

Towards Understanding **Learning-to-Communicate** in MARL: An **Information-Structure** Perspective

Kaiqing Zhang

University of Maryland, College Park

IMSI, Chicago, IL

May 21, 2026

Acknowledgement

- Joint work with



Haoyi You (UMD)



Xiangyu Liu (UMD)

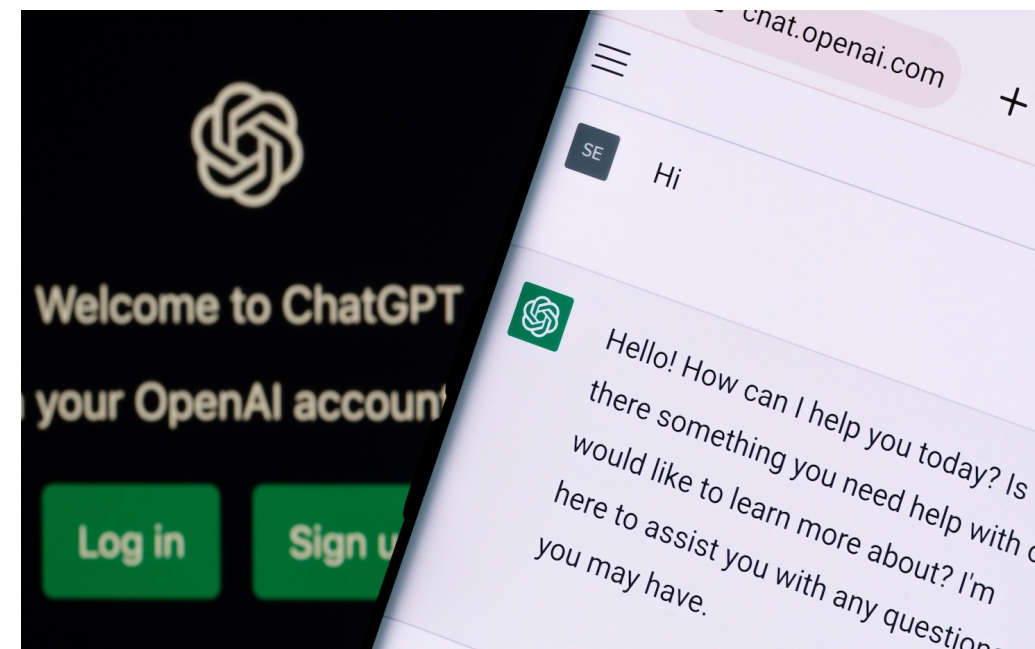
Background

Over the Years...

- Successes of reinforcement learning (RL) for decision-making and control



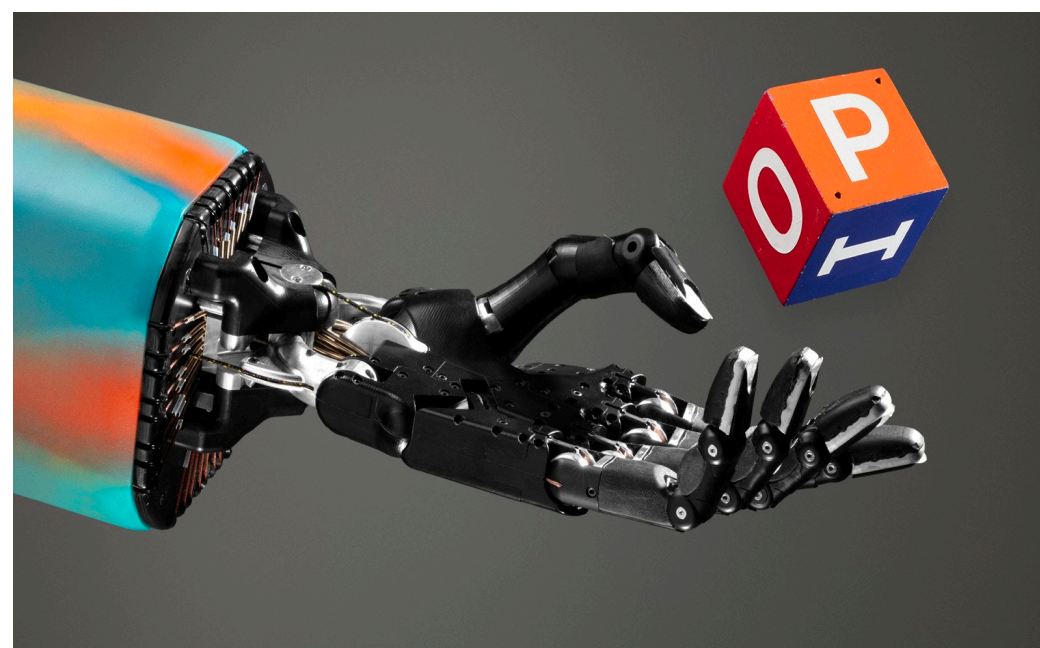
Board Games



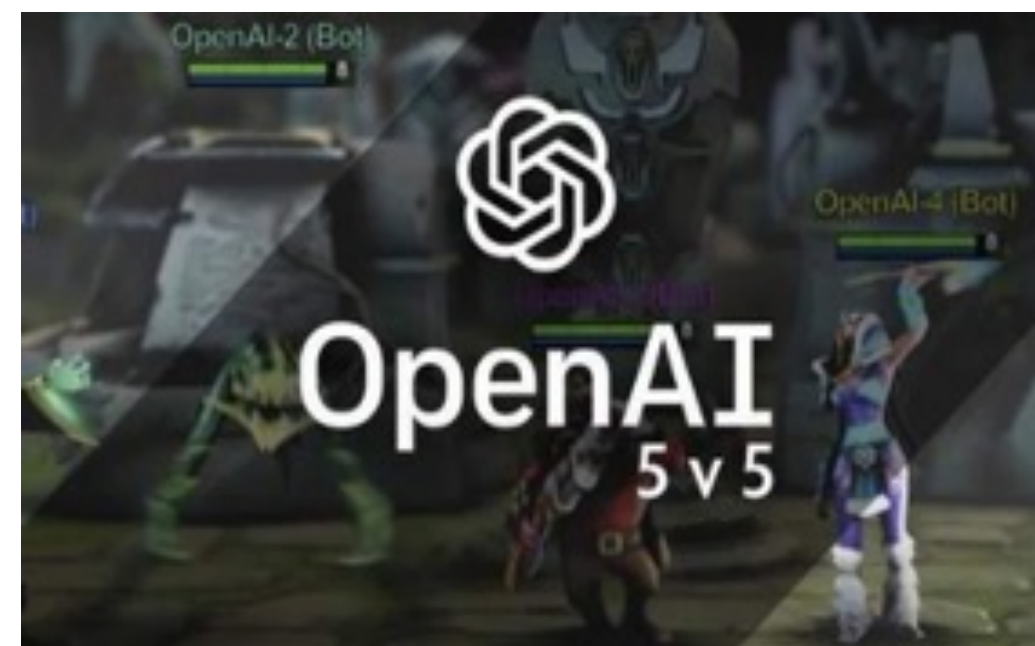
LLM Post-Training



Autonomous Driving



Robotics



Video Games



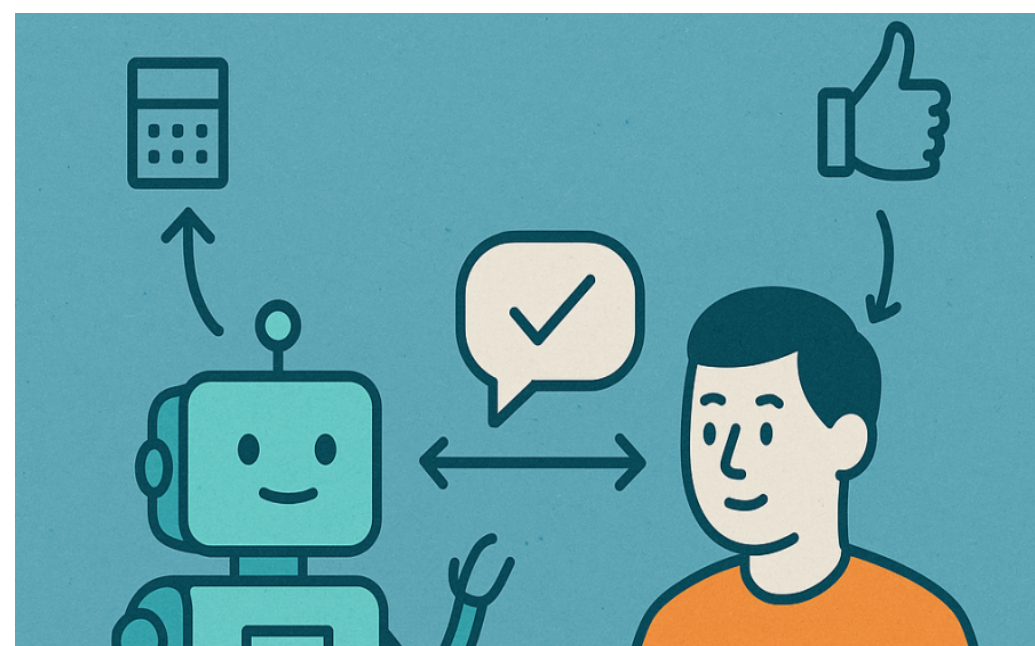
E-Commerce

Over the Years...

- Many of the success stories naturally involve **multiple agents** interacting with each other



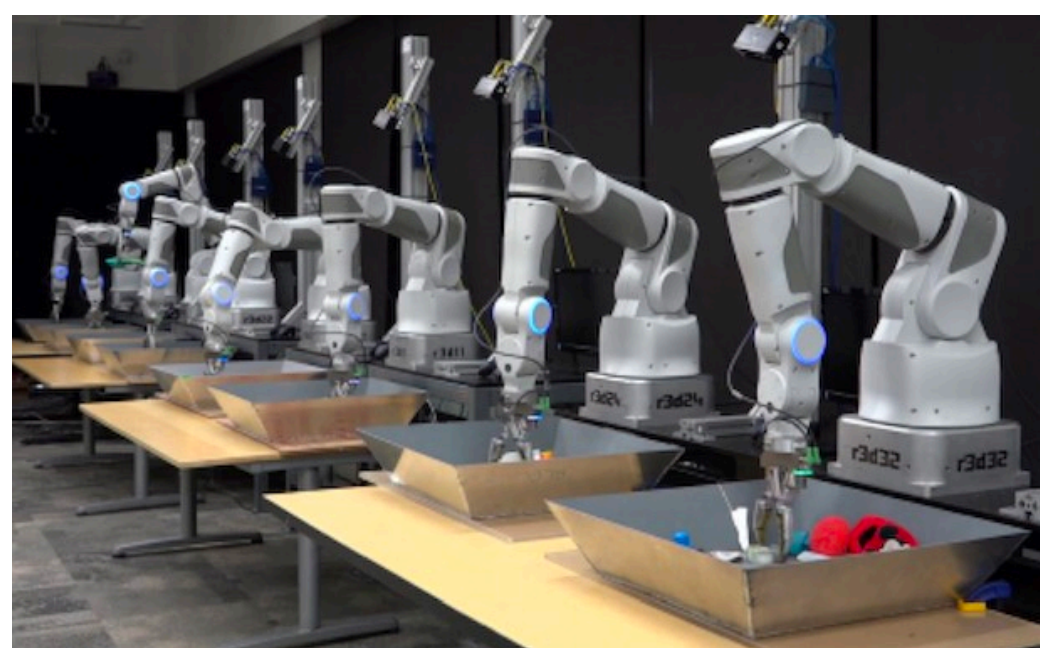
Board Games



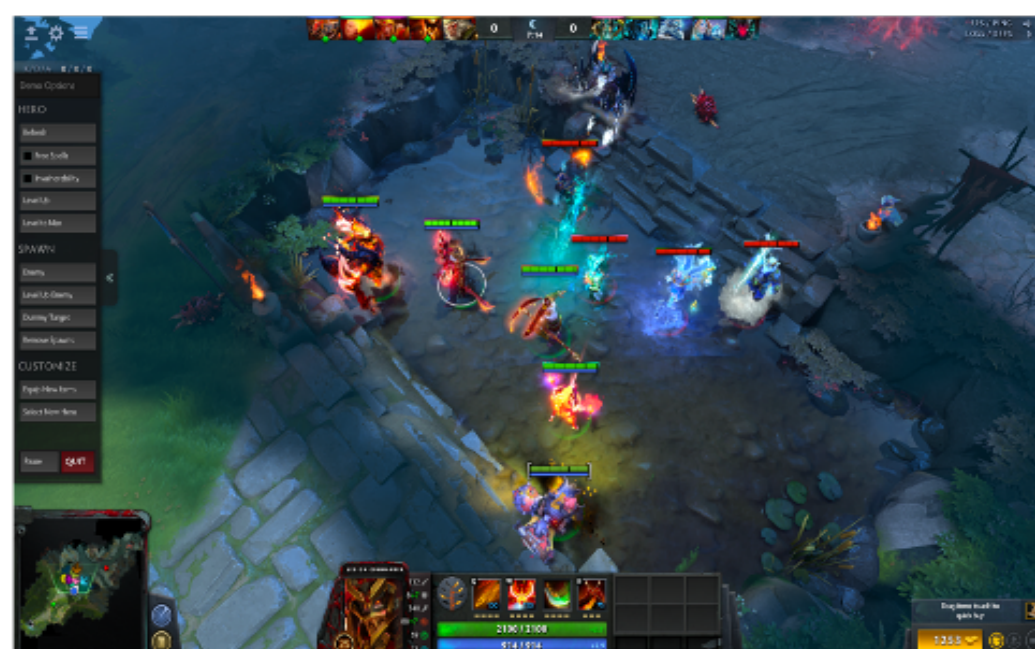
LLM Post-Training



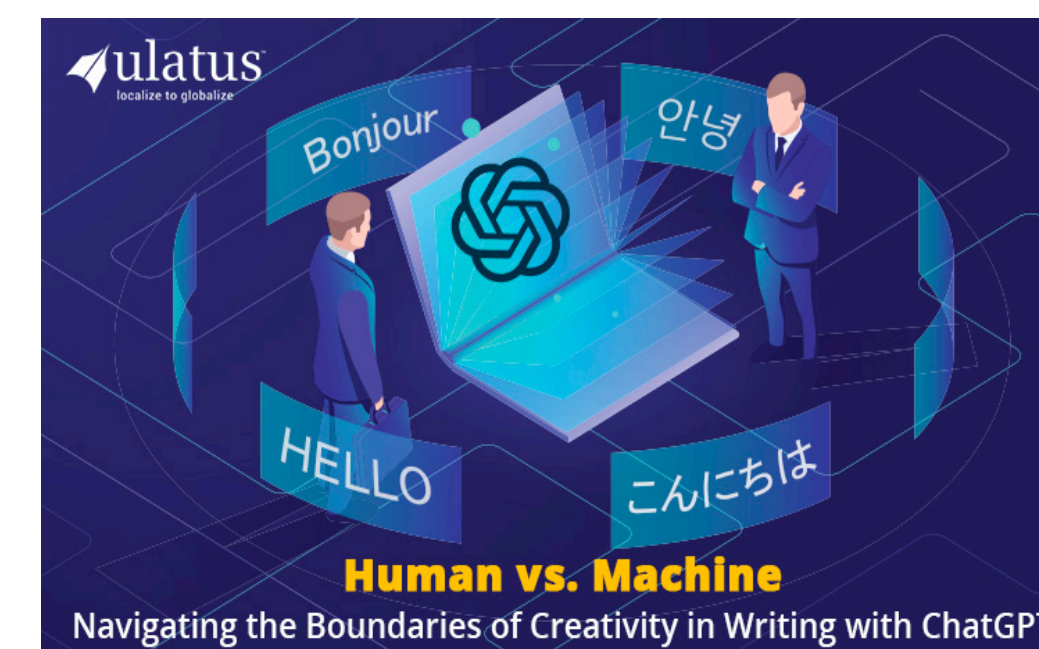
Autonomous Driving



Robotics



Video Games



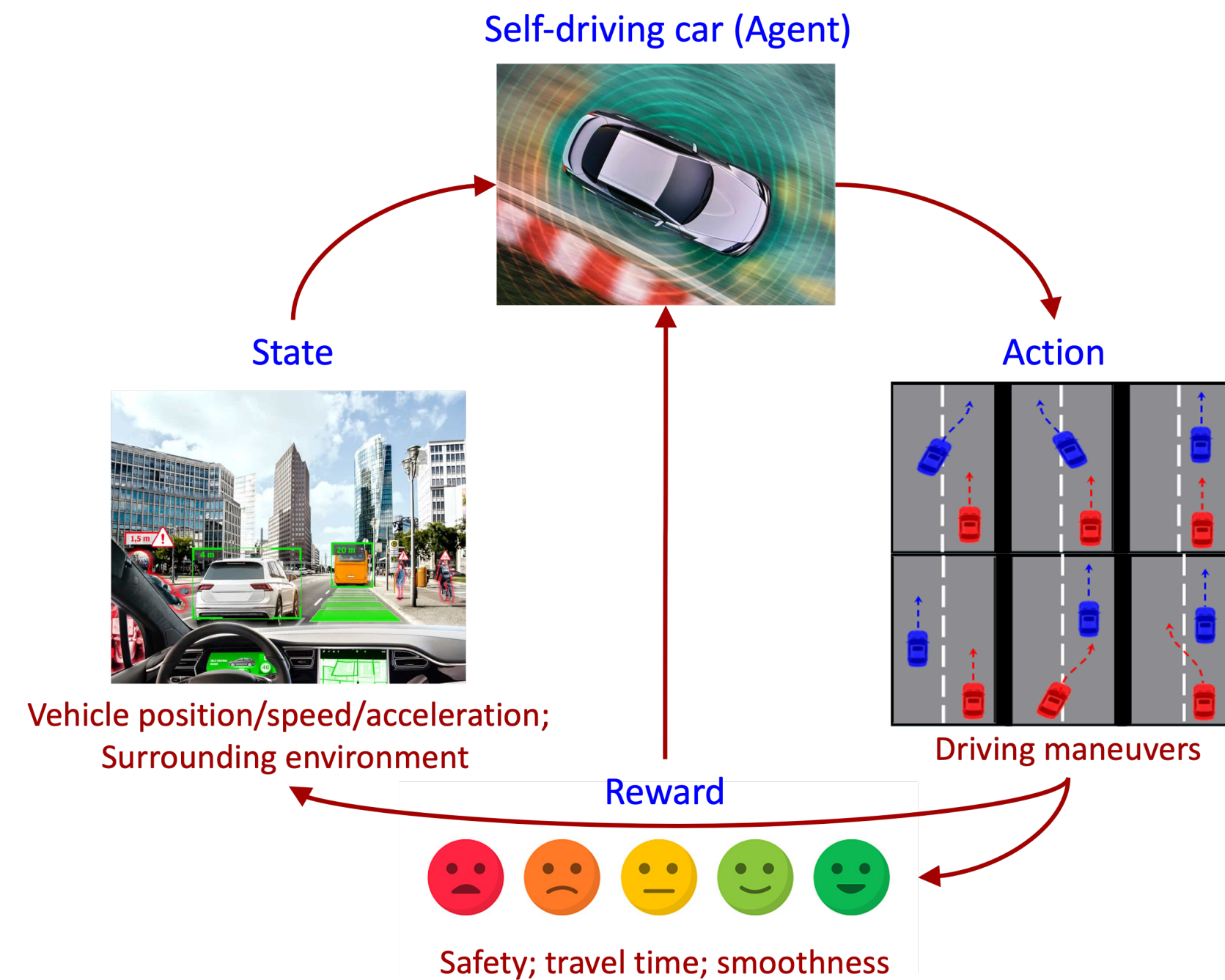
E-Commerce

- Significant advances on **foundations of RL** in **dynamic** (and **multi-agent**) environments

