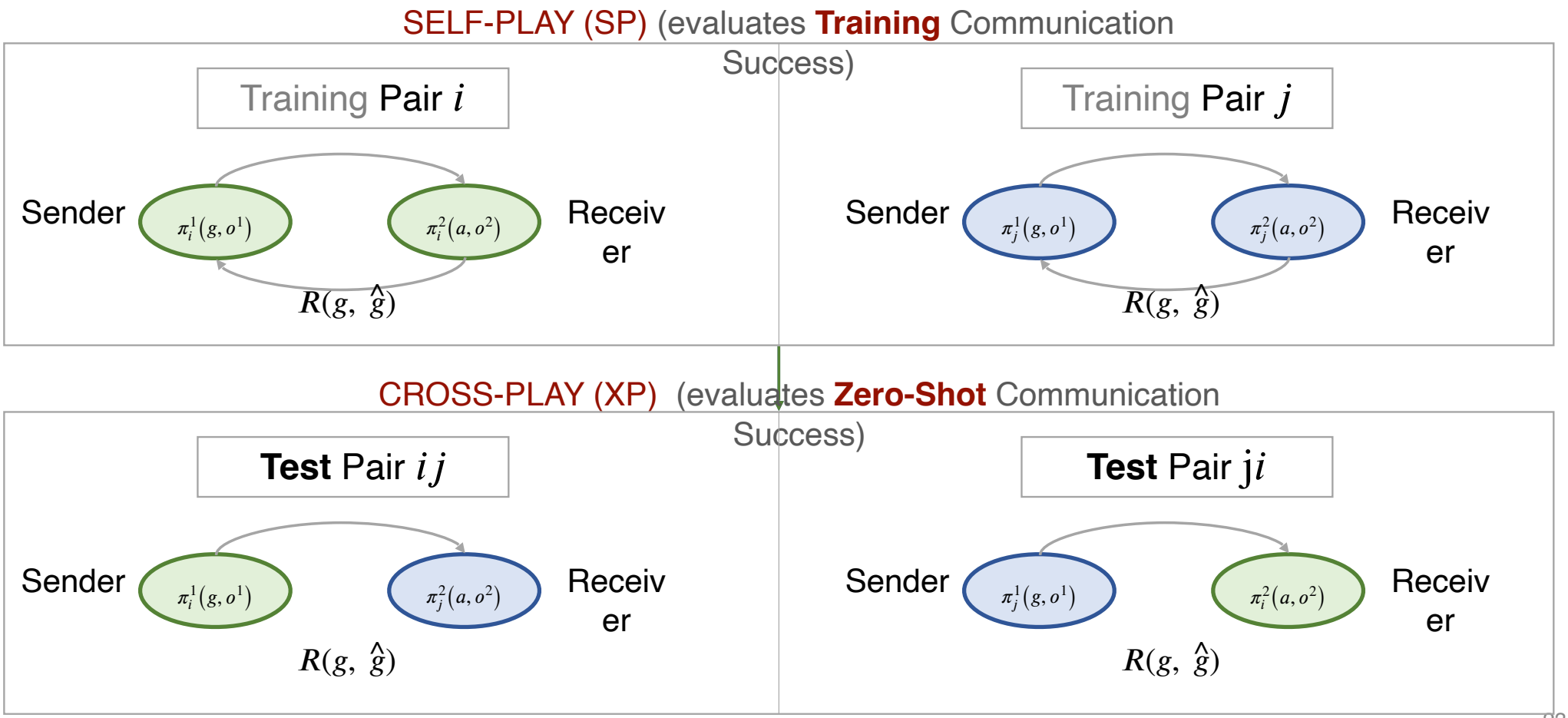
Hu et al, ICML 2020

- Zero-Shot Coordination: Agents must **coordinate** at test time with other agents who have been independently trained
- Introduce Equivalence Mappings (Symmetries) Φ of Policies
  - Bijective Mappings
- Formalize Symmetries for each element of underlying Dec-POMDP
  - States: S
  - Joint Actions: $\boldsymbol{A}$
  - Transition Model: $P(s' \mid s, \boldsymbol{a})$
  - Reward Function: $R(s, \boldsymbol{a}, s')$
  - Observation Function: $O(o \mid s, \boldsymbol{a}, i)$

[Bullard et al, arXiv 2021]

Formalize Problem Setting: **Zero-Shot** Communication

Coordination Goal: Learn communication protocol that effectively generalizes to *unseen* agents

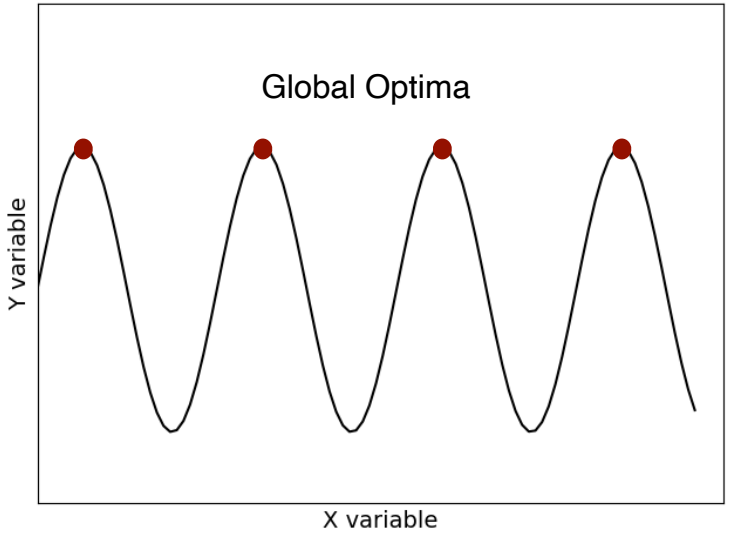


Problem with Training using **Self-Play**

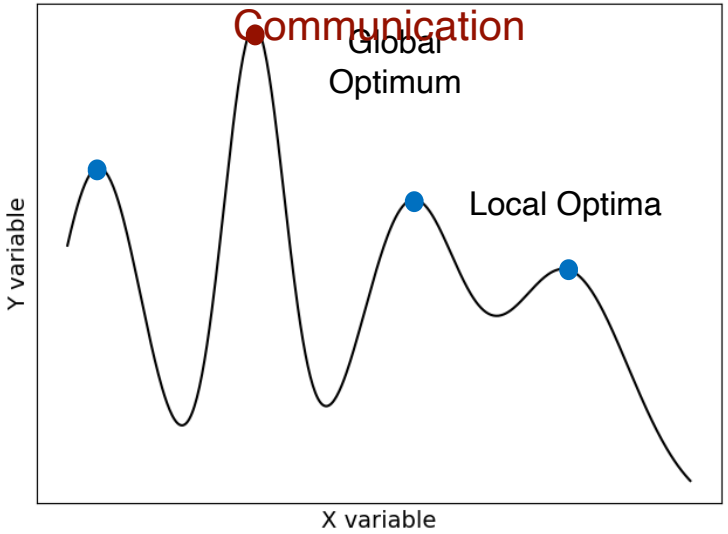
Self-Play (SP): MANY *Equivalent* but *Incompatible* Joint Policies (Protocols)

Successful Coordination := ***Agreement*** on Protocol (Convergence on SAME Optimum)

Typical with SP



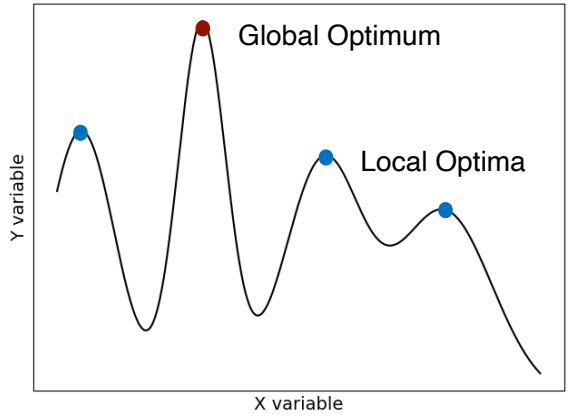
*What we **WANT:*** For ZS Communication



Our General Approach

Induce Bias in Protocol Learning with PRIORS on the Solution Space

- Use **Real-World Constraints** to engender a *Unique* Globally Optimal Protocol
  (Under ZSC Problem Setting)

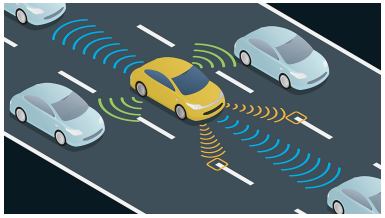


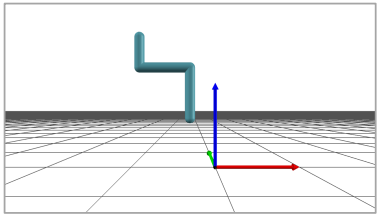
Adapt Learning Objective to Learn Distribution over **Equivalent** Global Optima

- Exploit **Equivalence Classes** of Communicative Actions

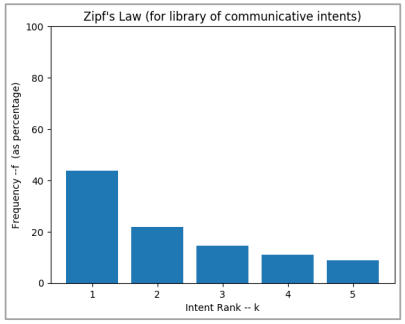
Inducing Bias: **Priors** from Real-World Settings

- Cost of Physical Energy Exertion
  - **Non**-Cheap Talk Setting
  - Communication is *Costly*





- Nonuniform Distribution over Words (and Communicative Goals)

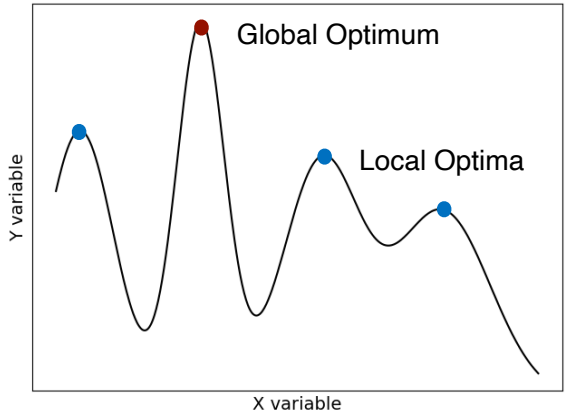


Zipfian Distribution

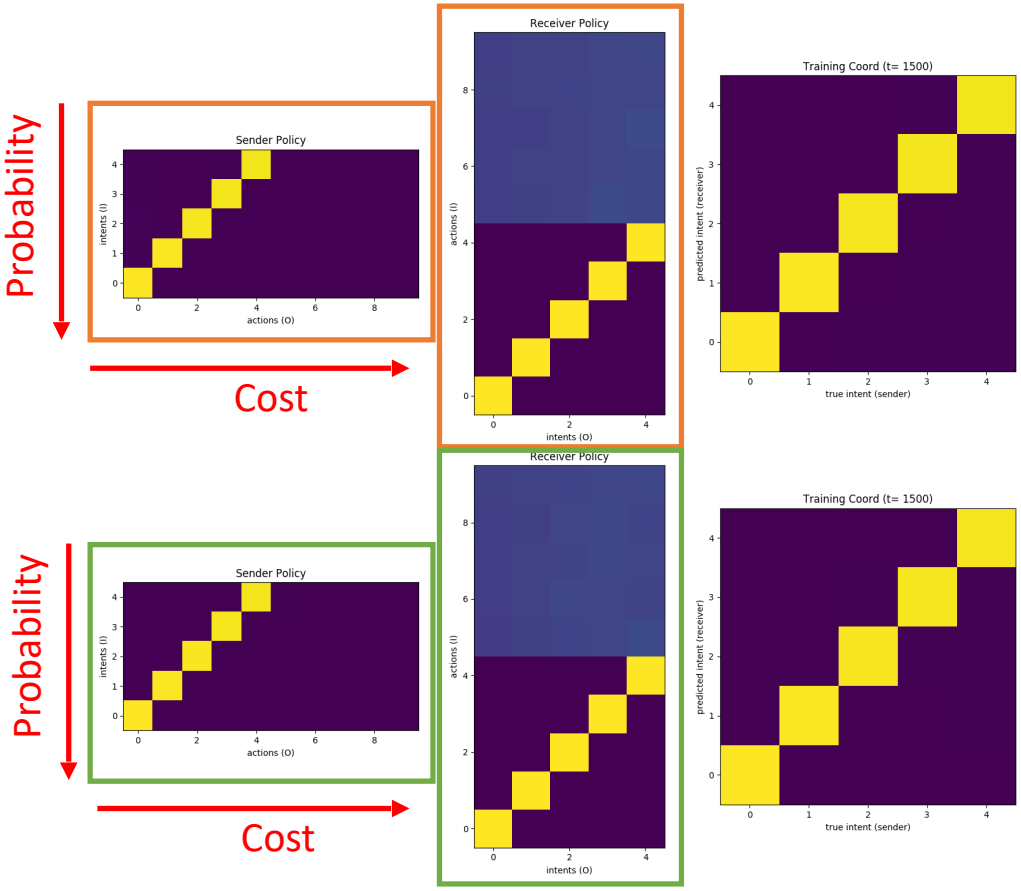
Energy

Zipf

Incentivizes **Minimal Effort** Protocol

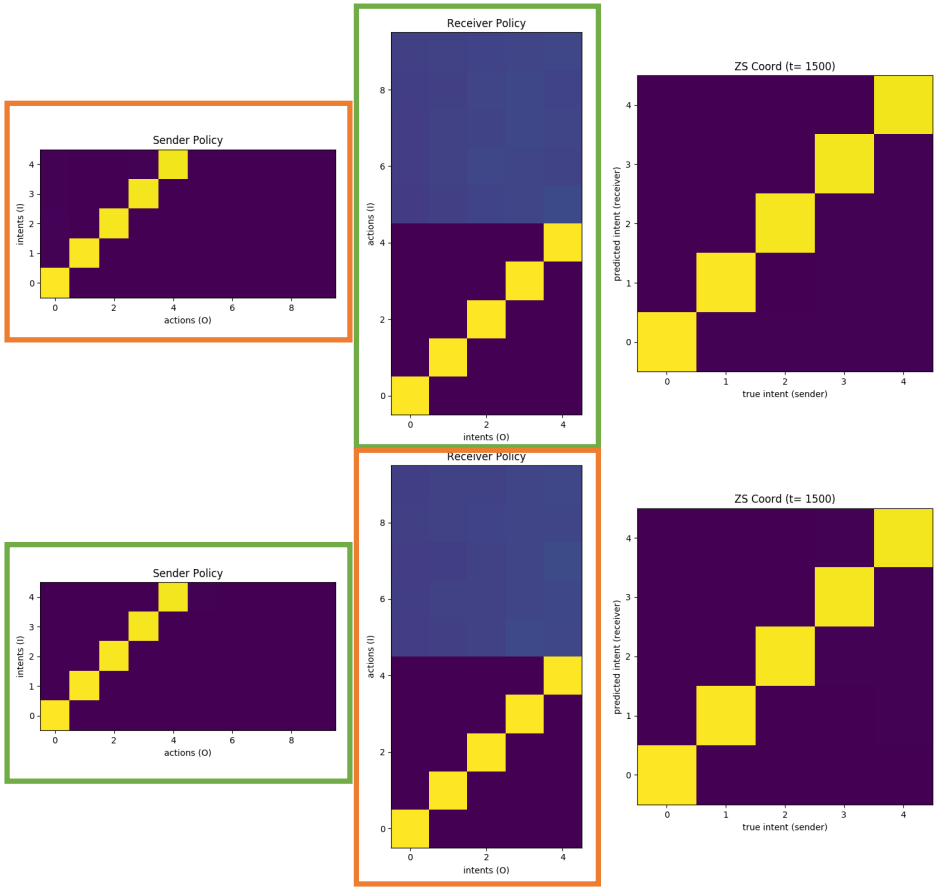


[George Zipf, *Human Behaviour and Principle of Least Effort*, 1949]



SP (Training Coordination)

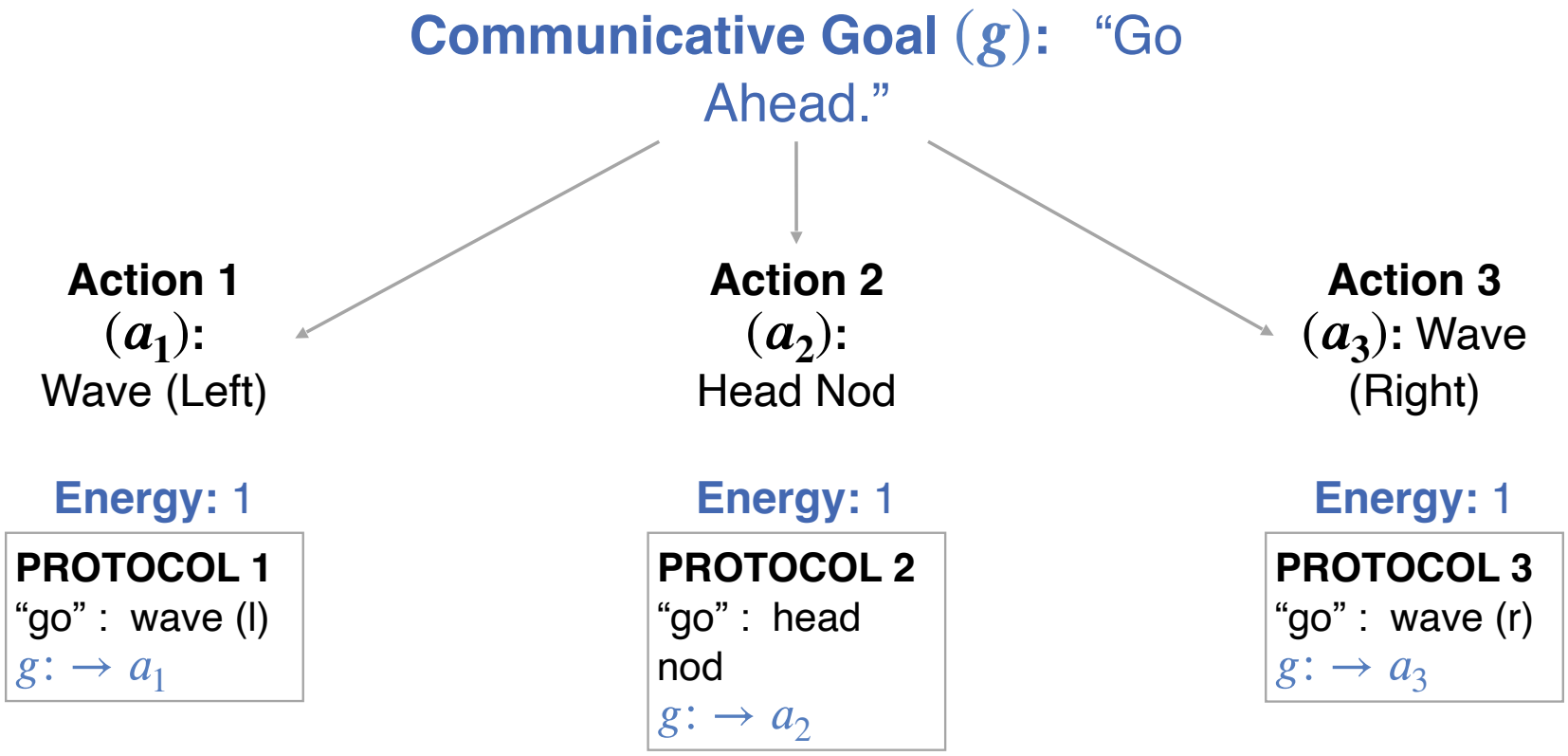
0.98 +- 3.7e-04



XP (ZS Coordination)

0.98 +- 3.4e-04

What about **Symmetries** in Realistic Communication?



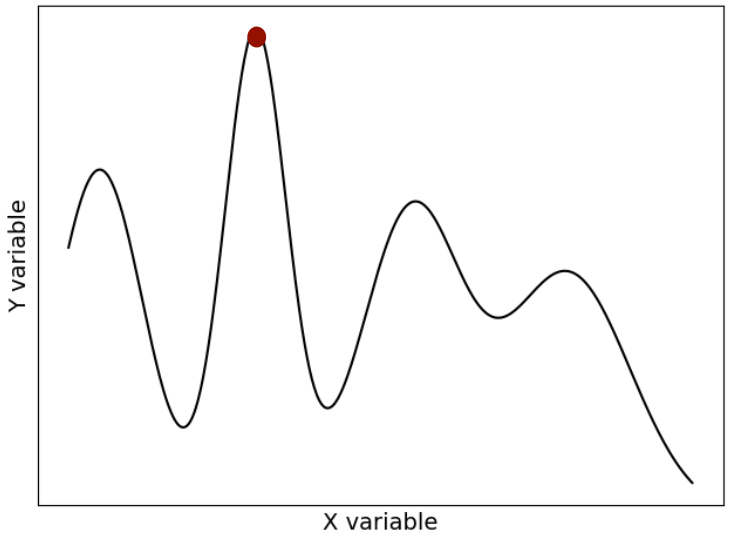
Challenge: All Equivalent (Minimal Effort) but INCOMPATIBLE Policies

How does **Symmetry** impact the *Optimization* Problem?