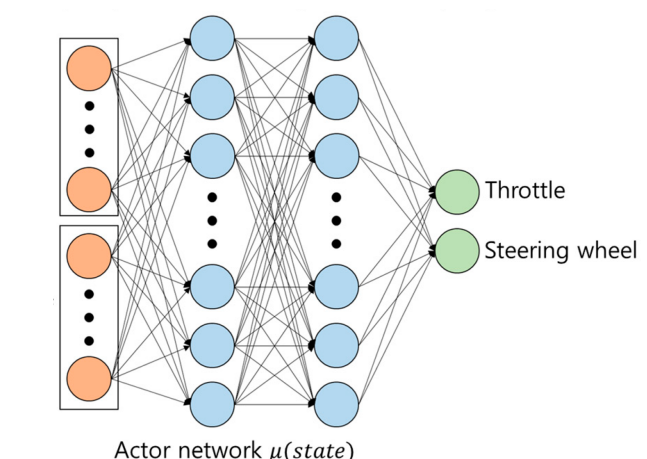
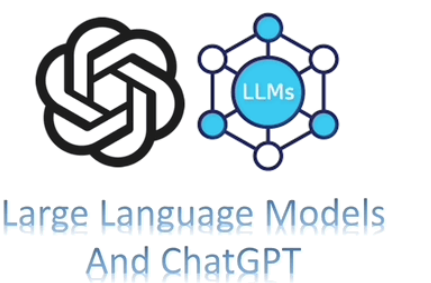
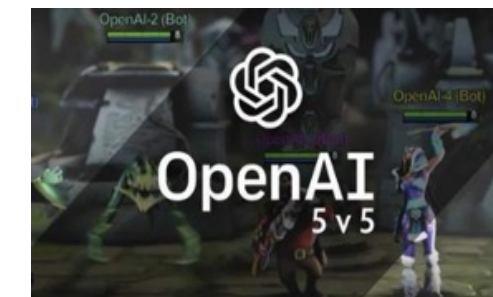
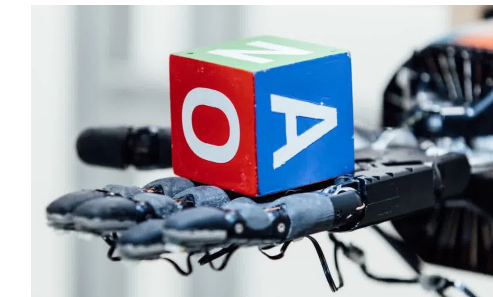
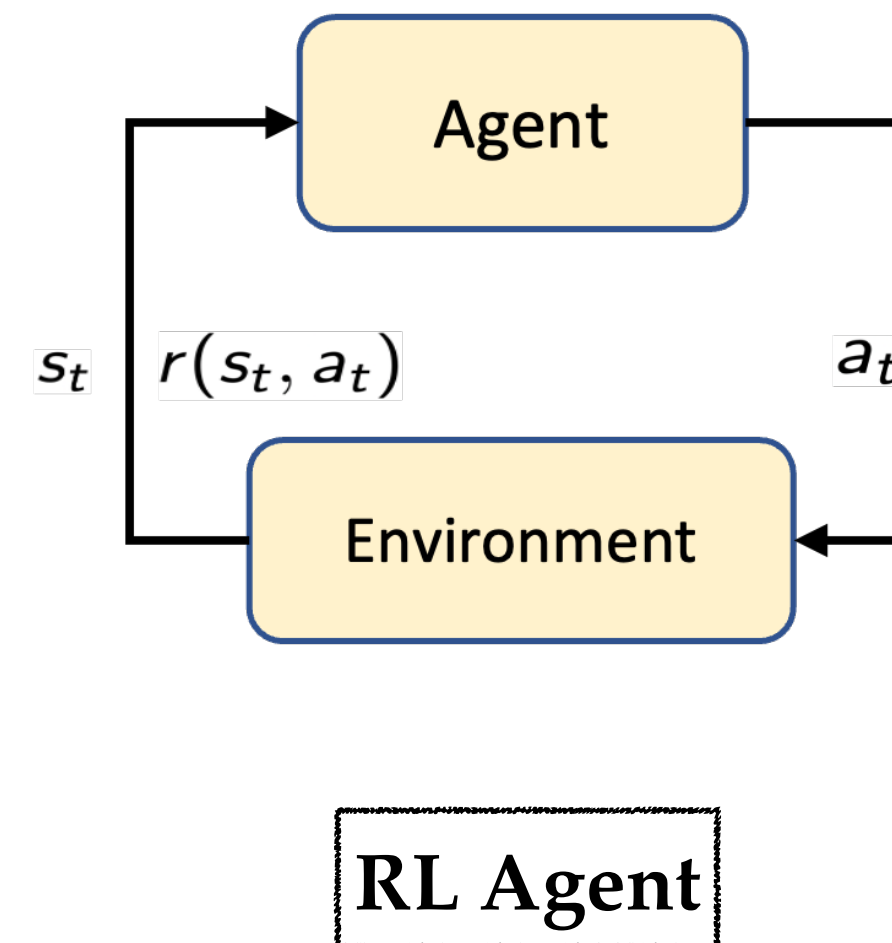
(Multi-Agent) Reinforcement Learning

- Reinforcement learning (RL): Data-driven optimal control

Goal: Arrive at the destination as soon as possible, safely



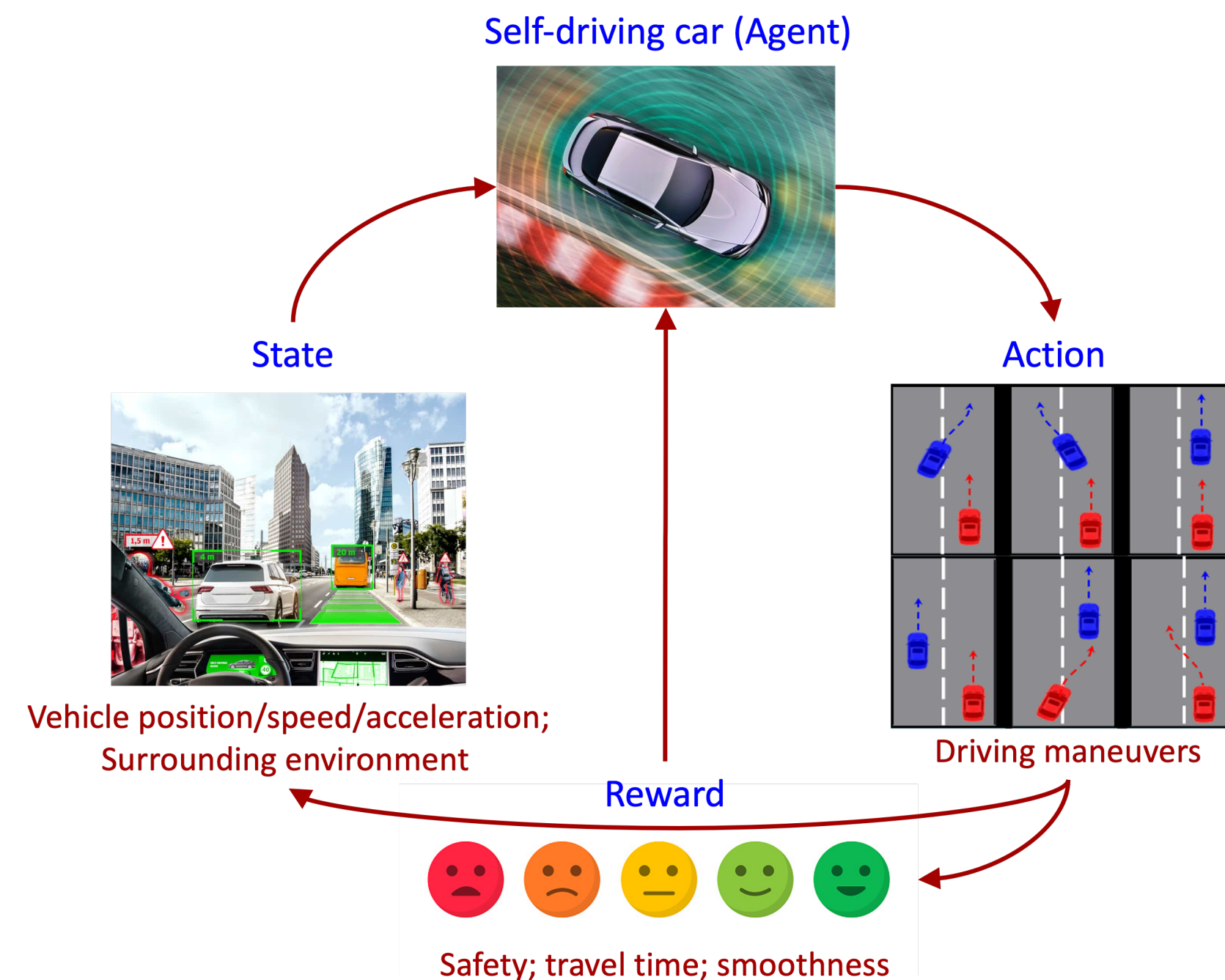
$$\text{Maximize } \mathbb{E} \left[\sum_t r(s_t, a_t) \right]$$



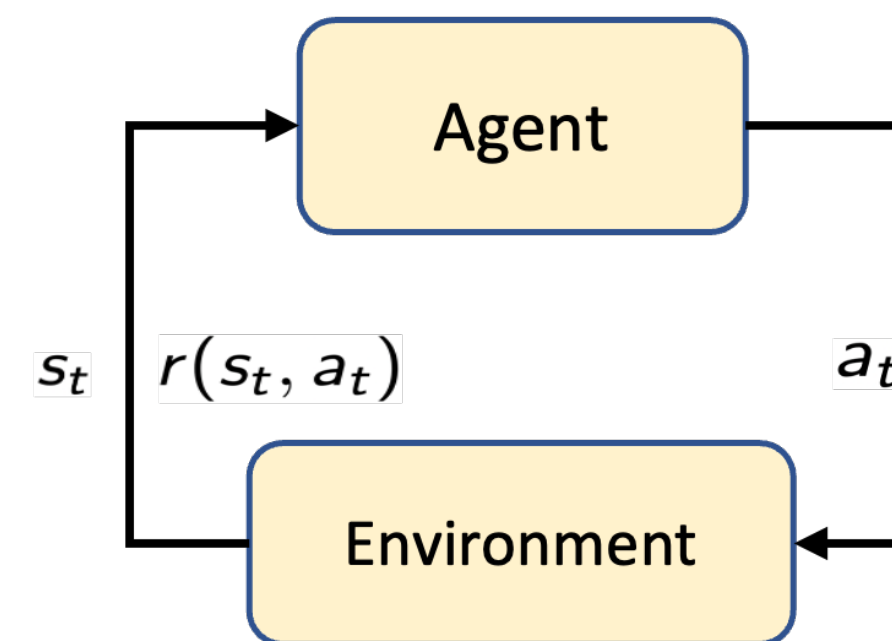
(Multi-Agent) Reinforcement Learning

- Reinforcement learning (RL): Data-driven optimal control
- Predominantly studied under the optimal control model of Markov decision processes (MDPs) [Puterman, '90][Bertsekas, '95] and Markov/stochastic games [Shapley '53][Fink, '64]

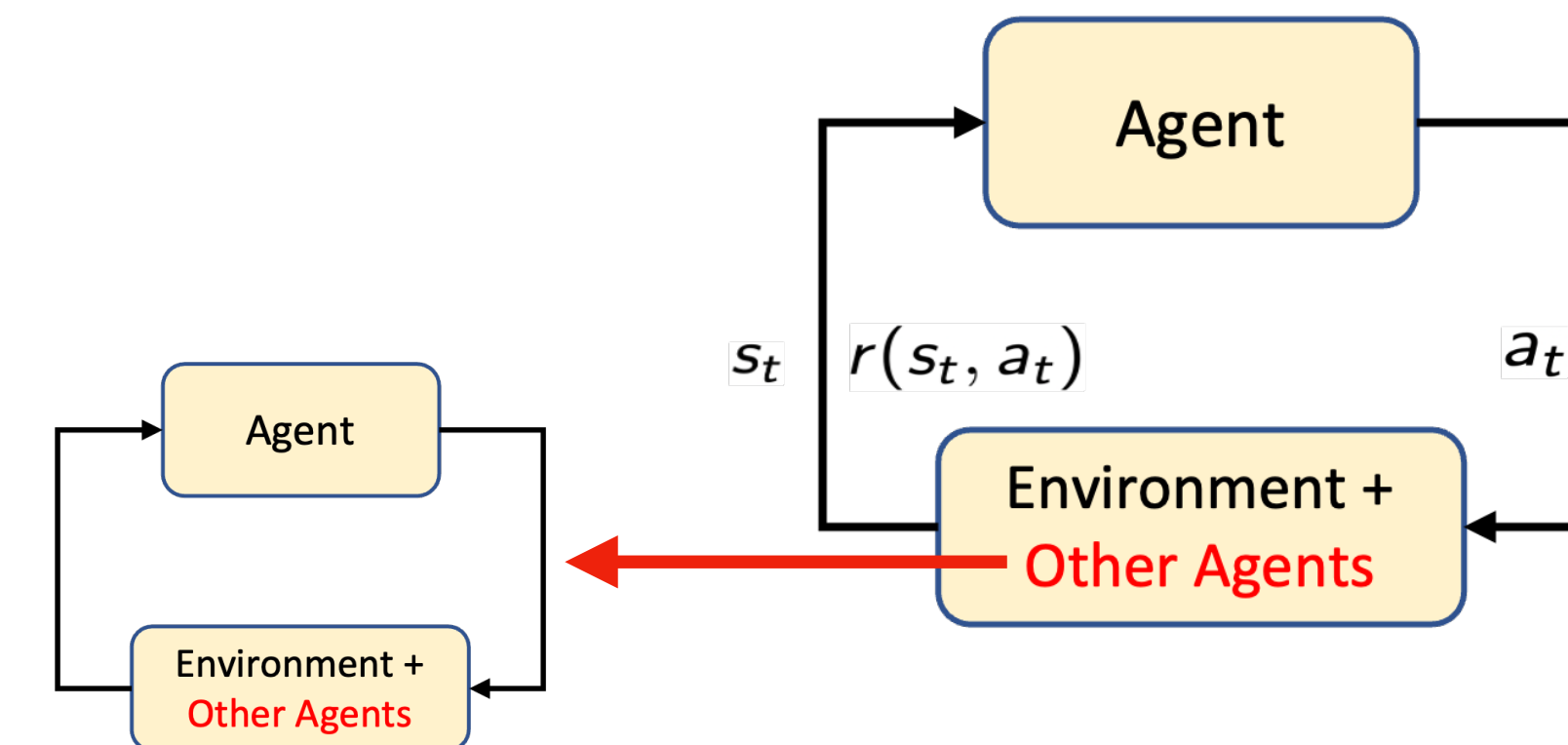
Goal: Arrive at the destination as soon as possible, safely