*Focus on **Action Symmetry** in this Work

CASE 1: **NO** Action or Goal Redundancy

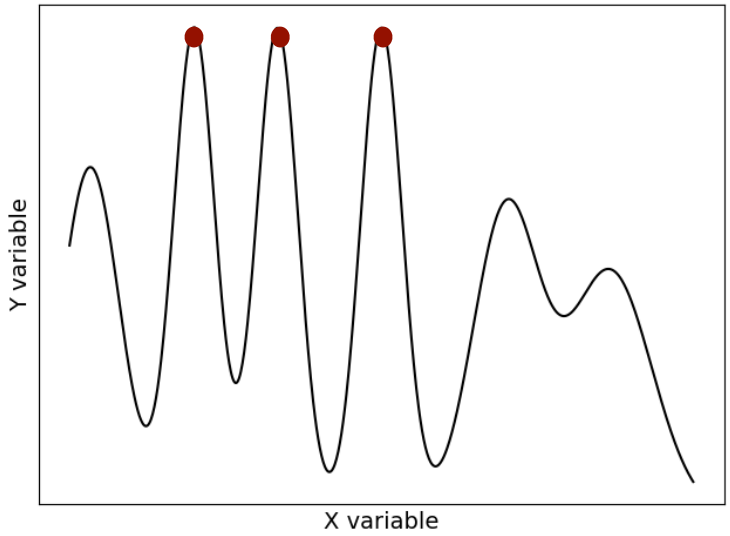
Every Action == *Distinct* Cost



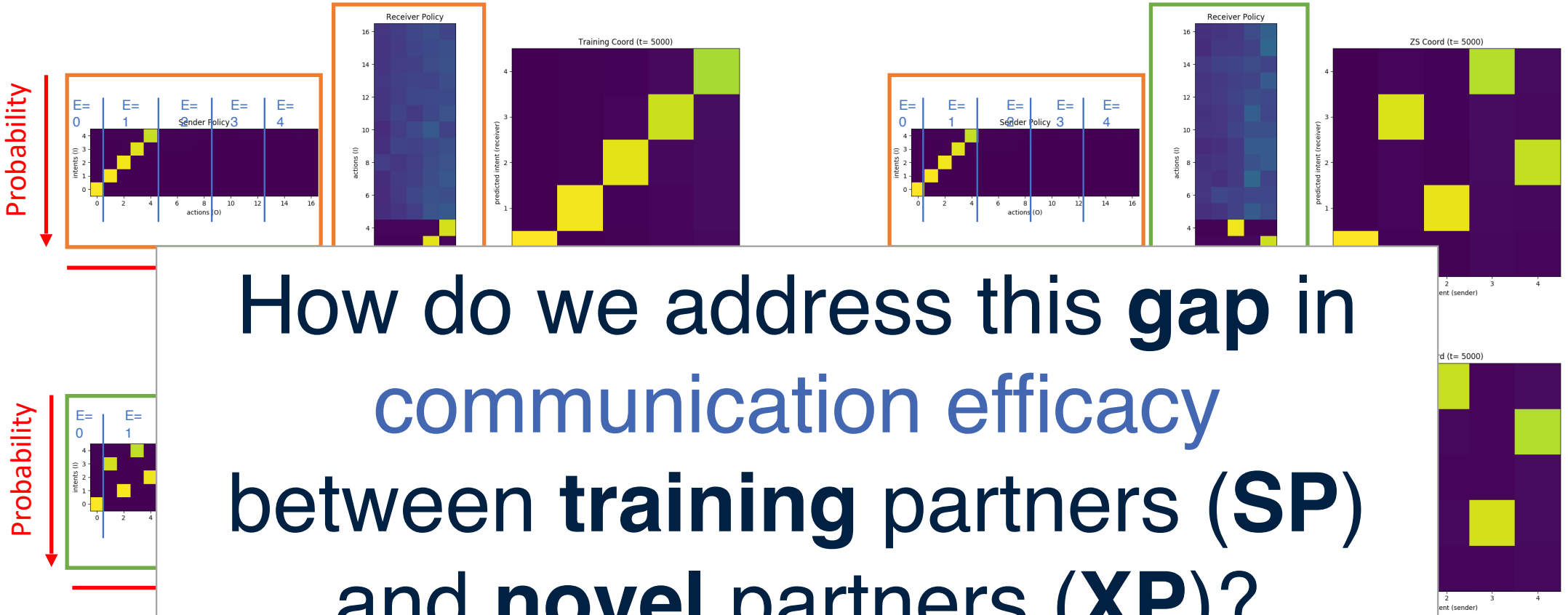
UNIQUE Minimal Effort Protocol

CASE 2: **Symmetry** in Action or Goal Space

Many Actions == *Same* Cost



MULTIPLE Minimal Effort Protocols



How do we address this **gap** in communication efficacy between **training** partners (**SP**) and **novel** partners (**XP**)?

SP (Training Coordination)

96% +- 0.14

XP (ZS Coordination)

45% +- 0.02

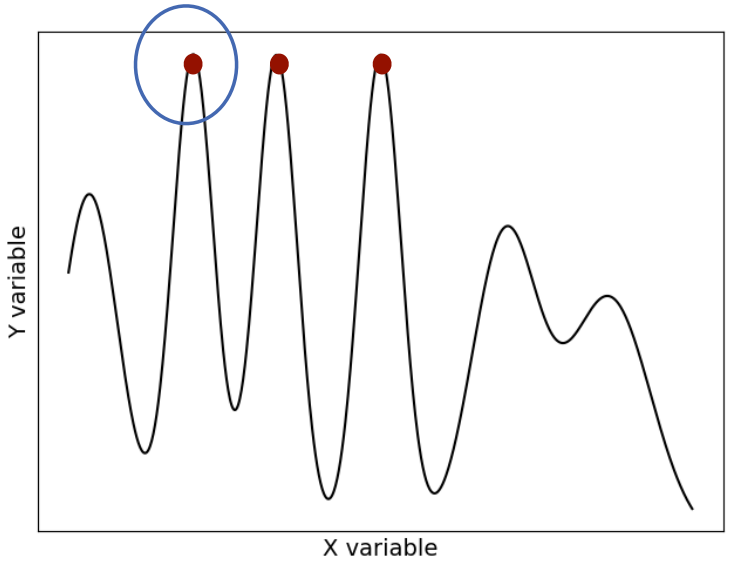
Method

Quasi-Equivalence Discovery (QED)

Baseline Training using SP

$$\pi^*_{SP} = \underset{\pi}{\text{argmax}}\ J(\pi^1, \pi^2)$$

Self Play (SP)

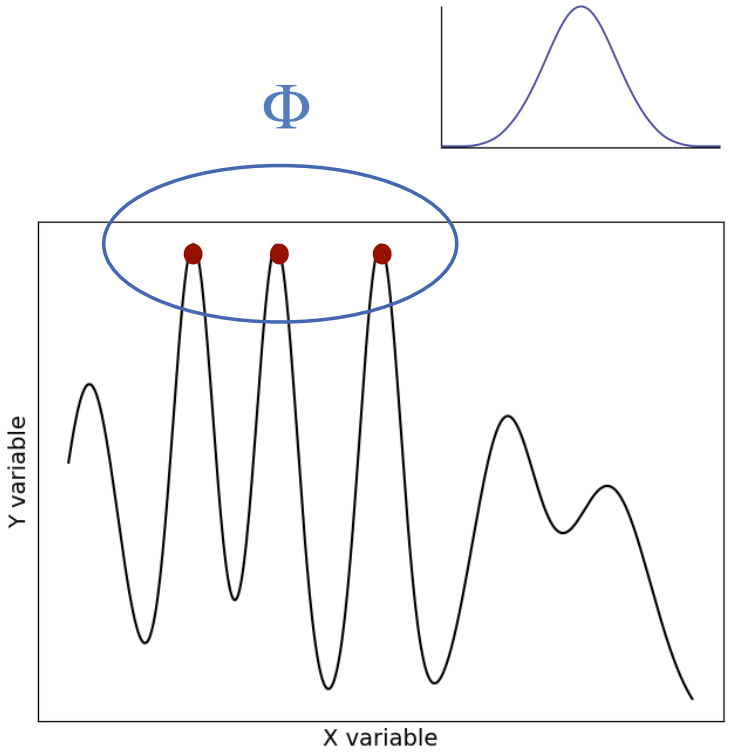


MULTIPLE Minimal Effort Protocols

Method: Exploit **Equivalence Mappings** using OP

$$\pi^*_{OP} = \underset{\pi}{\text{argmax}} \mathbb{E}_{\phi \sim \Phi} \; J\left(\pi^1, \phi\left(\pi^2\right)\right)$$

Other Play (OP) [Hu et al, ICML 2020)]



$\Phi :=$ equivalence mappings

MULTIPLE Minimal Effort Protocols

Method: Objective Function Analysis

[Bullard et al, arXiv 2021]

$$\pi^*_{OP} = \underset{\pi}{\text{argmax}}\ \mathbb{E}_{\phi\sim\Phi}\ J\left(\pi^1,\ \phi(\pi^2)\right)$$

Communication Efficacy — Action Cost

$$\mathbb{E}_{\phi\sim\Phi}\ J\left(\pi^1,\ \phi(\pi^2)\right) = \mathbb{E}_{\phi\sim\Phi,\ g\sim G}\ \log p_\pi(g|g) - \text{Cost}(\pi)$$

$$= \mathbb{E}_{\phi\sim\Phi} \sum_g \left[p(g)\log p_\pi(g|g)\right] - \sum_g C(a)\,\pi^1(a|g)\,p(g)$$

$$= \sum_{g,\phi} p(g,\phi)\log\left[\sum_a \pi^2(g|\phi(a))\ \pi^1(a|g)\right] - \sum_g C(a)\,\pi^1(a|g)\,p(g)$$

$$\geq \sum_{g,\phi,a}\left(p(g,\phi)\log\left[\pi^2(g|\phi(a))\ \pi^1(a|g)\right]\right) - \mathbb{E}_{a\sim\pi}\left[C(a)\right]$$

$$= \sum_{g,\phi,a} p(g,\phi)\log\pi^2(g|\phi(a)) + \sum_{g,a} p(g)\log\pi^1(a|g) - \mathbb{E}_{a\sim\pi}\left[C(a)\right]$$

$$\propto \sum_a I_\pi(G;\Phi(a)) + \mathbb{E}_G\left[\sum_a \log\pi^1(a|g)\right] - \mathbb{E}_{a\sim\pi}\left[C(a)\right]$$

(learned distributions)

π^1 := sender policy

π^2 := receiver policy

$\pi := \langle \pi^1,\ \pi^2 \rangle$

How can we **obtain** these *Equivalence Mappings*?

$\max_{g,\phi}$ Mutual Information

$\min_a$ Effort

[Bullard et al, arXiv 2021]

Method: **Learning Algorithm** inferring Equivalence Mappings

Other-Play (OP): Key Limitation

- **Assumes** equivalence mappings (symmetries) are given

Quasi-Equivalence Discovery (QED)

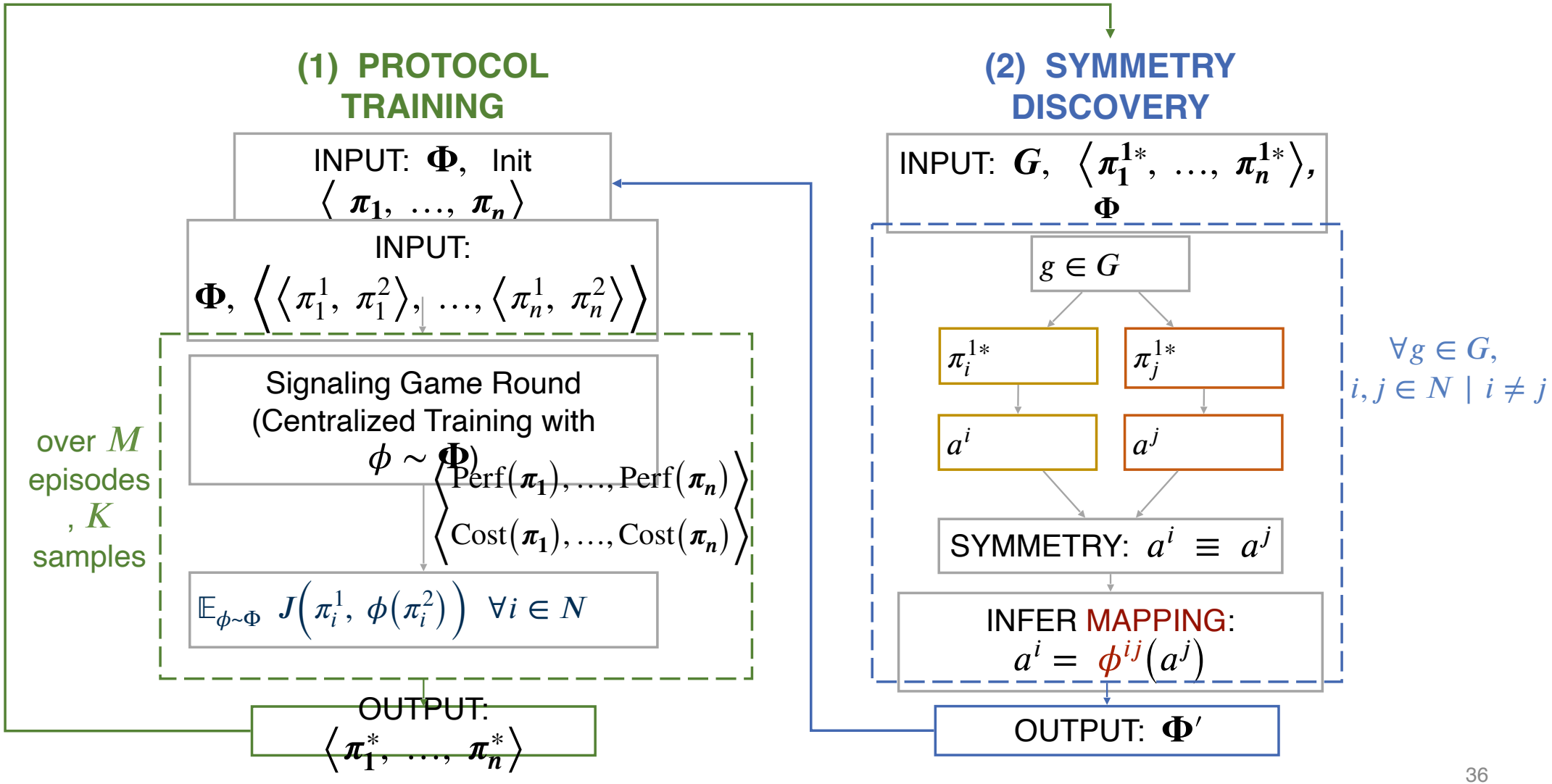
- **Introduce** Iterative Algorithm
- Automatically discovers symmetries

QED Algorithm Overview

- Initialize set of Equivalence Mappings with the *Identity Mapping*
- For each Iteration: [until Convergence]
  - **PROTOCOL TRAINING:** Train new set of optimal ZSC joint policies **given** Φ
    --> using $\mathbb{E}_{\phi \sim \Phi} \; J\left(\pi^1, \phi\left(\pi^2\right)\right)$
  - **SYMMETRY DISCOVERY:** Extract new symmetries Φ from optimal policies