$$\text{Maximize } \mathbb{E} \left[\sum_t r(s_t, a_t) \right]$$



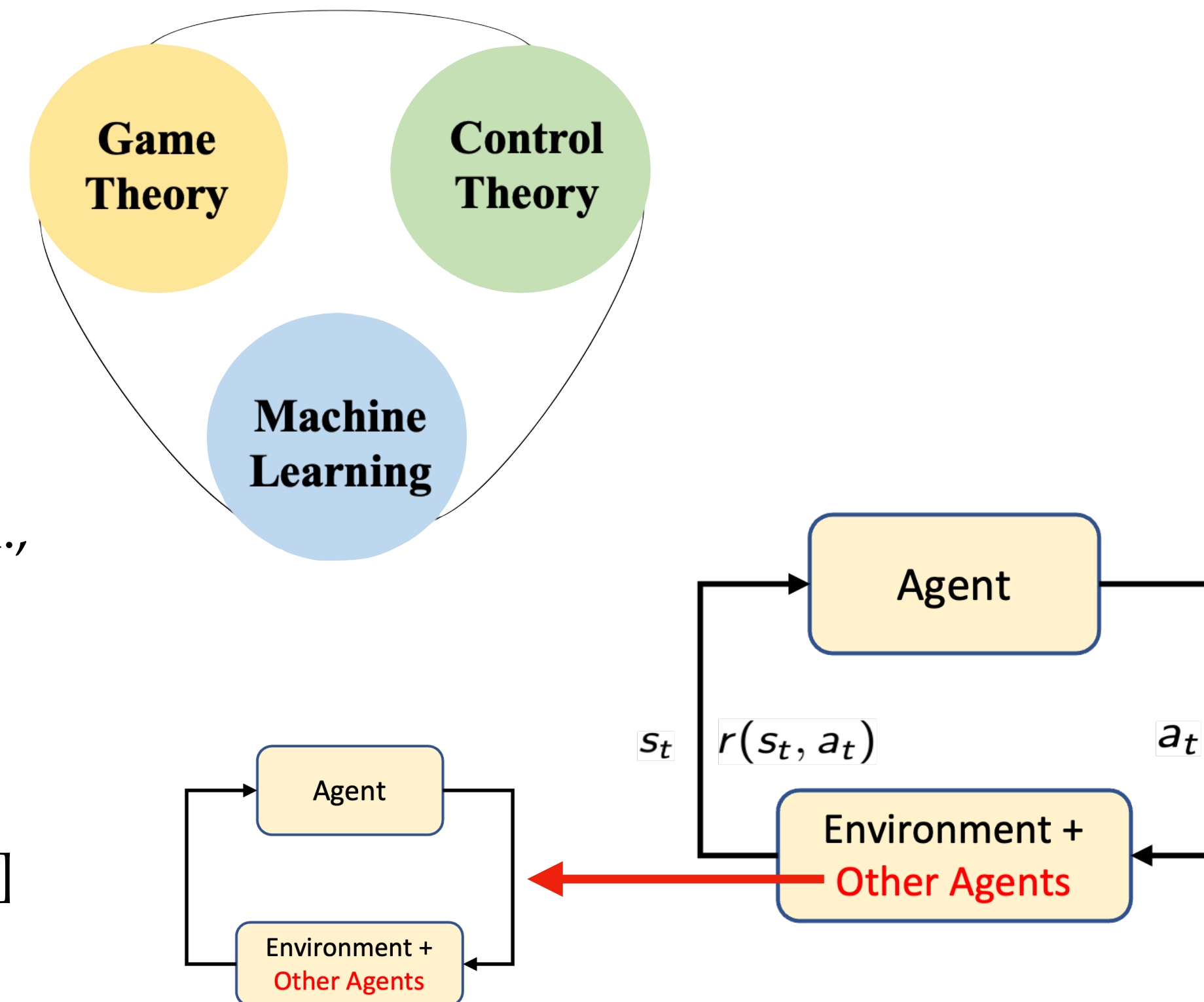
RL Agent



Interactive RL Agents

(Multi-Agent) Reinforcement Learning

- Reinforcement learning (RL): Data-driven optimal control
- Predominantly studied under the optimal control model of Markov decision processes (MDPs) [Puterman, '90][Bertsekas, '95] and Markov / stochastic games [Shapley '53][Fink, '64]
- Long (Incomplete!) list of (recent) results on RL theory (broad interest from Machine Learning, Game Theory, Control communities):
 - Single-agent RL: [Azar et al., '17][Jin et al., '18][Sidford et al., '18][Fazel et al., '18][ZKZB, '19][Dean et al., '19][Sarkar et al., '19][Yang and Wang, '19][Chen and Jiang, '19][Jin et al., '20][Cai et al., '20][LZBY, '20][Agarwal et al., '21][Jin et al., '21][Li et al., '24][Zhang et al., '25]...
 - Multi-agent RL: [Littman, '94][Hu and Wellman, '03][Arslan and Yüksel, '17][ZHB, '19, '20][Bai and Jin, '20][ZKBY, '20][Xie et al., '20][Daskalakis et al., '21][Jin et al., '21][OSZ, '22][Zhang et al., '22][Jin et al., '22][Li et al., '22][ZSO, '23][DGZ, '23][Zhang et al., '23, '24]...
- State is assumed to be observable, and shared among agents



Interactive RL Agents

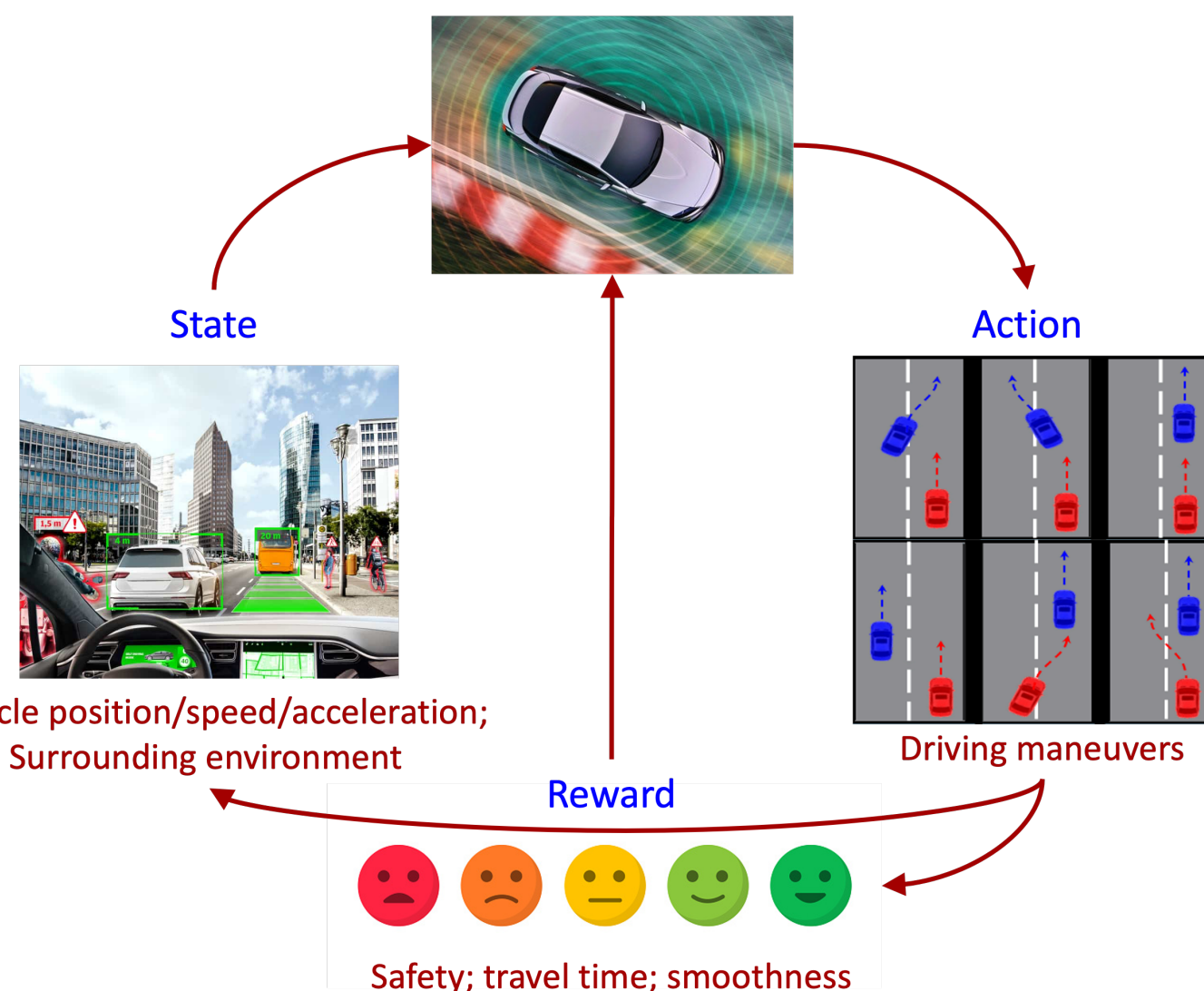
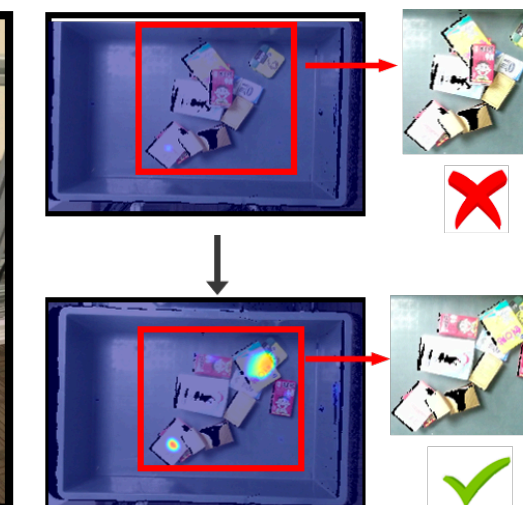
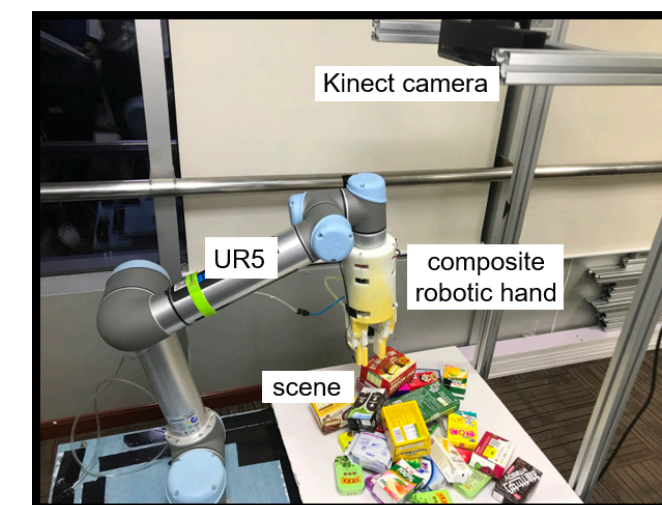
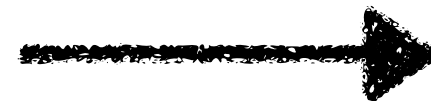
Agents (Only) Have **Partial, Asymmetric** Information

- In practice, however, agents only access **partial / noisy / delayed** observations almost always
 - Especially in **multi-agent** decision-making: Agents have not only **partial state observations**, but also partially information about **other agents** — **asymmetric** information & **decentralized** decision-making

- Many more...



Self-driving car (Agent)



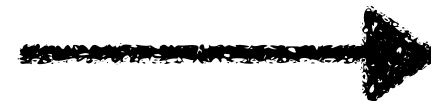
RL agents "in the wild"

Agents (Only) Have **Partial, Asymmetric** Information

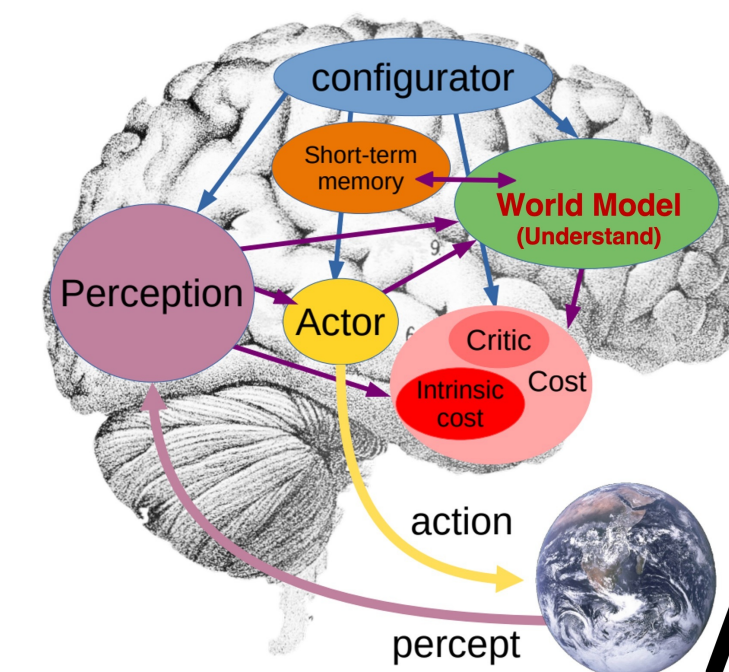
- In practice, however, agents only access **partial / noisy / delayed** observations almost always
 - Especially in **multi-agent** decision-making: Agents have not only **partial state observations**, but also partially information about **other agents** — **asymmetric** information & **decentralized** decision-making



Self-driving car (Agent)

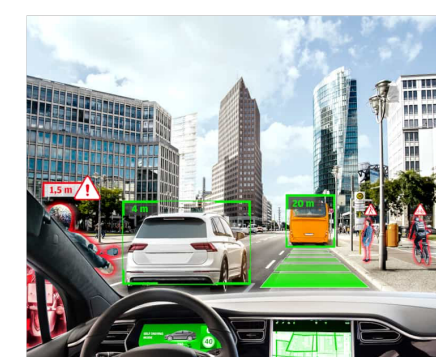
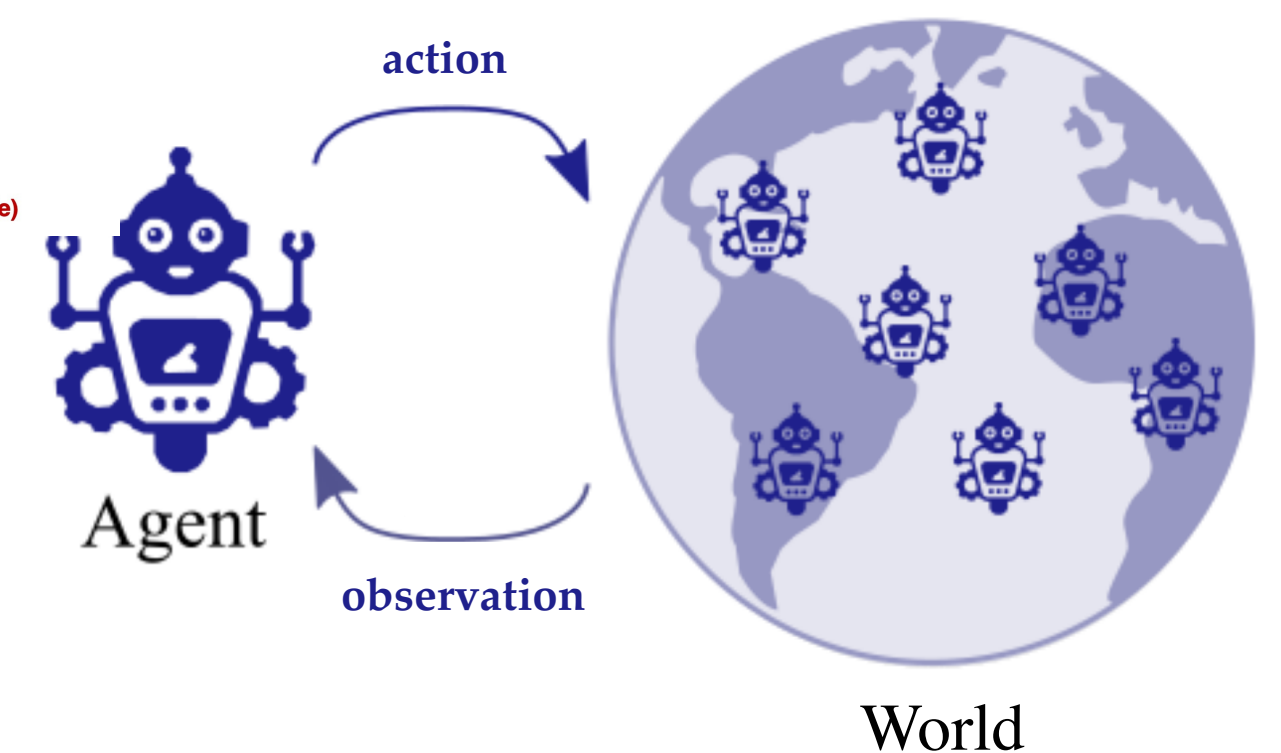


Each agent has their **own, partial view** of the world (state), as they “perceive, plan, and act” autonomously



The “world model” [Ha and Schmidhuber, '18][LeCun, '22]...