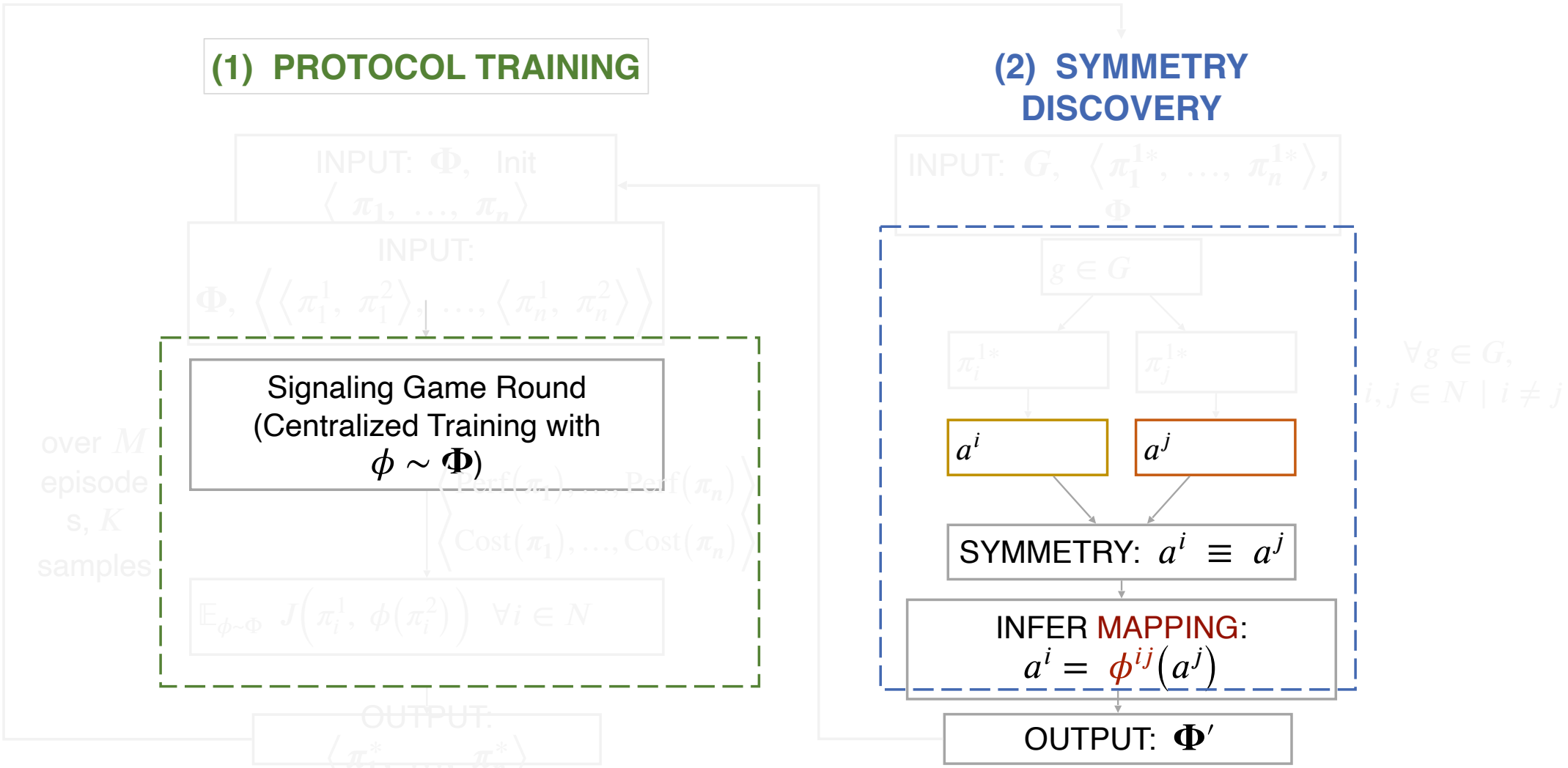
QED Method

Until **Convergence** …



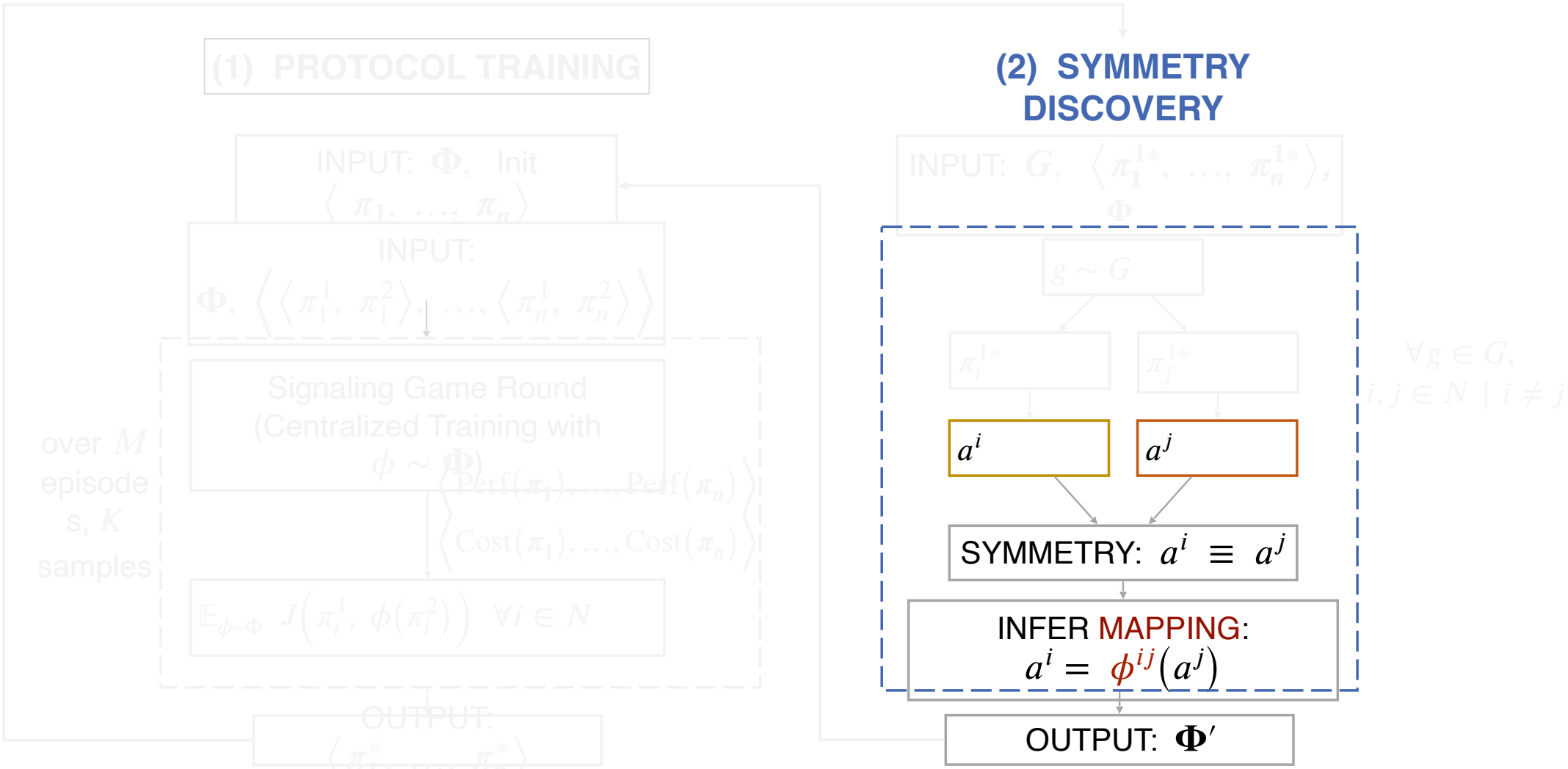
QED Method

Until **Convergence** …



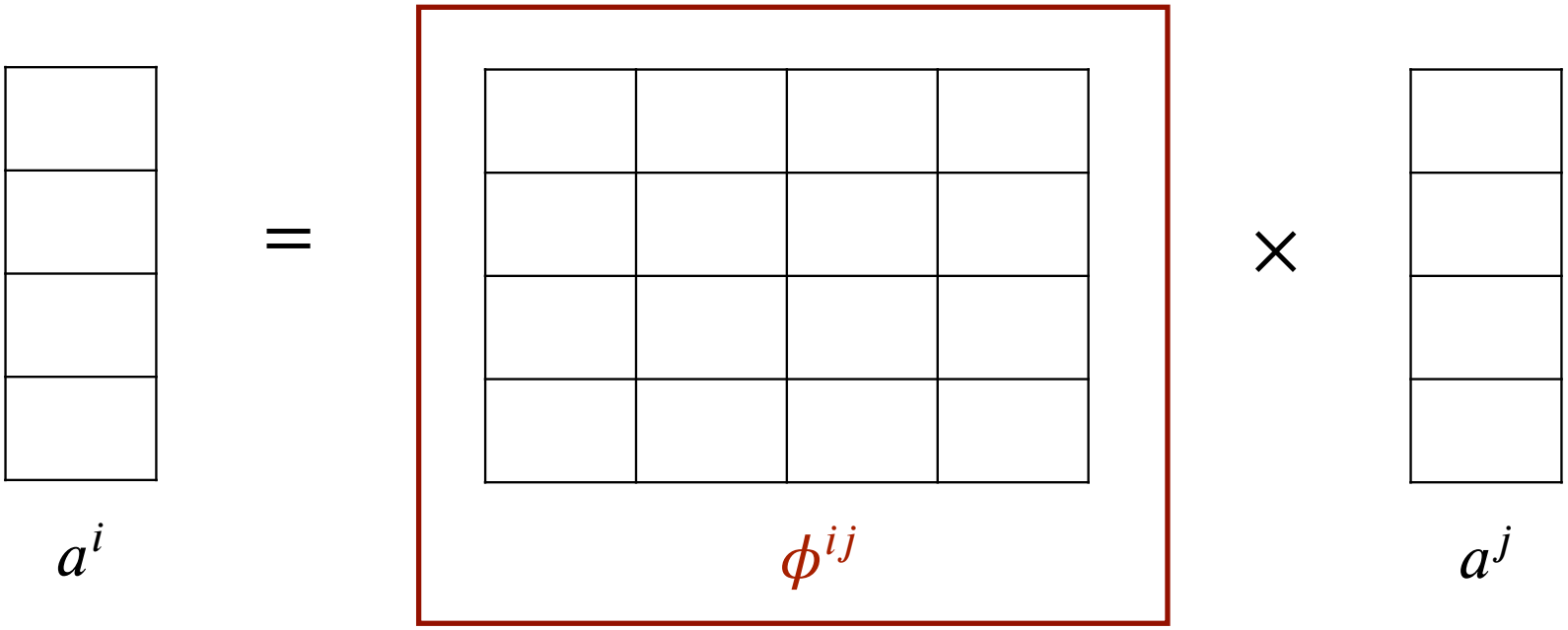
QED Method

Until **Convergence** …



QED Method

Inferring **Equivalence** Mappings (Symmetries)



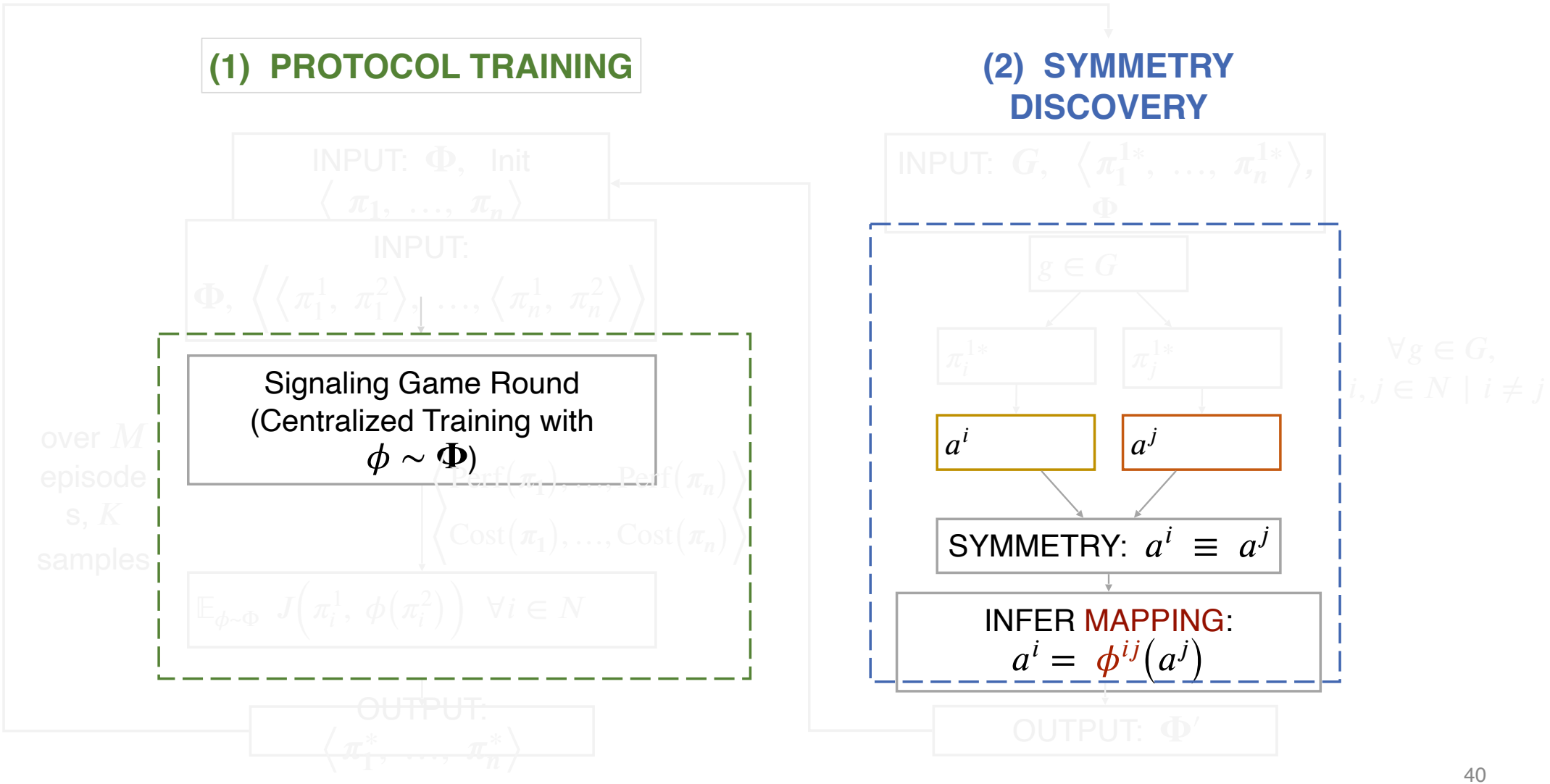
Learn: Equivalence Mapping ϕ := Transformation Matrix

Minimize

$$\mathrm{KL}\left(a^i,\ \phi^{ij}\left(a^j\right)\right) \longmapsto \Phi\colon \Phi + \left\{\phi^{ij}\right\}$$