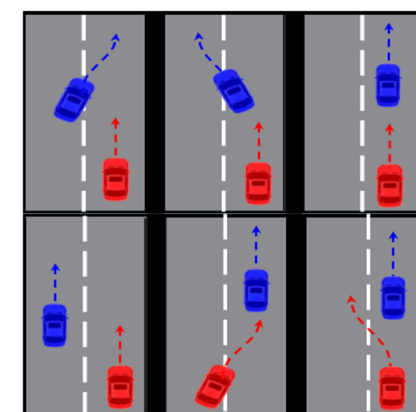
World Model (Simulate)



State

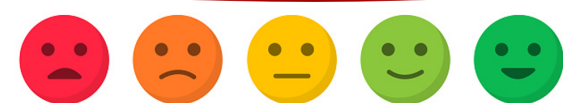


Action



Driving maneuvers

Reward



Safety; travel time; smoothness

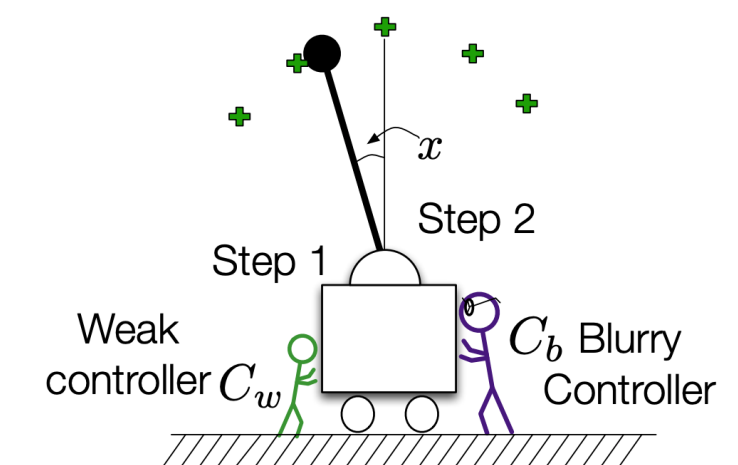
RL agent



RL agents “in the wild”

“In Theory”

- Single-agent case: **Computing** an optimal policy is **PSPACE-complete** [Papadimitriou and Tsitsiklis, '87] even with model knowledge; **Learning** requires **exponential** # of **samples** [Krishnamurthy et al., '16; Jin et al., '20]
 - Neither **computationally** nor **statistically** efficient algorithms exist in general
- Multi-agent case: Computing a **team-optimal policy** (even with **2 agents**) is NEXP-hard [Bernstein et al., '02]
 - Can take **doubly exponential time** in general
 - Even **single-step decentralized** decision-making is **NP-hard** [Tsitsiklis and Athans, '85]
 - Intriguing counterexample of Witsenhausen's in **decentralized stochastic control** [Witsenhausen '68] — **linear** systems but **nonlinear** optimal policy

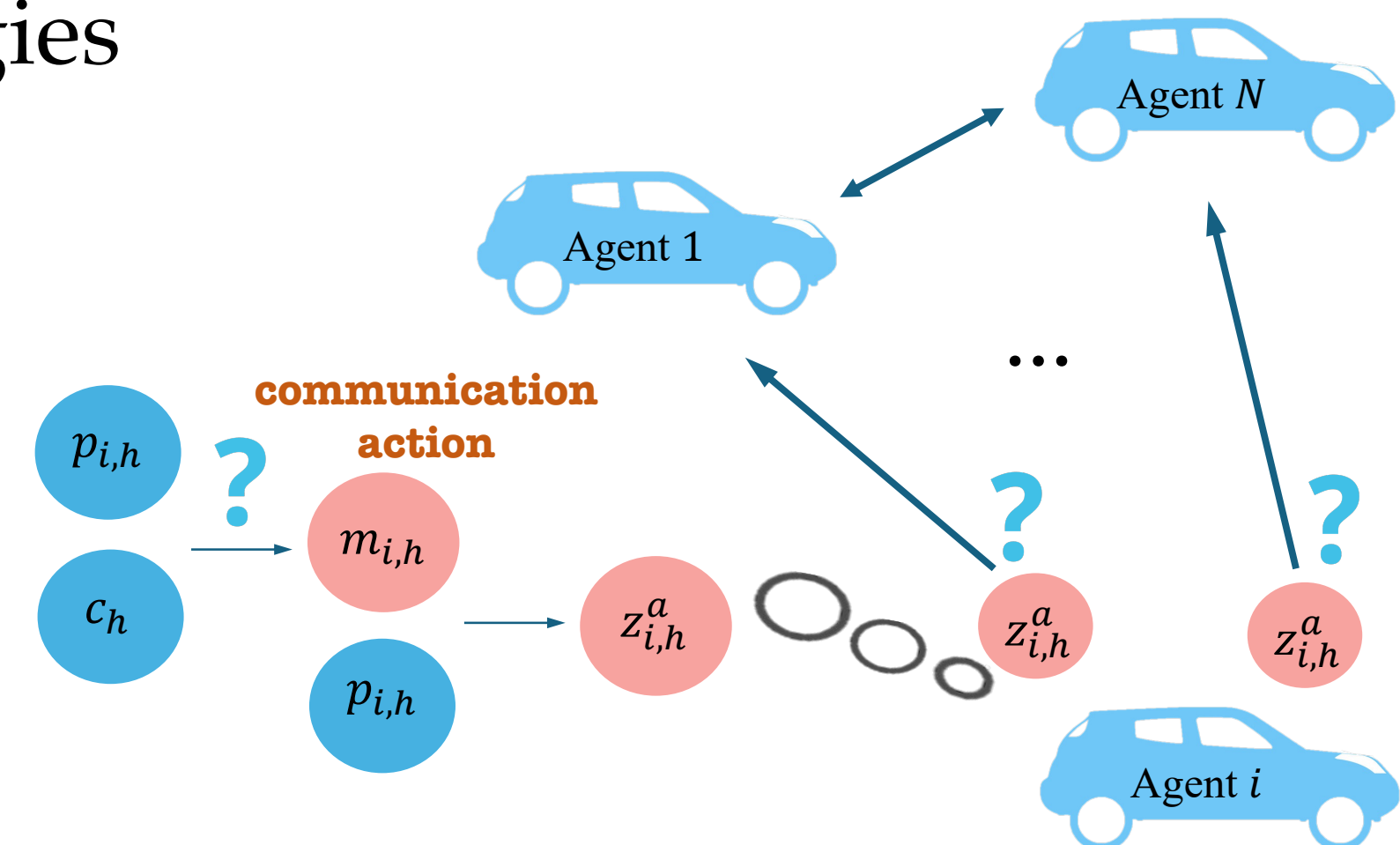
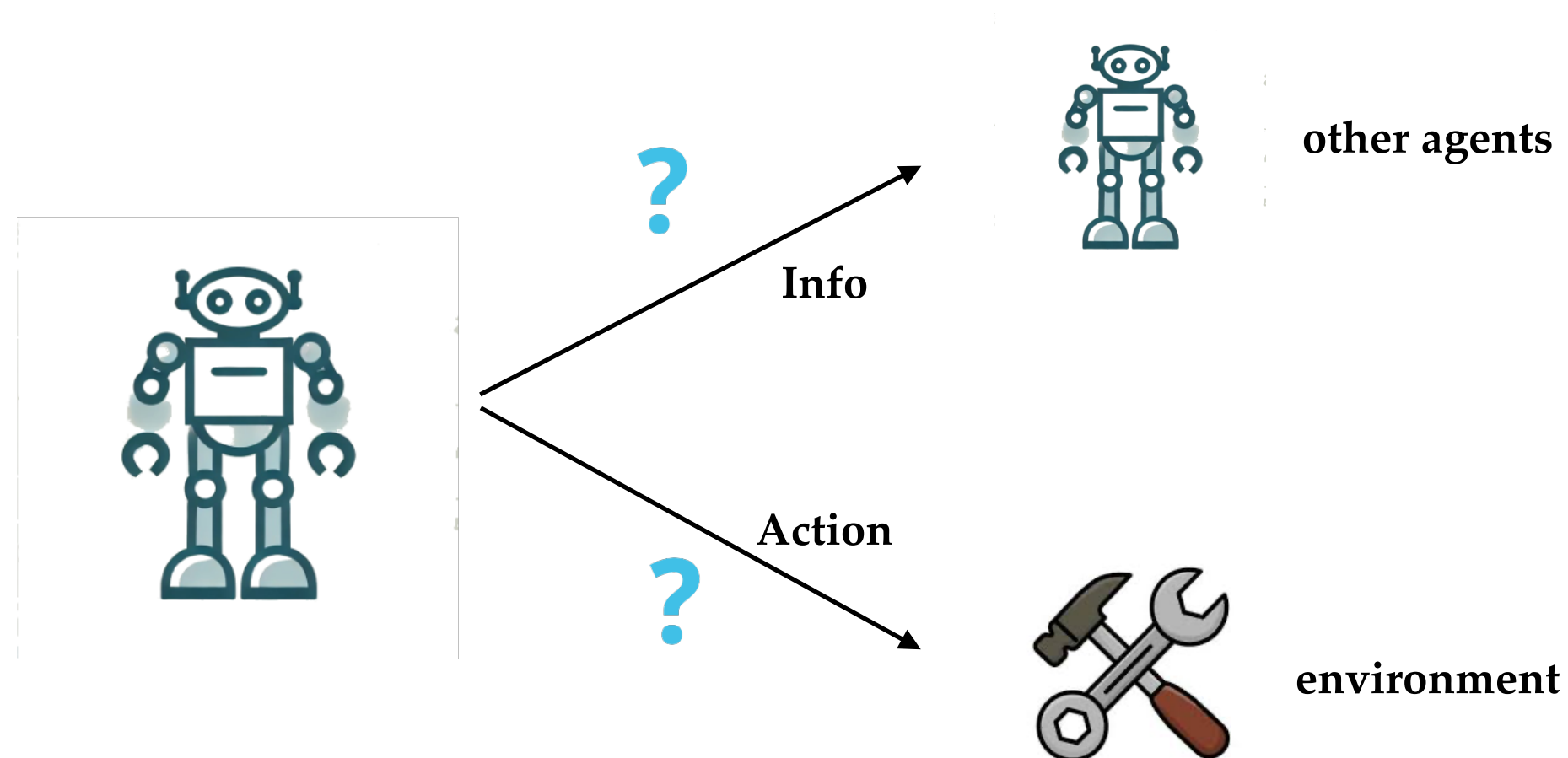


Witsenhausen's Counterexample

Control and learning with partial observations are fundamentally HARD!

An (Empirical) Remedy: **Learning-to-Communicate**

- A popular approach to address **partial observability** in deep / empirical multi-agent RL (MARL):
 - **Learning-to-Communicate** (LTC) [Foerster et al., '16][Sukhbaatar et al., '16][Jiang and Lu, '18]...
- In classical MARL, agents learn **control** strategies to maximize their expected accumulated rewards
- In LTC, agents need to learn both **control** and **communication** strategies to maximize the rewards
 - Agents **share information** via the communication strategies



“In (Control) Theory”

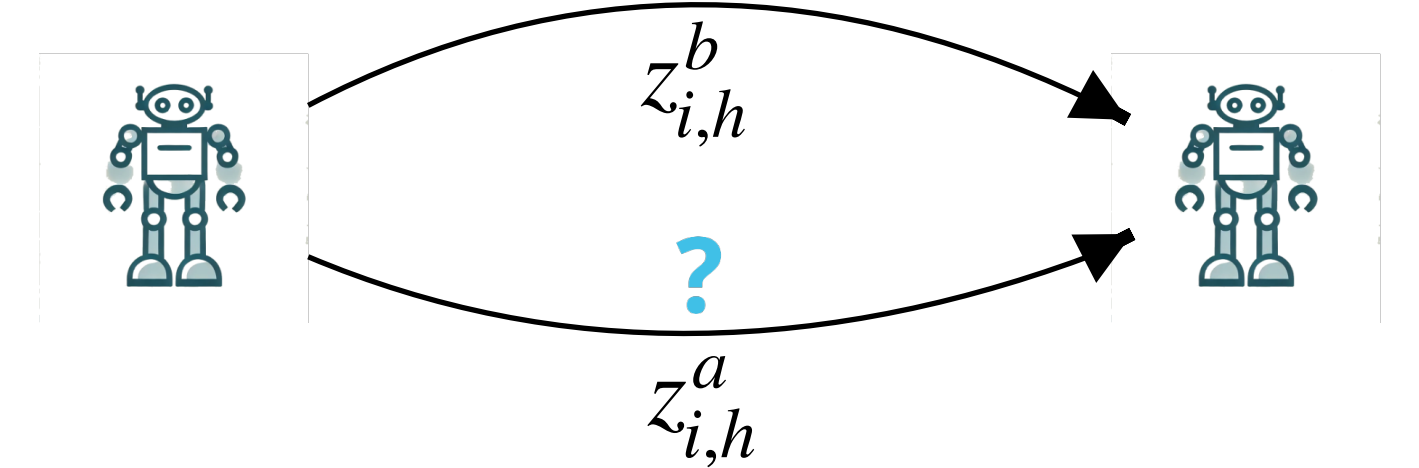
- The role of **information/communication** in **control/decision-making** has long been studied in **(decentralized) stochastic control** [Sandell et al., '78] and **dynamic games** [Başar and Olsder, '98]
 - **Information structures** [Witsenhausen, '71][Mahajan, et al., '12][Yüksel and Başar, '25]: Formally captures “Who knows what and when?”
 - Control under **communication constraints** & **networked control systems** [Tatikonda and Mitter, '04][Borkar and Mitter, '97][Jain et al., '02][Nair and Evans, '04][Hespanha et al., '07]...
 - Computational **intractability** for decision-making with **decentralized information** [Tsitsiklis and Athans, '85][Papadimitriou and Tsitsiklis, '86][Blondel and Tsitsiklis, '00]...
 - (Optimization) **tractability** conditions [Ho and Chu, '72][Lessard and Lall, '11][Rotkowitz and Lall, '06]...
 - **Sufficient statistics** and **structural results** (for **sequential decomposition**) [Mahajan, '08][Nayyar et al., '13a, '13b][Yüksel, '13][Lamperski and Lessard, '15]...

Can we better understand **learning-to-communicate**
through these **principles**?

Preliminaries

Formalizing Learning-to-Communicate

- Formalize LTC $\mathcal{L}$ under the framework of decentralized partially observable MDPs (Dec-POMDPs)
- Decision-making/Control: $\langle H, \mathcal{S}, \{\mathcal{A}_{i,h}\}_{i \in [n], h \in [H]}, \{\mathcal{O}_{i,h}\}_{i \in [n], h \in [H]}, \{\mathcal{R}_h\}_{h \in [H]}, \{\mathbb{T}_h\}_{h \in [H]}, \{\mathbb{O}_h\}_{h \in [H]} \rangle$
 - $n > 1$ agents and $H > 1$ timesteps
 - At timestep $h \in [H]$, agent $i \in [n]$ observes $o_{i,h} \sim \mathbb{O}_{i,h}(\cdot \mid s_h)$, selects $a_{i,h}$, and state transitions to $s_{h+1} \sim \mathbb{T}_h(\cdot \mid s_h, a_h)$
- Communication:
 - Baseline sharing z_h^b , based on existing information-sharing protocol (for generality and also for computational tractability, more on this later): h^- -step
 - Additional sharing $z_h^a = \bigcup_{i=1}^n z_{i,h}^a$, where agent i shares $z_{i,h}^a$: h^+ -step
- Common information: All the shared information so far: $c_{h-} = \bigcup_{t=1}^{h-1} z_t \cup z_h^b$, $c_{h+} = \bigcup_{t=1}^h z_t$ with $z_h := z_h^b \cup z_h^a$



Formalizing Learning-to-Communicate

- Additional sharing $z_{i,h}^a$ determined by the **communication action** $m_{i,h} \in \mathcal{M}_{i,h}$: $z_{i,h}^a = \phi_{i,h}(p_{i,h-}, m_{i,h})$

Private Information

Common Information

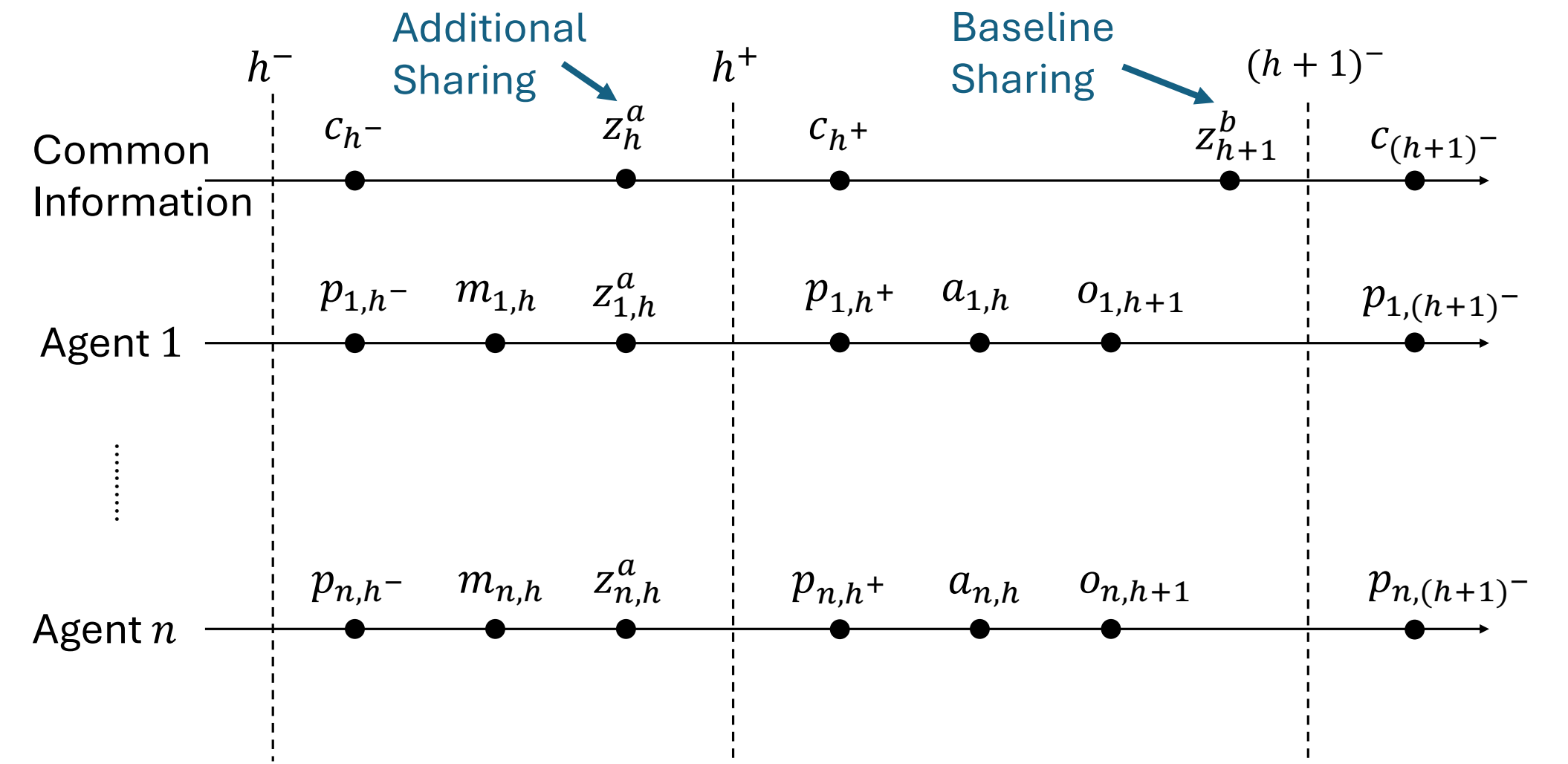
- Information available** to agent i at step h : $\tau_{i,h-} = p_{i,h-} \cup c_{h-}$, $\tau_{i,h+} = p_{i,h+} \cup c_{h+}$
 - Private information** evolves as $p_{i,h-} = \xi_{i,h}(p_{i,(h-1)+}, a_{i,h-1}, o_{i,h})$ and $p_{i,h+} = p_{i,h-} \setminus z_{i,h}^a$
 - Closed-loop** information structure: $o_{i,h} \in \tau_{i,h-}$; Information **recall**: $\tau_{i,h-} \subseteq \tau_{i,h+} \subseteq \tau_{i,(h+1)-}$

- An illustration: d -step delayed baseline sharing

$$p_{i,h-} = \{ \textcolor{red}{o}_{i,h-d}, \textcolor{blue}{a}_{i,h-d}, \textcolor{blue}{o}_{i,h-d+1}, \textcolor{red}{a}_{i,h-d+1}, \dots, \textcolor{blue}{o}_{i,h} \}$$

$$m_{i,h} = (\textcolor{red}{0}, \textcolor{blue}{1}, \textcolor{blue}{1}, \textcolor{red}{0}, \dots, \textcolor{blue}{1})$$

$$z_{i,h}^a = \{ \textcolor{blue}{a}_{i,h-d}, \textcolor{blue}{o}_{i,h-d+1}, \dots, \textcolor{blue}{o}_{i,h} \}$$



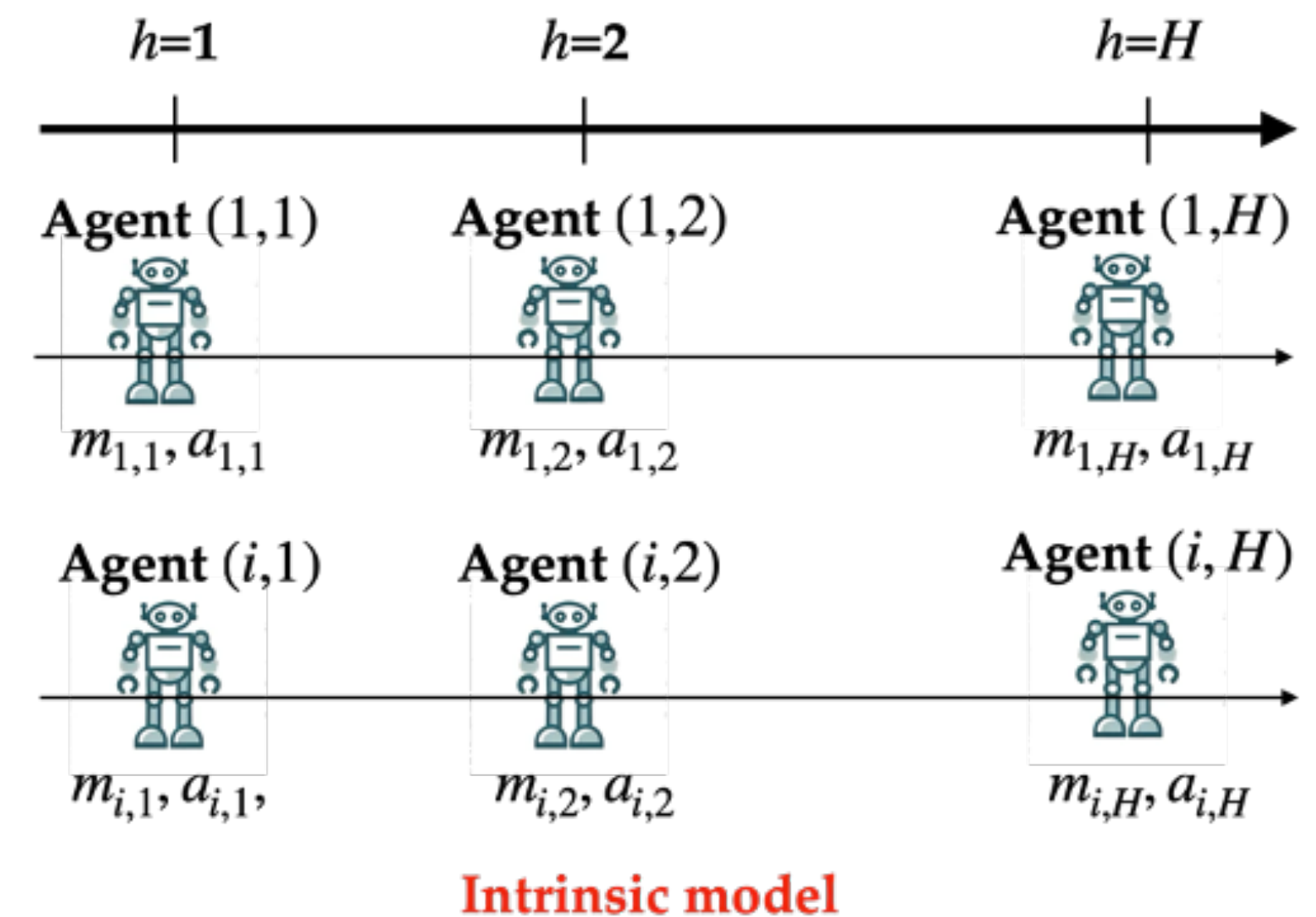
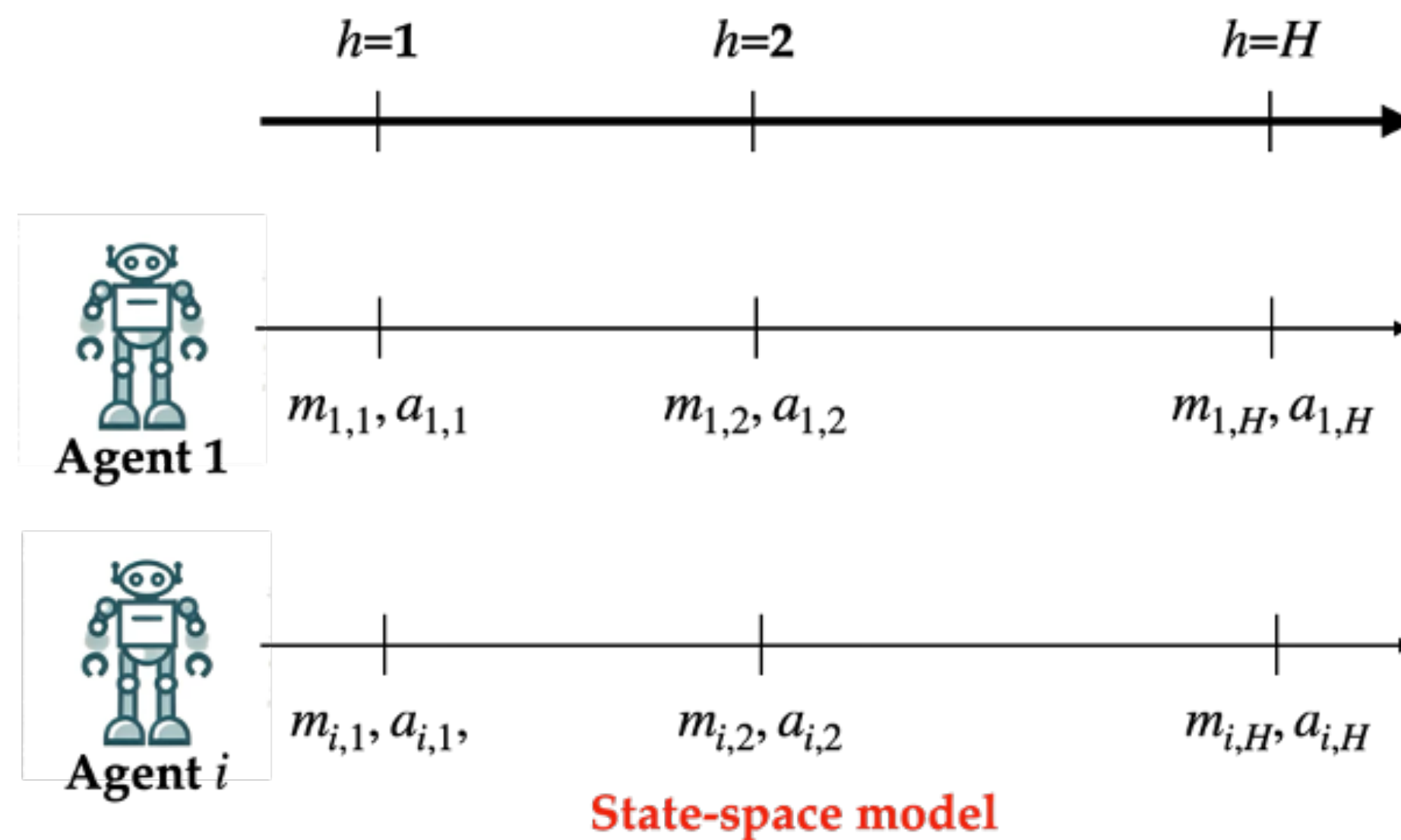
Formalizing Learning-to-Communicate

- **Communication strategy** $g_{i,h}^m$ chooses $m_{i,h}$ based on $\tau_{i,h-}$ (**before** additional sharing)
- **Control strategy** $g_{i,h}^a$ chooses $a_{i,h}$ based on $\tau_{i,h+}$ (**after** additional sharing)
- Control **reward** at time h : $r_h = \mathcal{R}(s_h, a_h)$
- Communication **cost** at time h : $\kappa_h = \mathcal{K}(z_h^a)$
- Overall goal: optimize **jointly** over $(g_{1:H}^a, g_{1:H}^m)$ to maximize

$$J_{\mathcal{L}}(g_{1:H}^a, g_{1:H}^m) := \mathbb{E}_{\mathcal{L}} \left[\sum_{h=1}^H (r_h - \kappa_h) \middle| g_{1:H}^a, g_{1:H}^m \right]$$

Information Structure & Intrinsic Model

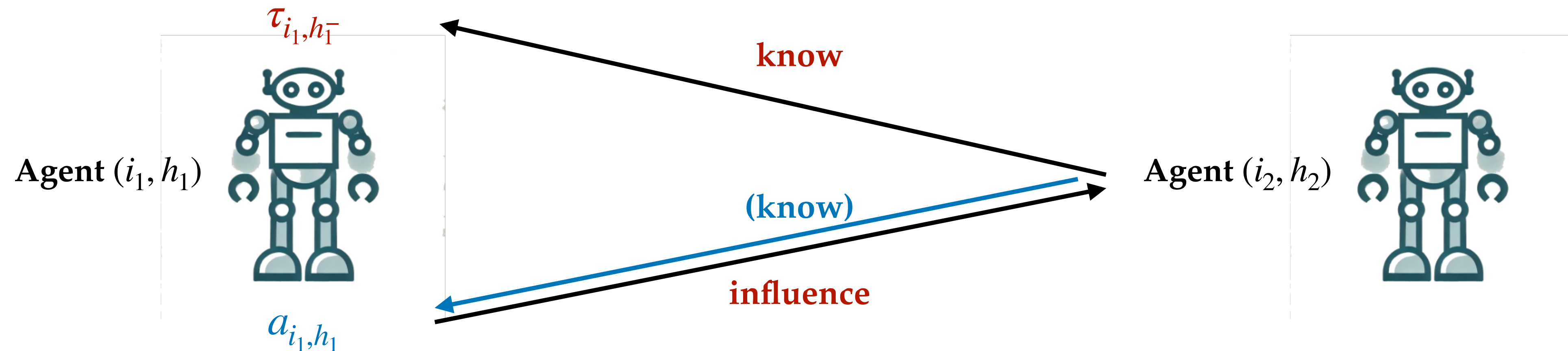
- Information structure (IS) is an important notion in (decentralized) stochastic control theory, which characterized *who knows what and when*
- Usually conveniently discussed under the equivalent **intrinsic models** [Witsenhausen, '75]
 - “Each agent only acts once” — in total $H \times n$ “agents”



- Without **additional sharing**, LTC $\mathcal{L} \rightarrow$ Dec-POMDP (with baseline sharing) **induced** by LTC $\overline{\mathcal{D}}_{\mathcal{L}}$
- We define the **IS of an LTC problem $\mathcal{L}$** as the **IS of $\overline{\mathcal{D}}_{\mathcal{L}}$** (an inherent property of the problem $\mathcal{L}$)

Information Structure & Intrinsic Model

- (Strictly) Quasi-classical ((s)QC) IS is an important subclasses of IS [Witsenhausen, '75]
- An LTC is (s)QC (i.e., the DecPOMDP induced by LTC is (s)QC), if when there is **no additional sharing**, agent (i_1, h_1) **influences** agent $(i_2, h_2) \Rightarrow$ agent (i_2, h_2) **knows** τ_{i_1, h_1^-} (and a_{i_1, h_1})



- An LTC that is not QC is called **non-classical** (following [Witsenhausen, '75])

Main Results

Hardness and Structural Assumptions

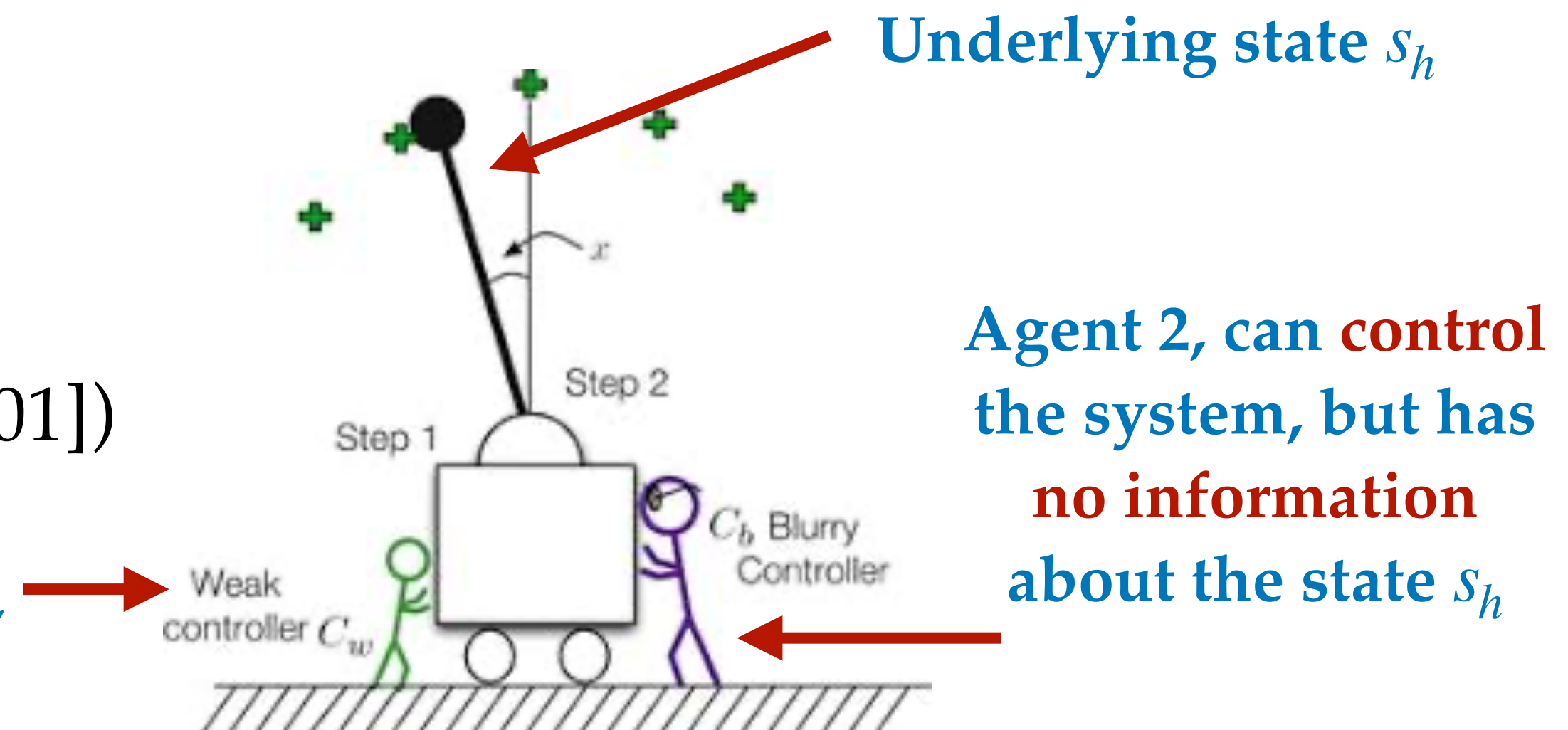
- If LTCs are **non-classical (NC)**, then finding an optimal solution is **PSPACE-hard**

Lemma (Non-classical LTCs are hard). For non-classical LTCs satisfying other assumptions for tractability (to be introduced later), finding an $\frac{\epsilon}{H}$ -team optimum is **PSPACE-hard**.