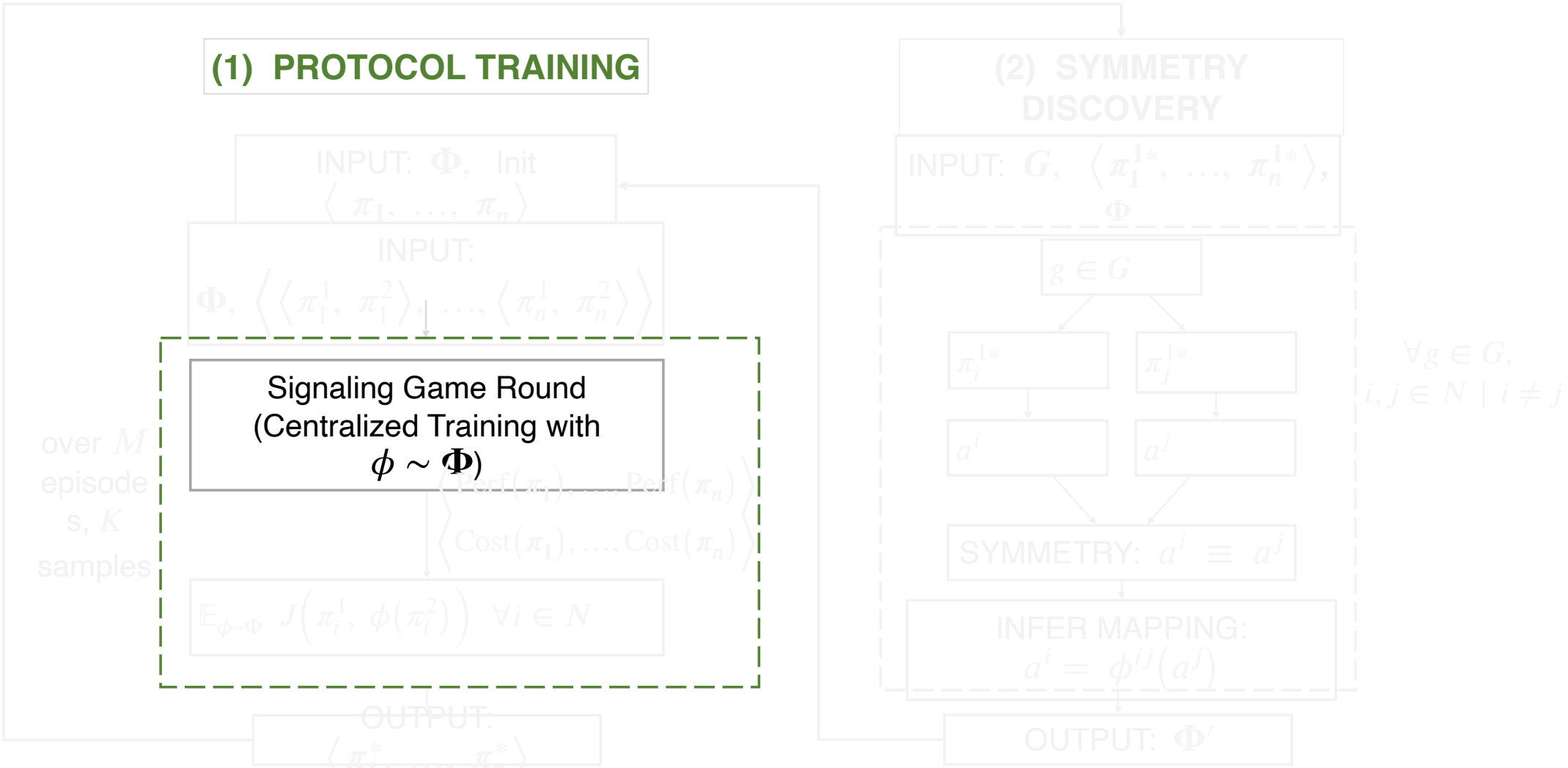
QED Method

Until **Convergence** …

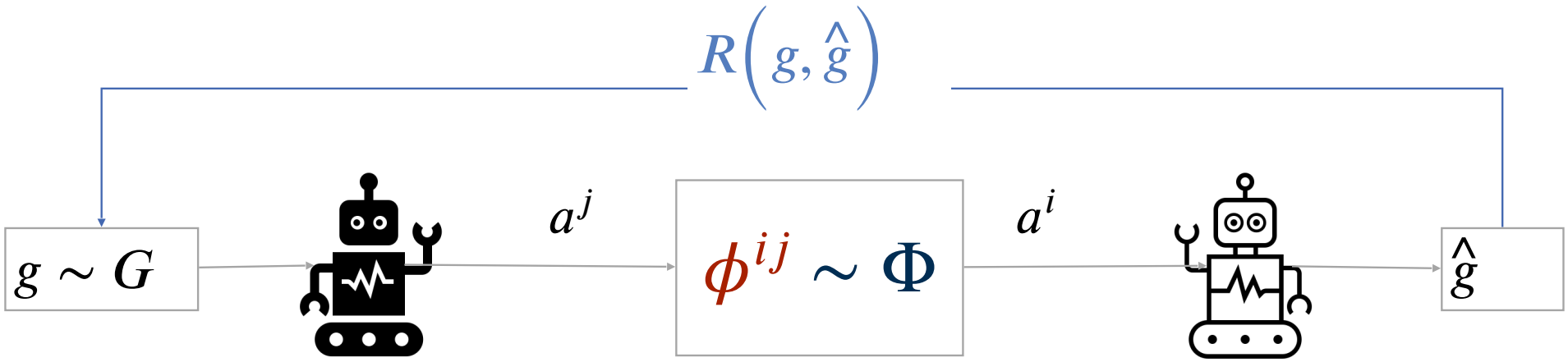


QED Method

Until **Convergence** …



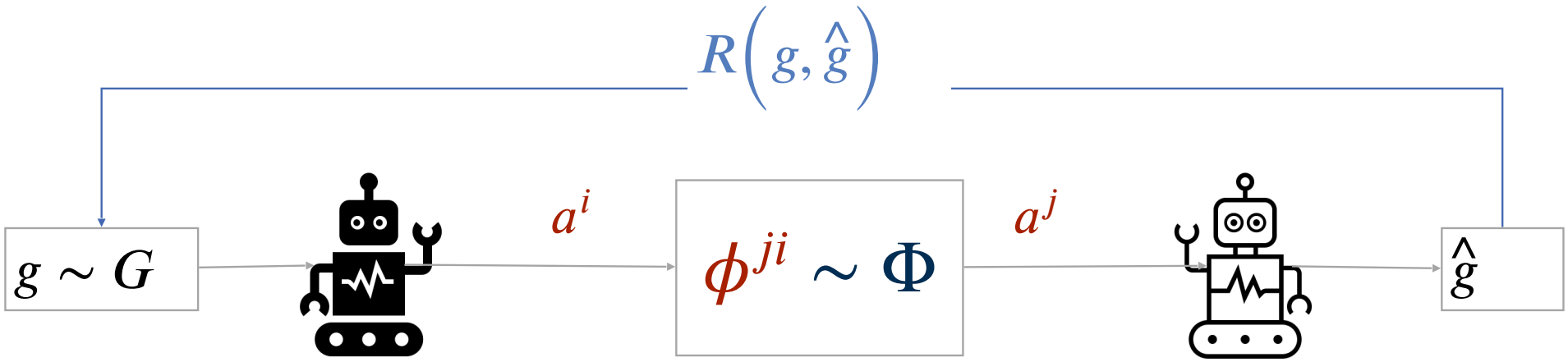
Playing the Signaling Game with **Symmetries**



If $\hat{g} == g$:

- **Sender** learns
  $g \longmapsto a^j$
- **Receiver** learns
  i

In a LATER Game… (when relationship is reversed)



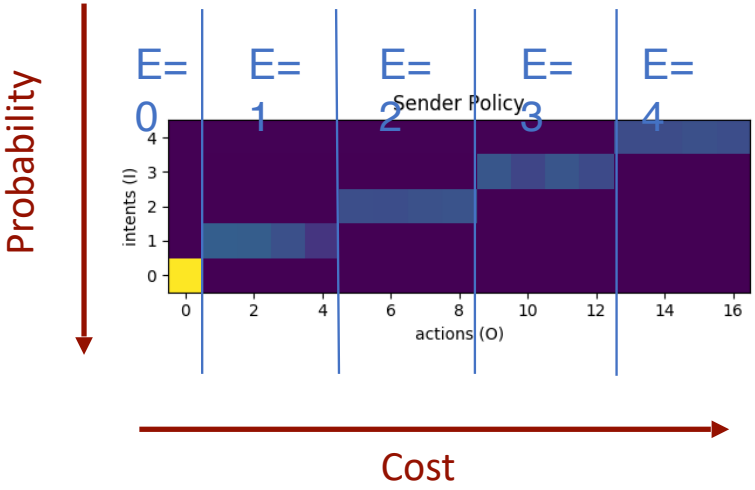
If $\hat{g} == g$:

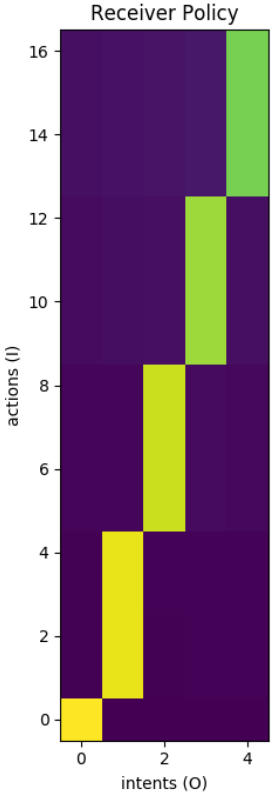
- **Sender** updates learning
  $g \longmapsto \{a^j, a^i\}$
- **Receiver** updates learning

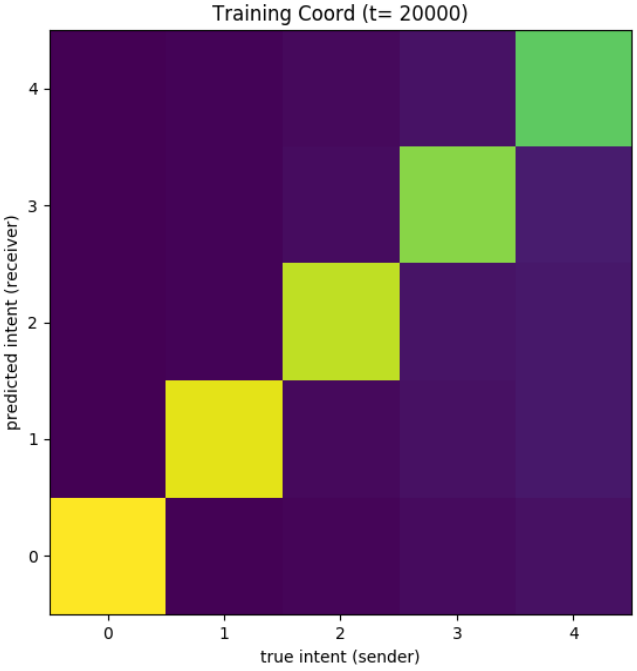
Some Empirical Results…

Domain: Discrete Signaling Game (the same one)

Optimal Sender and Receiver Policies (**QED**) – *unique!*







More Complex Task (Energy Degeneracy)

BASELINE:
SP (Zipf + energy)

SP Perf (Training):
96.2% +- 0.12

XP Perf (ZSC):
58.5% +- 11.12

Max Class
~44%

Energy Co
0.59 +- 0.

EXPERIMENTAL:
QED (Zipf + energy)