- Similar idea as the Wistenhausen's counterexample in control [Wistenhausen, '68]

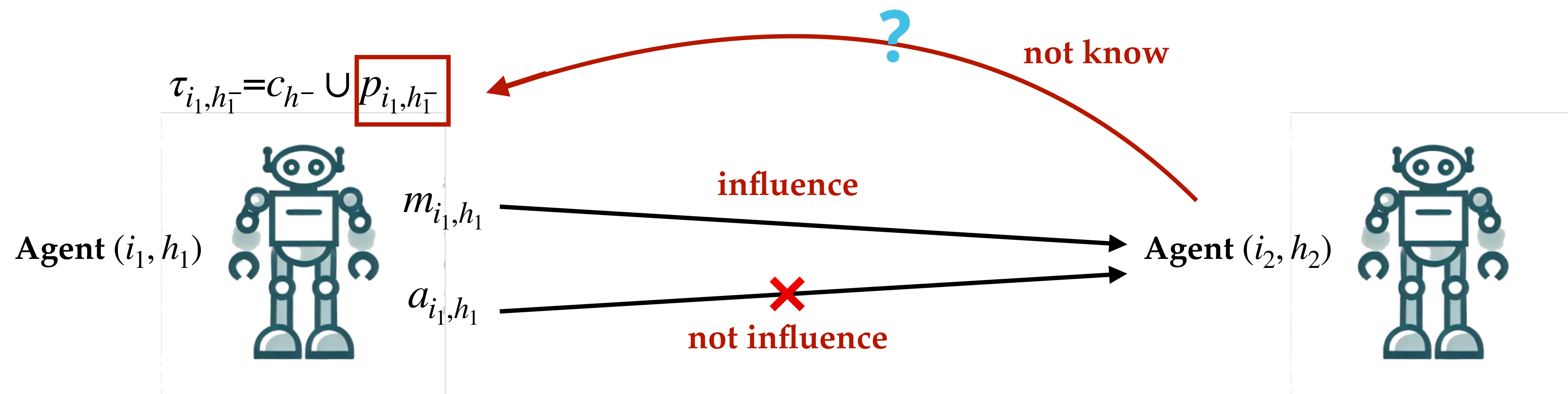
- ▶ One agent has **observability**, but **weak**
- ▶ The other is **strong**, but has **no observability**
- ▶ With high **communication costs**, NC-LTC will not share (thus inherits the hardness of **POMDPs** [Papadimitriou and Tsitsiklis, '87][Lusena et al., '01])

Agent 1, can **see** the underlying state s_h
But **cannot control** the system



Hardness and Structural Assumptions

- This inspired us to focus on **quasi-classical (QC)** LTCs
 - Indeed, **QC information structure** is known to be critical for efficiently solving decentralized control with **linear dynamics** [Ho and Chu, '72][Lamperski and Lessard, '15]...
 - Yet, additional sharing (due to **LTC**) may **break** such QC IS, and introduce **hardness: Two Cases**
- **Case 1:** QC may break due to the effect of **communication strategies**

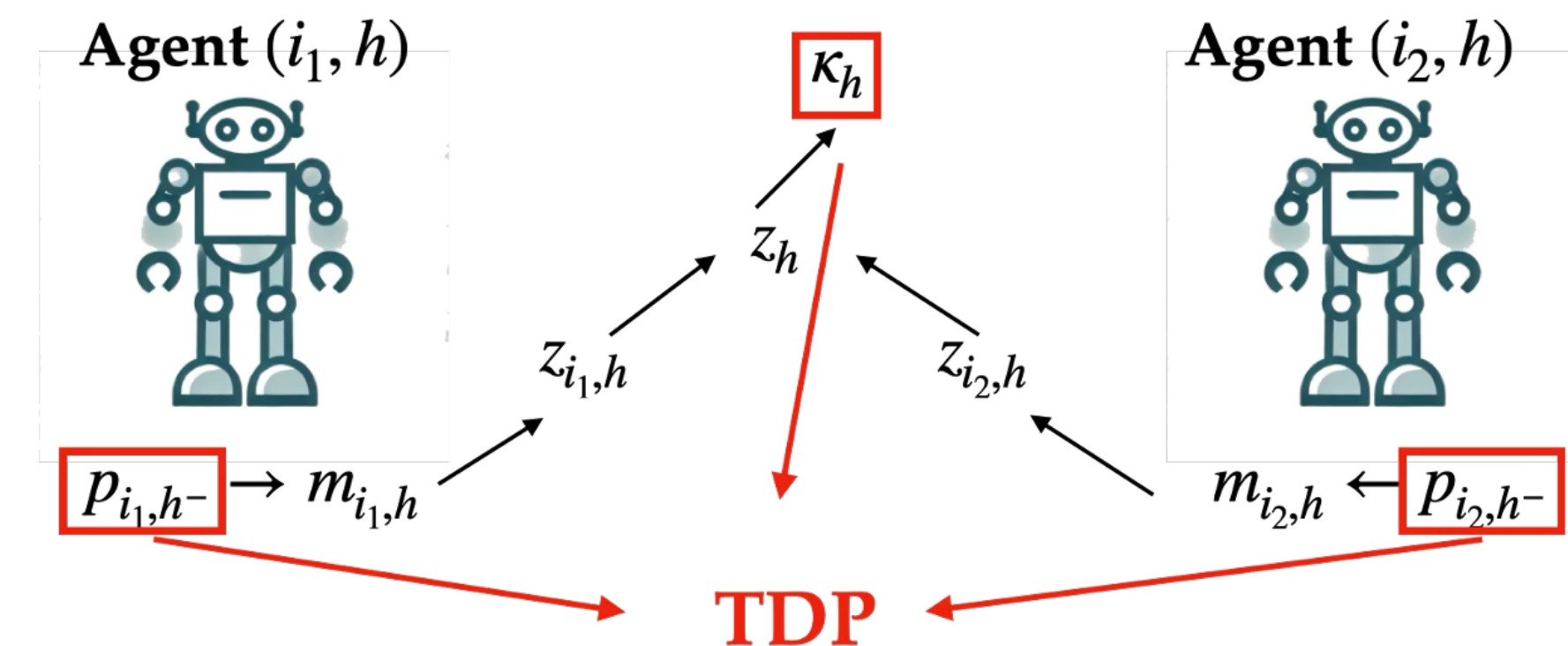


Hardness and Structural Assumptions

- QC breaks since agents use **private info** p_{i,h^-} to choose the communication action $m_{i,h}$

Lemma. For **QC LTCs** satisfying other assumptions for tractability (to be introduced later), finding a team optimum $(g_{1:H}^a, g_{1:H}^m)$ in the general space of $g_{1:h}^m : \mathcal{T}_{i,h^-} \rightarrow \mathcal{M}_{i,h}$ is **NP-hard**.

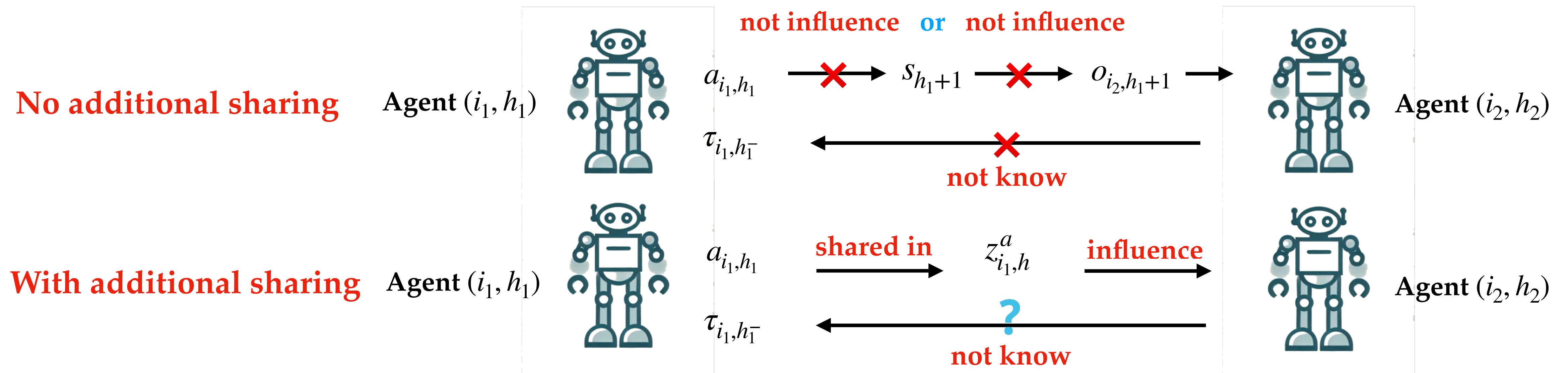
- Insight: **asymmetric/private information** makes the (communication) decision-space **exponentially large**
- Reduce LTC with **private information** as a (**one-stage**) **Team Decision problem** (TDP) [Tsitsiklis and Athans, '85]



Assumption 1 (Common-information-based communication strategies): The communication strategies take only **common information** as input: $\forall i \in [n], h \in [H], g_{i,h}^m : \mathcal{C}_{h^-} \rightarrow \mathcal{M}_{i,h}$.

Hardness and Structural Assumptions

- Case 2: QC may break also due to the effect of **control strategies**
- In **baseline sharing**, **control action** $a_{i,h}$ does **not influence** others since either: i) it does **not influence** state s_{h+1} ; or ii) their **observation** channel is **degenerate**, i.e., **not influencing** $o_{-i,h+1}$
- In **additional sharing**, however, the **control action** $a_{i,h}$ will be **shared** and thus influence others



Hardness and Structural Assumptions

- For the case of $a_{i_1, h_1} \xrightarrow{\times} s_{h_1+1}$:

Assumption 2 (**Control-useless** action is **not used**). $\forall i \in [n], h \in [H]$, if agent i 's action $a_{i,h}$ does not influence the state s_{h+1} , then $\forall h' > h$, it holds that $a_{i,h} \notin \tau_{h'-}, a_{i,h} \notin \tau_{h'+}$.

Lemma: For QC LTCs with other assumptions but **without Assumption 2**, finding a team optimum is **NP-hard**.

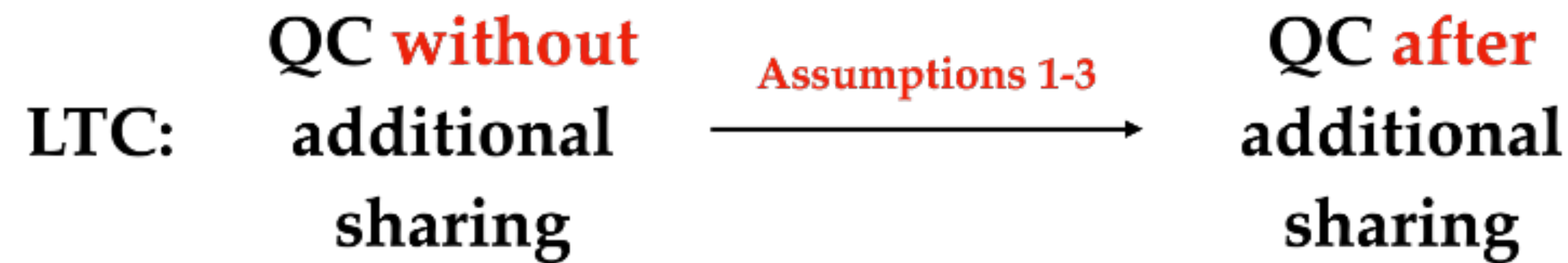
- For the case of $s_{h_1+1} \xrightarrow{\times} o_{i_2, h_1+1}$:

Assumption 3 (Other agent's emissions are **non-degenerate**). $\forall i \in [n], h \in [H]$, $\odot_{-i,h}$ satisfies that, for any $b, b' \in \Delta(\mathcal{S})$ such that $b \neq b'$, we have $\odot_{-i,h}^\top b \neq \odot_{-i,h}^\top b'$.

Lemma: For QC LTCs with other assumptions but **without Assumption 3**, finding an $\frac{\epsilon}{H}$ -team optimum is **PSPACE-hard**.

Hardness and Structural Assumptions

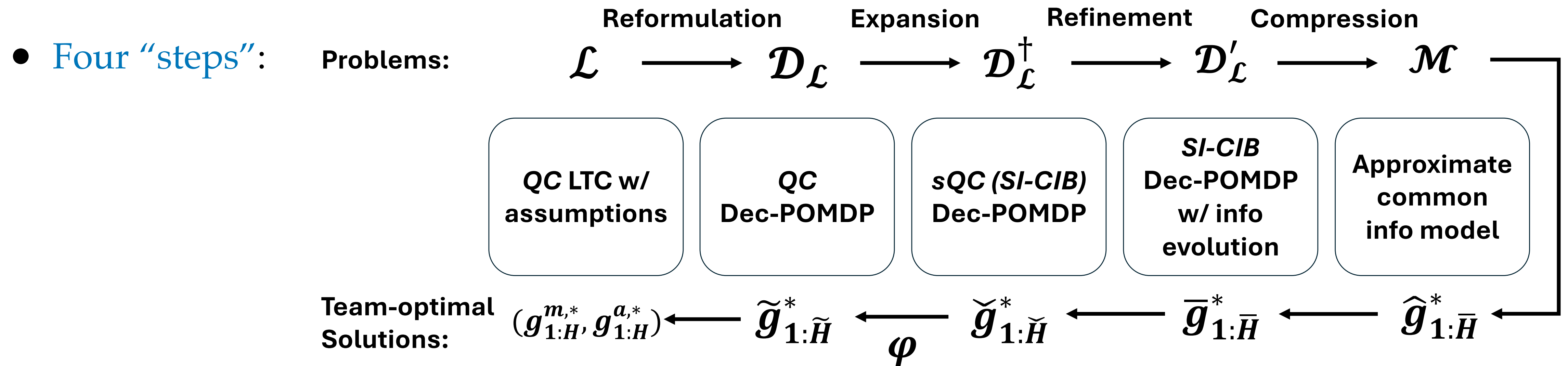
Lemma: For QC LTCs satisfying Assumptions 1-3, additional sharing preserves the QC/sQC information structure.



- In sum, **two justifications** of these structural assumptions:
 - Missing one of them creates **computational hardness** in general
 - Using them (together), one can **preserve** the benign (s)QC information structure

Solving LTCs Efficiently

- But still, how to **solve** LTCs efficiently **even under these assumptions**?
- Planning: **with** model knowledge
- Learning: **without** model knowledge



Step A: Reformulation

- Given any LTC $\mathcal{L}$, we can construct a Dec-POMDP $\mathcal{D}_{\mathcal{L}}$ from $\mathcal{L}$:
 - Split H timesteps into $\tilde{H} = 2H$ timesteps
 - Odd steps $2h - 1$ are **communication** steps: $\tilde{o}_{i,2h-1} = o_{i,2h-1}$, $\tilde{a}_{i,2h-1} = m_{i,h}$, $\tilde{r}_{2h-1} = -\kappa_h$, $\tilde{s}_{2h} = \tilde{s}_{2h-1}$
 - Even steps $2h$ are **control** steps: $\tilde{o}_{2h} = \emptyset$, $\tilde{a}_{i,2h} = a_{i,h}$, $\tilde{r}_{2h} = r_h$, $\tilde{s}_{2h+1} \sim \mathbb{T}_h(\cdot \mid \tilde{s}_{2h}, \tilde{a}_{2h})$
 - Set $\tilde{p}_{i,2h-1} = \emptyset$ due to **Assumption 1** (common-information-based communication)
- $\mathcal{L}$ and $\mathcal{D}_{\mathcal{L}}$ are equivalent $\Rightarrow$ **Solving** $\mathcal{D}_{\mathcal{L}}$ can give a solution of $\mathcal{L}$
- Similar “splitting” was also used in [Sudhakara et al., '21][Kartik et al., '22]

Lemma: For QC LTCs satisfying Assumptions 1-3, the reformulation step **preserves (s)QC IS**, i.e., if $\mathcal{L}$ is (s)QC, then $\mathcal{D}_{\mathcal{L}}$ is (s)QC.

Steps B & C: Expansion & Refinement

- Yet, QC IS of $\mathcal{D}_{\mathcal{L}}$ does not imply computational tractability (esp. discrete-space setting)
 - Even for linear control setting: i) QC $\implies$ convex reformulation [Ho and Chu, '72]; ii) one-step-delay / multi-tree / state-feedback special case $\implies$ sufficient statistics and/or dynamic programming [Yoshikawa, '74][Nayyar and Lessard, '14][Lamperski and Lessard, '15]
- Recently, [Liu and Zhang, '23] solved $\mathcal{D}_{\mathcal{L}}$ without intractable oracles, under the condition of having strategy-independent common-information-based beliefs (SI-CIBs) [Nayyar et al., '13]:

For any c_h and two different strategies $g_{1:h-1}, g'_{1:h-1}$ that can reach c_h :

$$\mathbb{P}_h^{\mathcal{D}_{\mathcal{L}}}(s_h, p_h | c_h, g_{1:h-1}) = \mathbb{P}_h^{\mathcal{D}_{\mathcal{L}}}(s_h, p_h | c_h, g'_{1:h-1})$$

- “One-way” separation: estimation of (s_h, p_h) is independent of control [Nayyar et al., '13]
- Computationally, this is critical as well [Liu and Zhang, '23] — no need to enumerate the strategies in backward induction, which are prohibitively many

Steps B & C: Expansion & Refinement

- Can we connect the (s)QC IS and the SI-CIB condition?
 - **Step B:** We first **expand** the QC $\mathcal{D}_{\mathcal{L}}$ to an sQC $\mathcal{D}_{\mathcal{L}}^{\dagger}$, by additionally sharing the **actions** of the agents whose **information** was already **shared**:

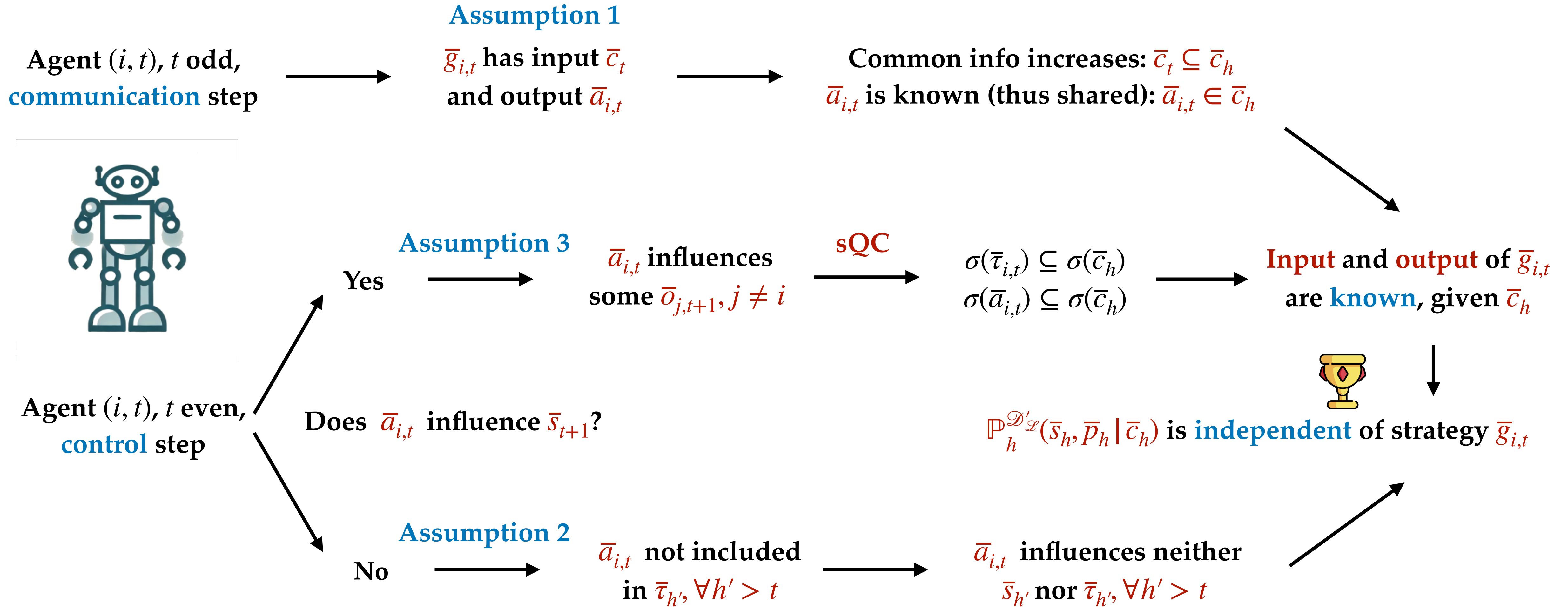
$$\check{c}_h = \tilde{c}_h \cup \left\{ \tilde{a}_{j,t} \mid j \in [n], t < h, \tilde{a}_{j,t} \text{ influences } \tilde{s}_{t+1}, \tilde{\tau}_{j,t} \text{ is included in } \tilde{c}_h \right\}$$
 - **Step C:** We then **refine** the sQC $\mathcal{D}_{\mathcal{L}}^{\dagger}$ by **adding back the private information** at odd timesteps $2h - 1$, obtaining $\mathcal{D}'_{\mathcal{L}}$

Theorem: For a QC LTC $\mathcal{L}$ satisfying Assumptions 1-3, the Dec-POMDP obtained after **Steps A-C**, $\mathcal{D}'_{\mathcal{L}}$, has **SI-CIBs**. Specifically, for any $h \in [\bar{H}]$ and $\bar{g}_{1:h-1}, \bar{g}'_{1:h-1} \in \bar{\mathcal{G}}_{1:h-1}$ that can reach $\bar{c}_h$:

$$\mathbb{P}_h^{\mathcal{D}'_{\mathcal{L}}}(\bar{s}_h, \bar{p}_h \mid \bar{c}_h, \bar{g}_{1:h-1}) = \mathbb{P}_h^{\mathcal{D}'_{\mathcal{L}}}(\bar{s}_h, \bar{p}_h \mid \bar{c}_h, \bar{g}'_{1:h-1}).$$

Moreover, solving the **best-in-class strategy** for the **SI-CIB** $\mathcal{D}'_{\mathcal{L}}$ yields an optimal strategy of $\mathcal{L}$.

Proof Idea: $\mathcal{D}'_{\mathcal{L}}$ has SI-CIBs (**sQC** $\Rightarrow$ **SI-CIB**)



Planning and Learning in QC LTCs

- Now, $\mathcal{D}'_{\mathcal{L}}$ can be solved by the **planning** and **learning** algorithms in [Liu and Zhang, '23]
 - Under the **γ -observability assumption** [Even-Dar et al., '07][Golowich et al., '23][Liu and Zhang, '23], one can **truncate the common information** of length $\Theta(\log(SH/\epsilon))$ to construct an **expected** approximate-information-state model $\widehat{\mathcal{M}}$
 - Due to **SI-CIB** (we just proved), one can solve $\widehat{\mathcal{M}}$ via a **dynamic program** of **tractable size**
 - For **each step** of the backward induction in $\widehat{\mathcal{M}}$, it is a **Team Decision Problem** in general
 - For **odd** (communication) steps, they are tractable since there is **no private information**
 - For **even** (control) steps, one need a “**one-step tractability**” assumption
 - The **combined** effect of both $\mathbb{T}_h$ and $\mathbb{O}_h$
 - Can be satisfied by several **QC LTC** examples, e.g., one-step delayed information/observation sharing, single-controller with asymmetric delays, one-step delayed observation and two-step delayed control... see [Liu and Zhang, '23] and the paper

Planning and Learning in QC LTCs

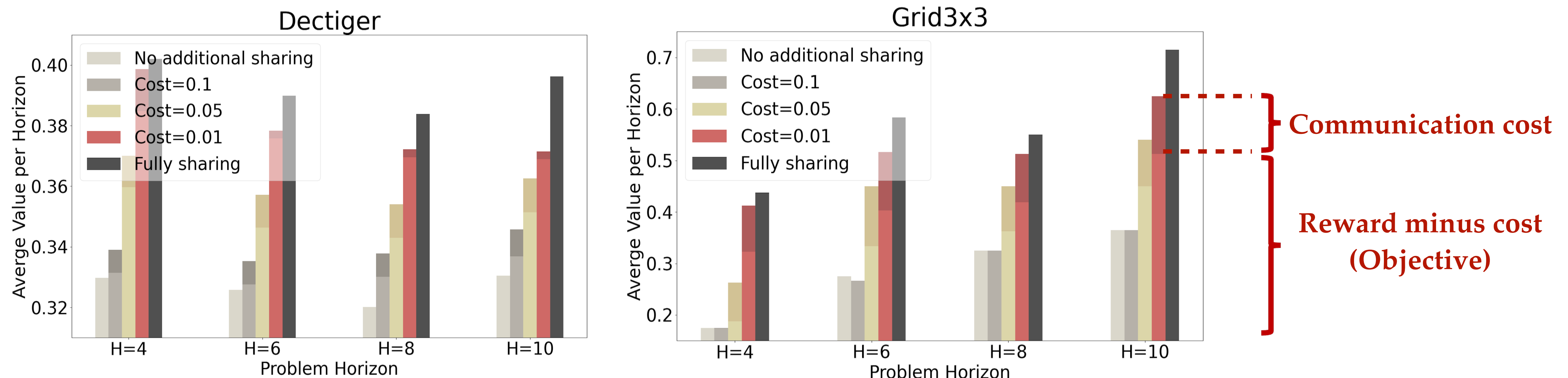
- Now, $\mathcal{D}'_{\mathcal{L}}$ can be solved by the **planning** and **learning** algorithms in [Liu and Zhang, '23]

Theorem: Suppose a QC LTC $\mathcal{L}$ satisfies Assumptions 1-3, and the joint observation channel is γ -observable, then there **exists an LTC algorithm** that can learn an ϵ -optimal strategy with high probability, **with time and sample complexities polynomial in the parameters of $\widehat{\mathcal{M}}$** , if the learned $\widehat{\mathcal{M}}$ satisfies one-step tractability.

In particular, for the **QC LTC examples** above, such an LTC algorithm can learn an ϵ -optimal strategy with high probability, with **quasi-polynomial time** and **sample complexities**, thus **without computationally intractable oracles**.

Experimental Results

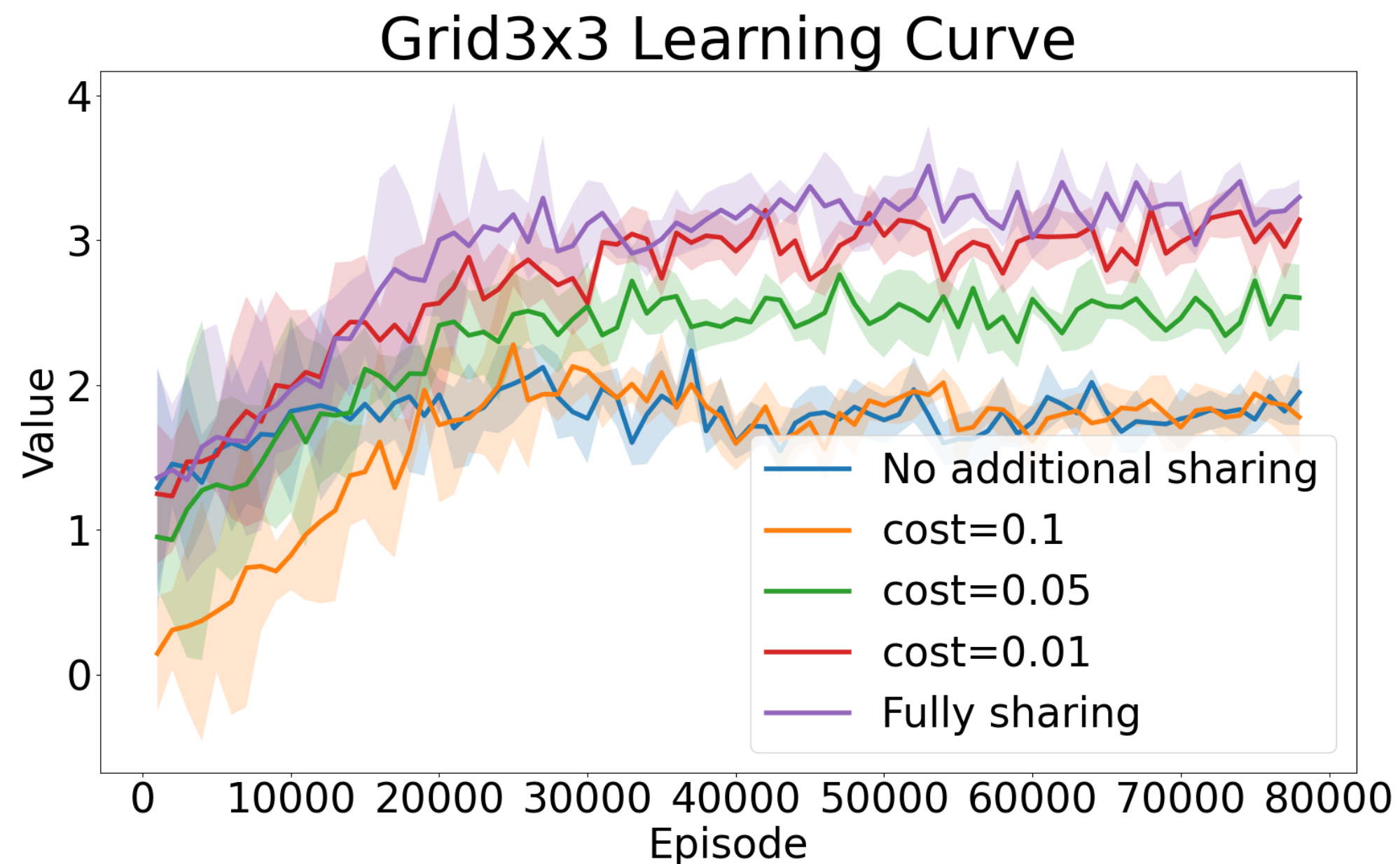
- We implement our LTC algorithm in the two benchmark Dec-POMDP environments: [Dectiger](#) [Nair et al., '03] and [Grid3x3](#) [Amato et al., '09]
- With different **communication cost** functions and **horizons**



Cheaper communication encourages agents to **share more information**,
and achieves **higher rewards** and overall **objectives**

Experimental Results

- We implement our LTC algorithm in the two benchmark Dec-POMDP environments: [Dectiger](#) [Nair et al., '03] and [Grid3x3](#) [Amato et al., '09]
- With different **communication cost** functions and **horizons**



Results for Dectiger

Horizon/Cost	No Sharing	Cost=0.1	Cost=0.05	Cost=0.01	Fully Sharing
H=4 w/ cost	1.32±0.025	1.33±0.044	1.44±0.034	1.54±0.013	1.57±0.004
H=4 w/o cost	-	1.36±0.032	1.48±0.034	1.59±0.002	-
H=6 w/ cost	1.95±0.009	1.97±0.07	2.08±0.068	2.26±0.012	2.29±0.002
H=6 w/o cost	-	2.01±0.047	2.14±0.072	2.27±0.011	-
H=8 w/ cost	2.56±0.041	2.64±0.078	2.74±0.118	2.96±0.021	3.0±0.002
H=8 w/o cost	-	2.7±0.044	2.83±0.117	2.98±0.02	-
H=10 w/ cost	3.31±0.024	3.37±0.135	3.51±0.153	3.69±0.029	3.87±0.007
H=10 w/o cost	-	3.46±0.069	3.63±0.152	3.71±0.026	-

Cheaper communication encourages agents to **share more information**,
and achieves **higher rewards** and overall **objectives**

Extension to General Dec-POMDPs

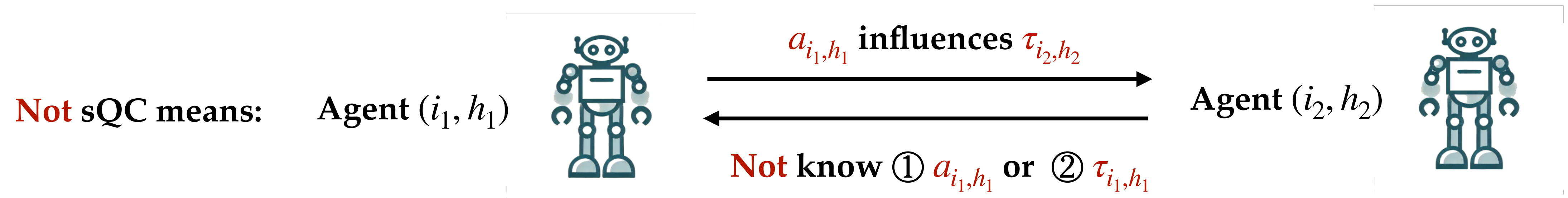
- We have shown **QC** $\implies$ **SI-CIB** for LTC problems
- Are **sQC** and **SI-CIB** conditions related for **general** Dec-POMDPs?

Theorem: Consider a Dec-POMDP $\mathcal{D}$ with $\tau_{i,h} \subseteq \tau_{i,h+1}$ (τ -recall) and $o_{i,h} \in \tau_{i,h}$ (**closed-loop** information structure) for all $h \in [H]$ and $i \in [n]$. If $\mathcal{D}$ is **sQC** and satisfies **Assumptions 2-3**, then it **has SI-CIBs**; if $\mathcal{D}$ **has SI-CIBs** and **perfect recall**, then $\mathcal{D}$ is **sQC** (up to null sets).

- **Perfect recall** [Kuhn, '53] here requires to remember **all past actions** $\neq$ τ -recall
- “Up to null sets” means except the null sets generated by any strategy $g_{1:H}$ in $\mathcal{D}$
- Proof for **QC** $\implies$ **SI-CIB** is similar as the LTC case

Proof Idea: **SI-CIB** $\Rightarrow$ **sQC** (up to null set)

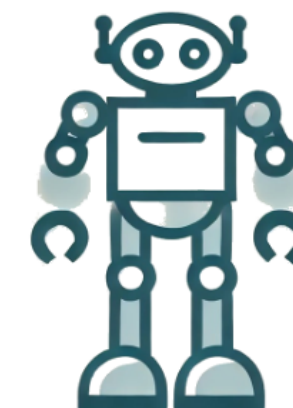
Prove **by contradiction**: Suppose the problem is SI-CIB but **not** sQC.



① Agent (i_2, h_2) does not know a_{i_1, h_1}

Two realizations

$a_{i_1, h_1}, a'_{i_1, h_1}$



Agent (i_2, h_2)

?

a_{i_1, h_1}

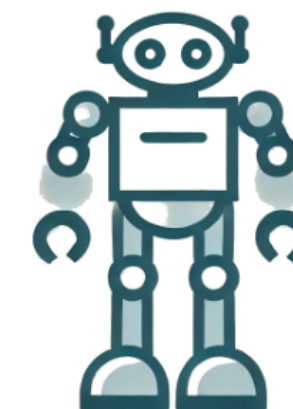
Or

a'_{i_1, h_1}

② Agent (i_2, h_2) does not know τ_{i_1, h_1}

Two realizations

$\tau_{i_1, h_1}, \tau'_{i_1, h_1}$



Agent (i_2, h_2)



a_{i_1, h_1}

?