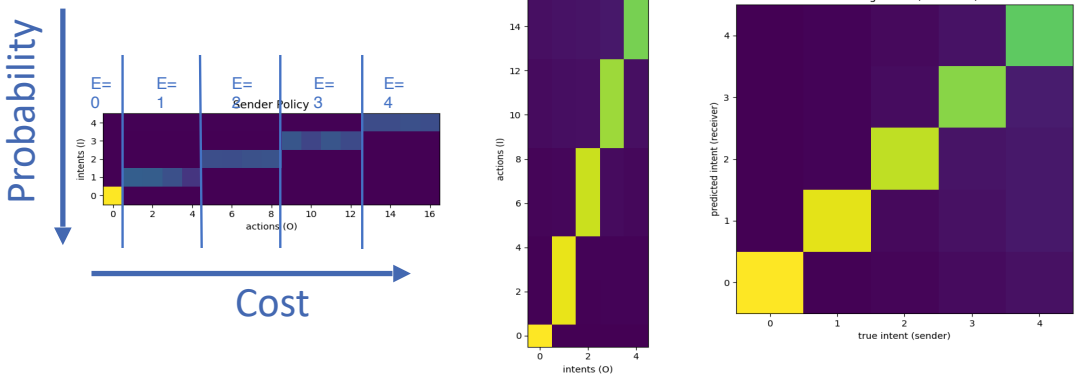
SP Perf (Training):
93.4% +- 0.03

XP Perf (ZSC):
93.4% +- 0.02

Increased Communication Performance with **Novel** Agents (ZSC) by **~60%**

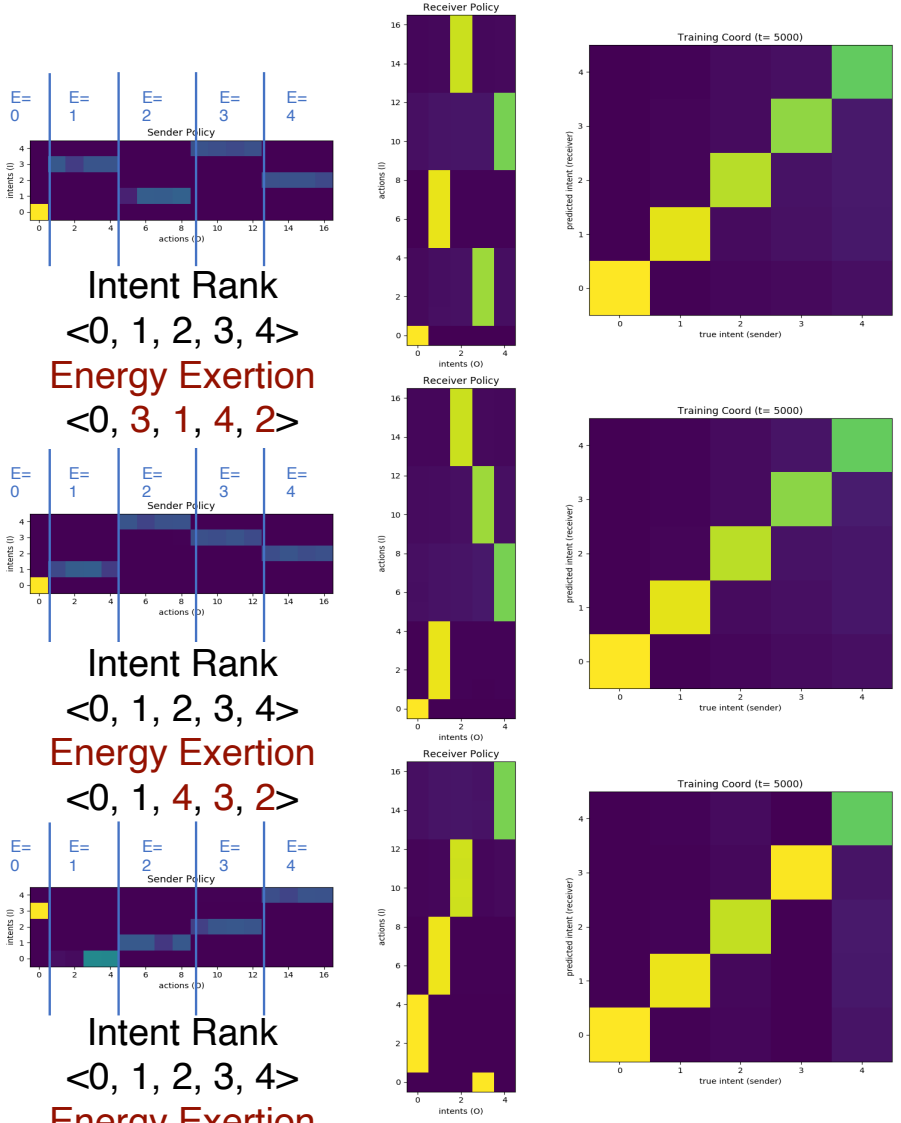
Open Challenge:

Prevalence of Converging on *Suboptimal* Protocols



Intent Rank
<0, 1, 2, 3, 4>
Energy Exertion
<0, 1, 2, 3, 4>

OPTIMAL



Intent Rank
<0, 1, 2, 3, 4>
Energy Exertion
<0, 3, 1, 4, 2>

Intent Rank
<0, 1, 2, 3, 4>
Energy Exertion
<0, 1, 4, 3, 2>

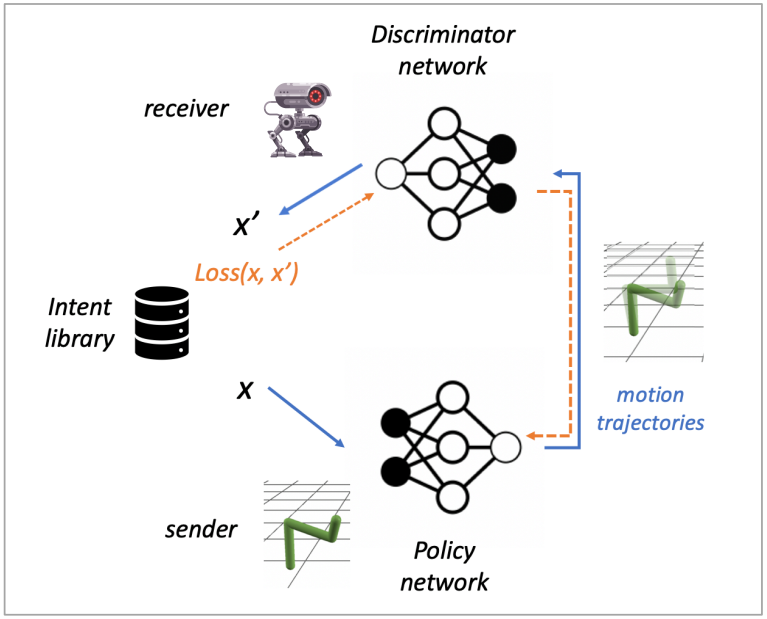
Intent Rank
<0, 1, 2, 3, 4>
Energy Exertion
<3, 0, 1, 2, 4>

SUBOPTIMAL

What about the **Continuous Channel** Setting?

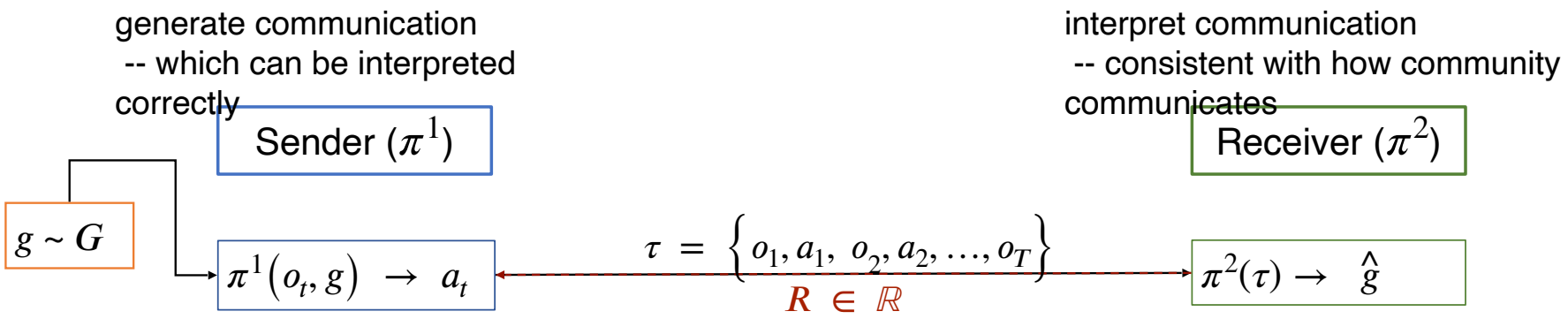
[Bullard et al, NeurIPS Deep RL & EmeComm Workshops 2020]

- **Underexplored** in Multi-Agent Communication!
- Message == Motion Trajectory $\{\tau\}$
  - High-Dimensional Continuous State & Action Spaces
  - Continuous Energy Values $:= C(\tau) \in \mathbb{R}$
- Message Space becomes **Intractably Large**
- Method: SP Training + *Third-Party Receiver* Training + Induced Bias (Energy + Zipf)
- Evaluation: Is ZSC *possible* in **continuous** setting?



Embodied Referential Game

Method: **SP** Protocol Training



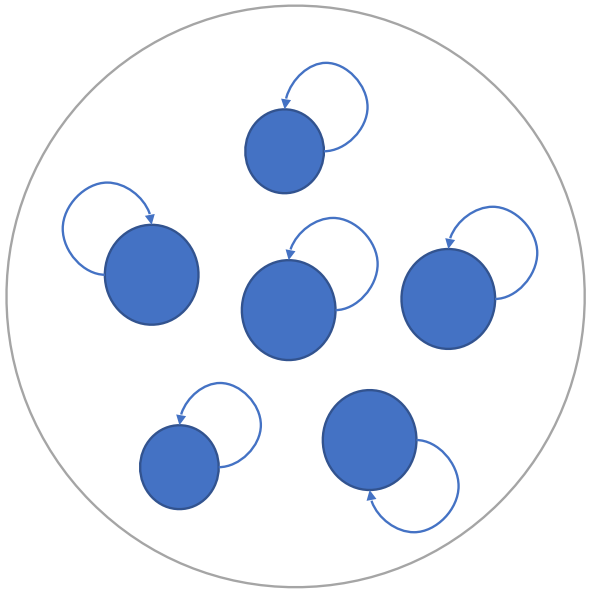
Shared Return

$$R = p(g) \log p(g|g) - \|I * (a_{2:T} - a_{1:T-1})\|_2^2$$

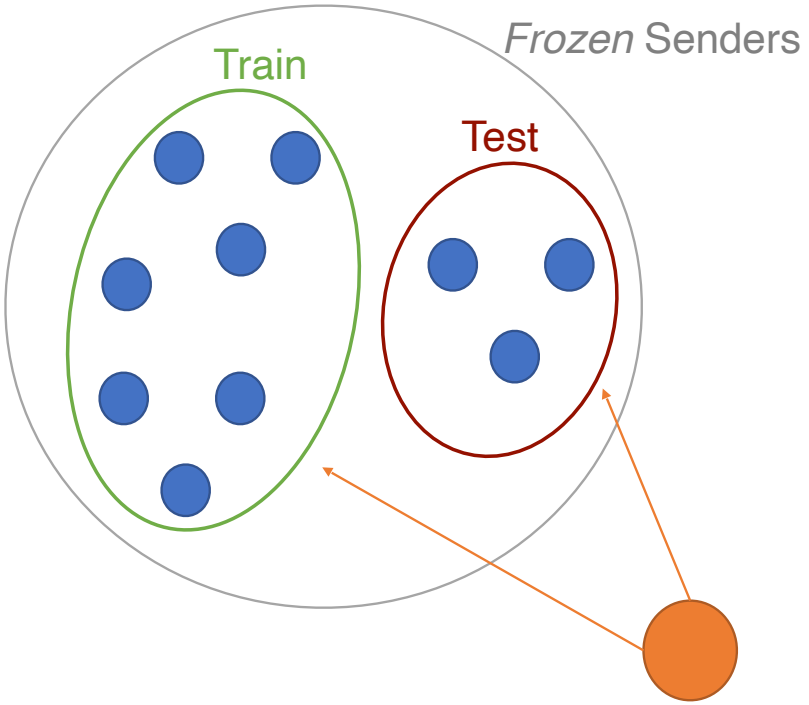
Communication Efficacy

Cost/ Energy Penalty

Method: SP Protocol + Third-Party Receiver Training



Self-Play



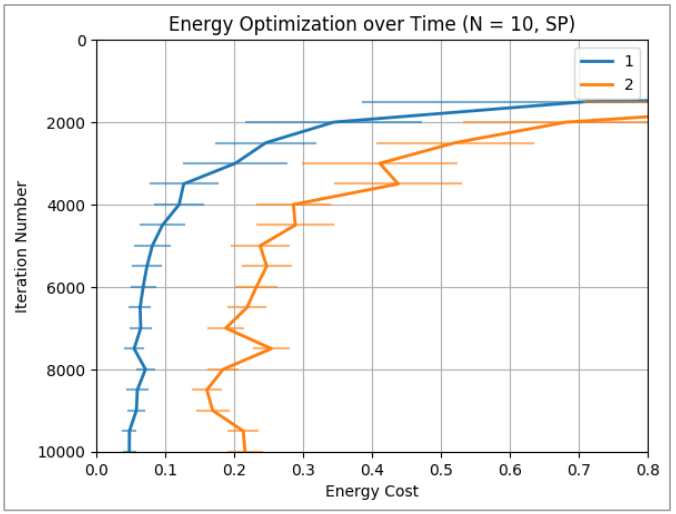
Protocol
Training

Third-Party Receiver
Training + ZS
Communication

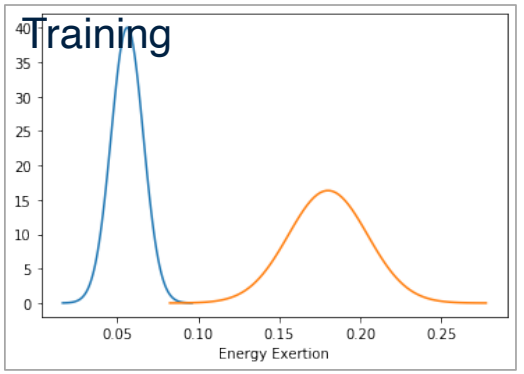
Qualitative Analysis
(*N=2* Intents Task)

Key Finding: Energy of Message *sufficient* for Prediction of Intent

ZSC possible *in principle* for continuous-action protocols!



At END of Protocol Training

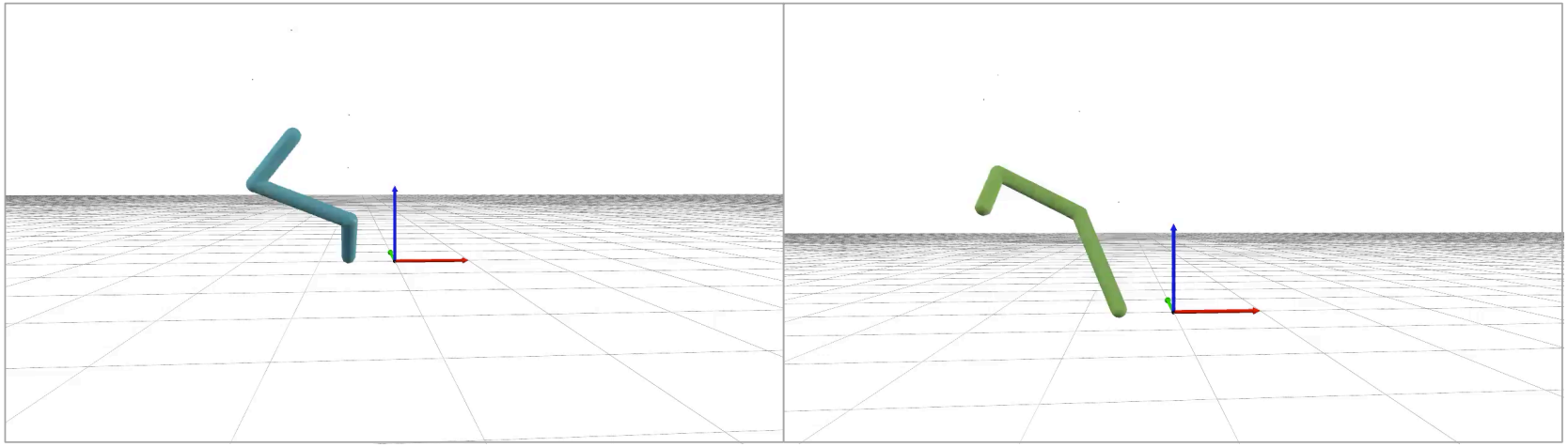


Achieves $I(G, C(\tau)) \gg 0$!

ZSC Performance (*Test* Input)

0.75	**0.97**

Max Class: ~ 0.67



Communicative Intent 1

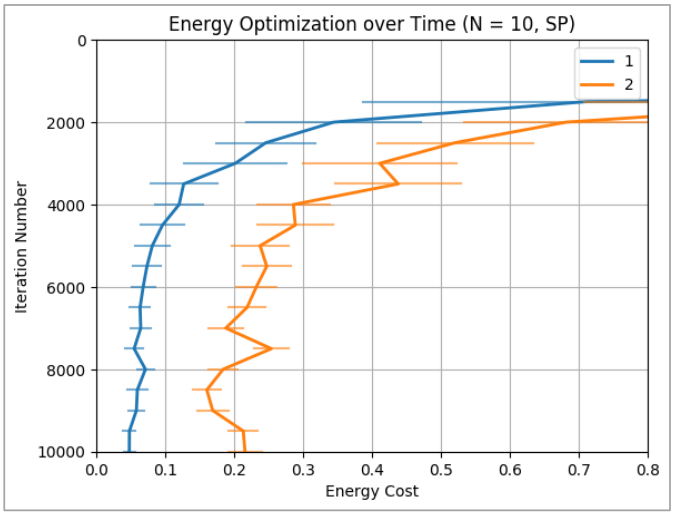
Communicative Intent 2

Visualization of Learned Policy for ASSOCIATED Sender Agent: $|G|$ = 2 Intents

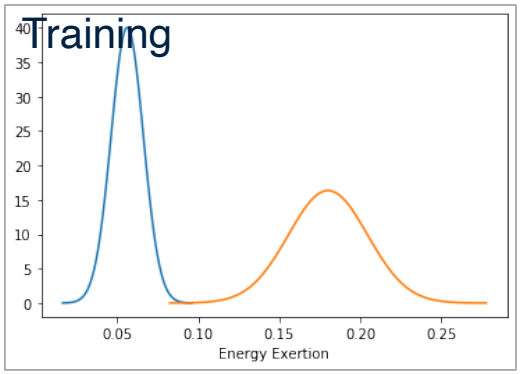
Qualitative Analysis
(*N=2* Intents Task)

Key Finding: Energy of Message *sufficient* for Prediction of Intent

ZSC possible *in principle* for continuous-action protocols!



At END of Protocol Training



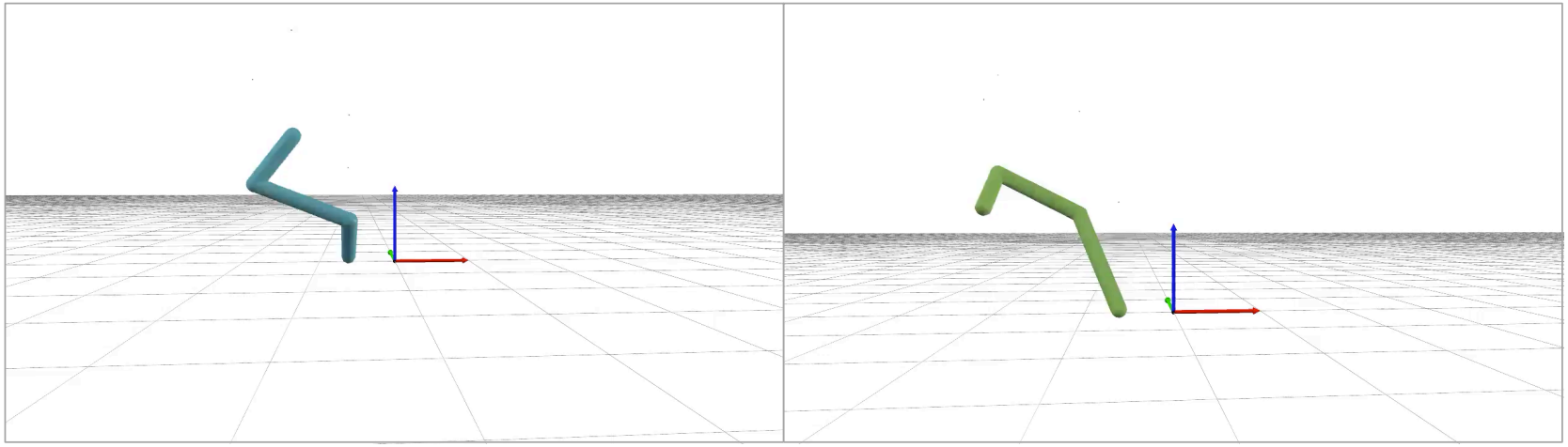
Achieves $I(G, C(\tau)) \gg 0$!

Open Challenges:

Difficulty *Inferring Latent Energy (Cost)* Variable from High-D Trajectories

Performance *Degrades* as $|G|$ Increases

Discovery of *Symmetries* for Continuous Channels



Communicative Intent 1

Communicative Intent 2

Visualization of Learned Policy for ASSOCIATED Sender Agent: $|G|$ = 2 Intents

Recap of Zero-Shot Communication Work

Motivating Goal: Emergence of Communication Protocols that **Generalize**

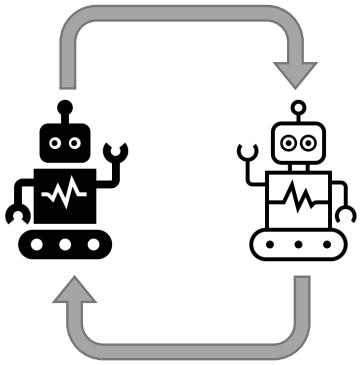
Novel Challenges

- Costly Channels (derived from *Energy Exertion*)
- High-Dimensional Continuous Channels (motivated by *Embodied* Agents)

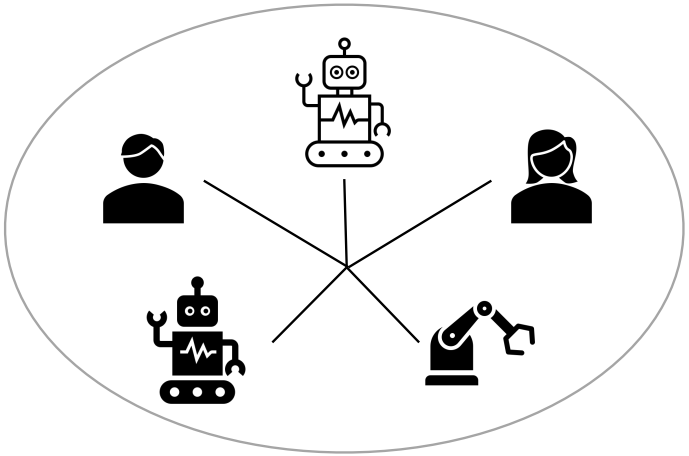
Key Contributions

- Formally introduced Zero-Shot Communication Problem Setting
- Incorporated Realistic Priors and Cost Constraint to Induce Learning Bias
- Theoretical Analysis of Communication Objective with Equivalence Classes
- QED Method for *Inferring* and *Generalizing* over Equivalence Classes
- Exploring Optimization of Continuous-Action Protocols

Research Vision:
Interactive Learning towards Multi Human-AI Cooperation



To Enable



Interactive Learning (methods)

Multi Human-AI Cooperation (long-term goal)

Open Problems for ZS Communication

Inferring *Optimal* Zero-Shot Communication Protocols

- Theoretical Analysis of QED Sample Complexity [*ongoing* work]
- Guiding policy search to *more robustly* converge on global optimum

Learning *Human-Compatible* Protocols

- Combining: MARL + supervisory signal / data
- Other realistic priors and inductive biases that can exploited

Mixed-Motive Settings (Social Dilemmas := Prosociality + Self-Interest)

- *Communication* Protocols [info exchange for *efficient* coordination]
  --> *Negotiation* Protocols [aligning incentives to *enable* cooperation]

https://www.kaleshabullard.com/
ksbullard@google.com