τ_{i_1, h_1}

Or

τ'_{i_1, h_1}

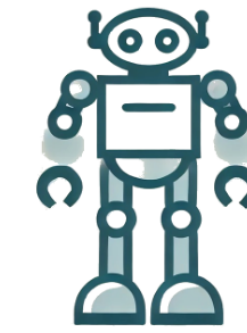
Proof Idea: **SI-CIB** $\Rightarrow$ **sQC** (up to null set)

① Agent (i_2, h_2) does not know a_{i_1, h_1}

- Consider some $g_{1:h_2-1}$ that can reach c_{h_2}

Two realizations

$a_{i_1, h_1}, a'_{i_1, h_1}$



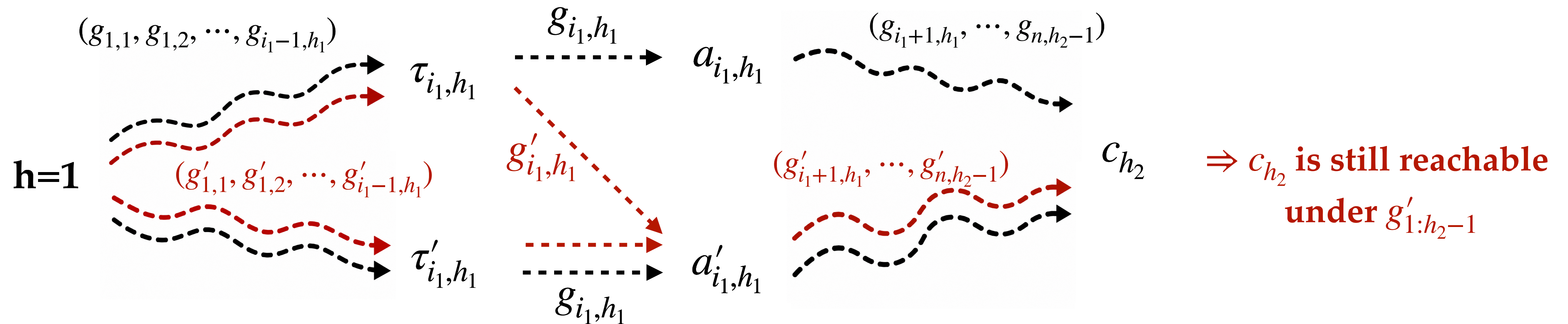
Agent (i_2, h_2) ?

a_{i_1, h_1}

Or

a'_{i_1, h_1}

- We construct another $g'_{1:h_2-1}$ as: $g'_{i_1, h_1}(\cdot) \equiv a'_{i_1, h_1}$; others remain the same



- Perfect recall $\Rightarrow a_{i_1, h_1} \in p_{h_2}$ and $\mathbb{P}_{h_2}^{\mathcal{D}}(s_{h_2}, p_{h_2} | c_{h_2}, g_{1:h_2-1}) > 0$

- However, $\mathbb{P}_{h_2}^{\mathcal{D}}(s_{h_2}, p_{h_2} | c_{h_2}, g'_{1:h_2-1}) = 0$ (a_{i_1, h_1} does not occur under $g'_{1:h_2-1}$)

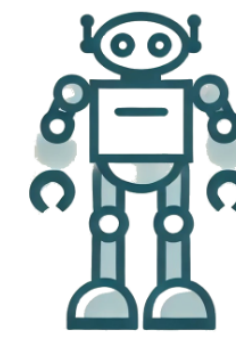
} **✗** SI-CIB

Proof Idea: **SI-CIB** $\Rightarrow$ **sQC** (up to null set)

② Agent (i_2, h_2) does not know τ_{i_1, h_1}

Two realizations

$\tau_{i_1, h_1}, \tau'_{i_1, h_1}$



Agent (i_2, h_2)  a_{i_1, h_1} ?

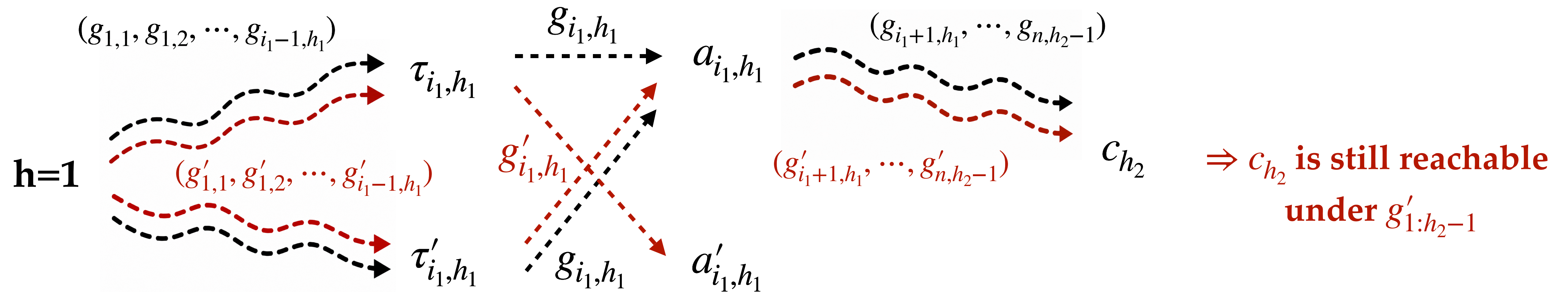
τ_{i_1, h_1}

Or

τ'_{i_1, h_1}

• Consider some $g_{1:h_2-1}$ that can reach c_{h_2}

• We construct another $g'_{1:h_2-1}$ as: $\begin{cases} g'_{i_1, h_1}(\tau_{i_1, h_1}) = a'_{i_1, h_1} \\ g'_{i_1, h_1}(\tau'_{i_1, h_1}) = a_{i_1, h_1} \end{cases}$; others remain the same



• Perfect recall $\Rightarrow (\tau_{i_1, h_1} \setminus c_{h_2}) \in p_{h_2}$ and $\mathbb{P}_{h_2}^{\mathcal{D}}(s_{h_2}, p_{h_2} \mid c_{h_2}, g_{1:h_2-1}) > 0$

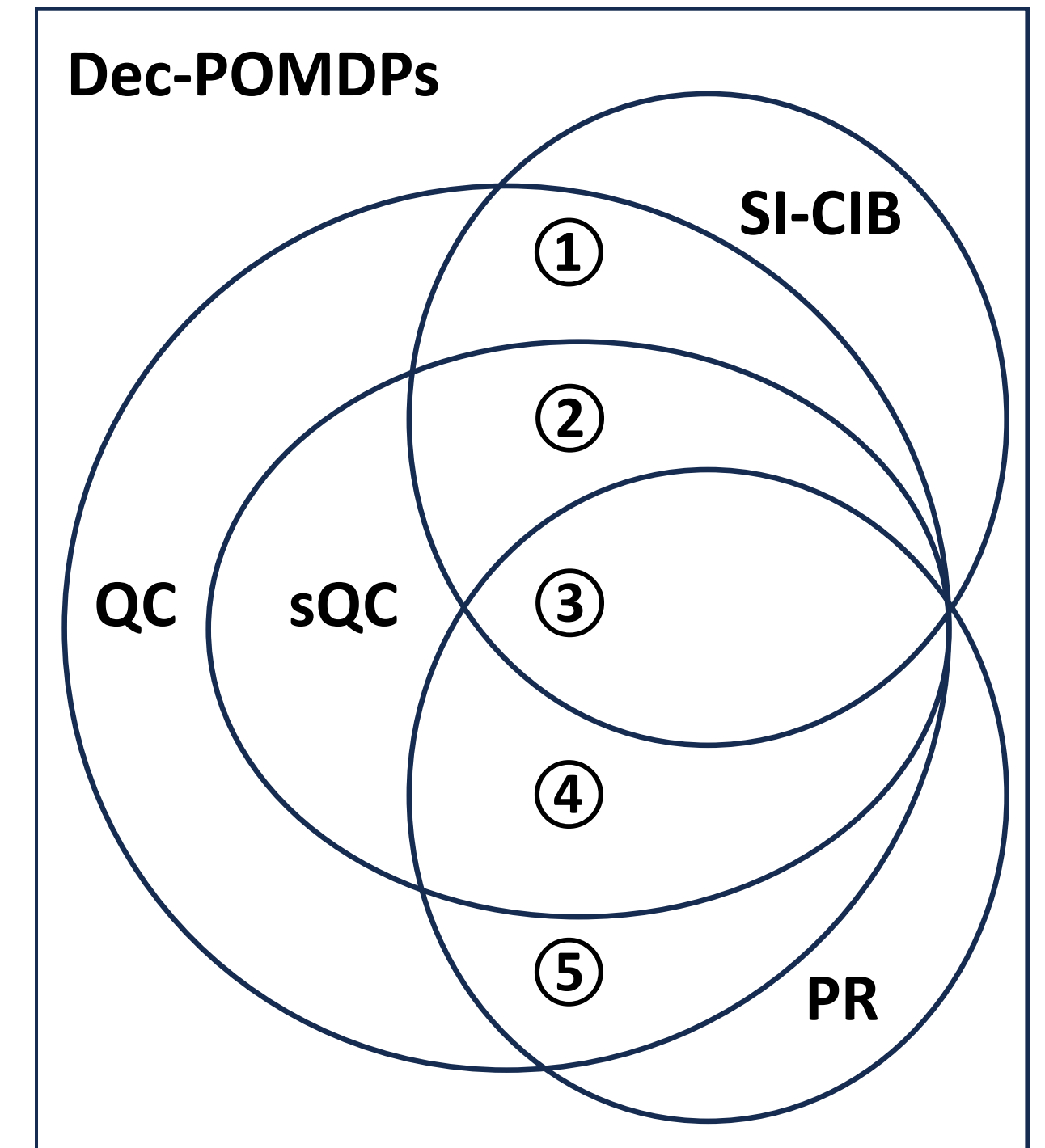
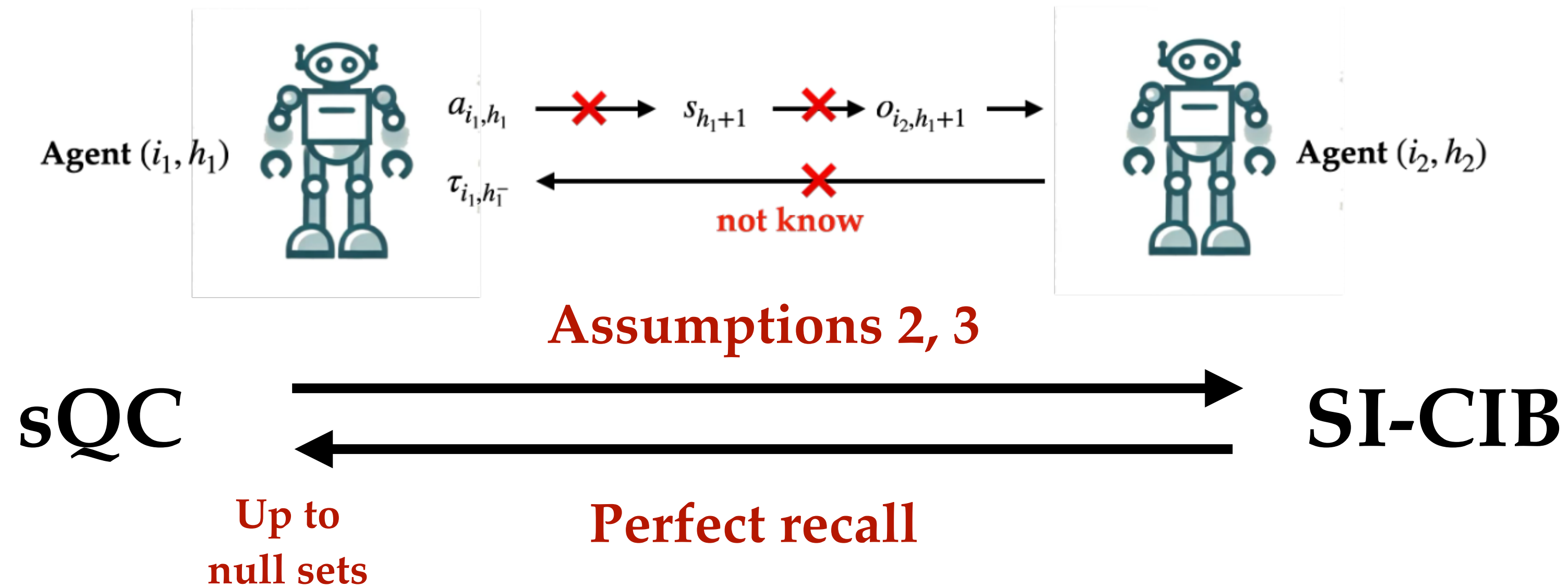
• However, $\mathbb{P}_{h_2}^{\mathcal{D}}(s_{h_2}, p_{h_2} \mid c_{h_2}, g'_{1:h_2-1}) = 0$ (τ_{i_1, h_1} and a_{i_1, h_1} cannot occur simultaneously)

SI-CIB

Additional Remark on **SI-CIB** $\Rightarrow$ **sQC**

- Importance of **Perfect Recall**:
 - Ensure $\tau_{i,t} \subseteq \tau_h$ and $a_{i,t} \in \tau_h$ for $t \leq h$
 - The information of $\tau_{i,h}, a_{i,h}$ is either in c_h or p_h
- In general, **without** Perfect Recall, **SI-CIB** $\nRightarrow$ **sQC**:
 - Counterexample: Multi-agent **Markov** decision process
 - $\tau_{i,h} = c_h = s_h, \quad p_h = \emptyset$
 - **Non-sQC**: $\tau_{i,h-1} \not\subseteq \tau_{j,h}$ and $a_{i,h-1} \notin \tau_{j,h}$
 - **SI-CIB**: $\mathbb{P}(s_h, p_h | c_h) = \mathbf{1}[s_h = c_h]$, **independent** of $g_{1:h-1}$

Extension to General Dec-POMDPs: Summary



- Relationship between conditions
 - (s)QC were defined under **intrinsic models** [Witsenhausen, '75]
 - SI-CIBs were defined under **state-space models** [Nayyar et al., '13]
 - Perfect recall (PR) were introduced for **extensive-form games** [Kuhn, '53]

Extension to Linear-Quadratic-Gaussian Setting

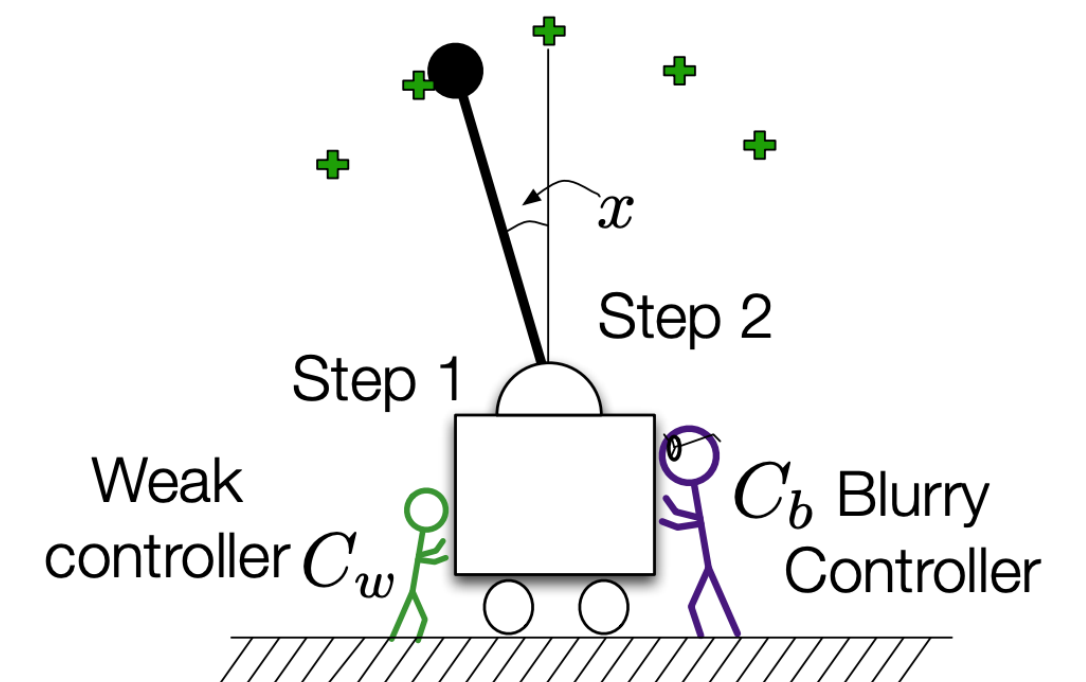
- Consider LTC $\mathcal{L}$ under **Decentralized Linear-Quadratic-Gaussian** (Dec-LQG) control framework
 - State, action, and observation spaces are continuous and real-vector valued:
 - $s_h \in \mathbb{R}^{d_s}, a_h \in \mathbb{R}^{d_a}, o_h \in \mathbb{R}^{d_o}$
 - **Linear** dynamics with **Gaussian** noises:
 - $s_{h+1} = A_h s_h + B_h a_h + w_{0,h}; \forall i \in [n], o_{i,h} = E_{i,h} s_h + w_{i,h}$
 - $\forall i \in \{0, 1, \dots, n\}, w_{i,h}$ is **Gaussian** noise
 - **Quadratic** reward (minus cost):
 - $r_h(s_h, a_h) = - (s_h^\top Q_h^1 s_h + a_h^\top Q_h^2 a_h)$

Hardness and Assumptions for LQG Setting

- **Tractability** is now oftentimes associated with the **linearity** of optimal strategies: $g_{i,h}^{a,\star}(\tau_{i,h+}) \stackrel{?}{=} G_{i,h}^{\star} \tau_{i,h+}$
- **Assumptions** are required and similar **hardness** results hold:
 - **Assumption 1:** **QC** information structure

Assumption 1: $\mathcal{L}$ has a **QC** information structure (with baseline sharing).

- Without it, the optimal (control) strategy $g_{1:h}^{a,\star}$ can be **non-linear**
- Reduced to the Witsenhausen's counterexample [Witsenhausen '68] under **high communication costs**



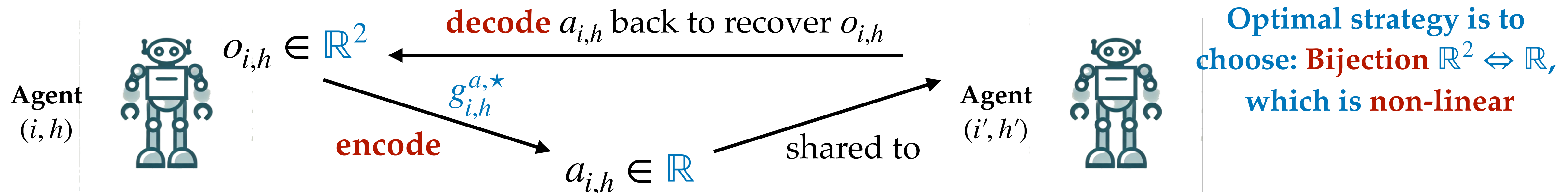
Witsenhausen's Counterexample

Hardness and Assumptions for LQG Setting

- **Tractability** is now oftentimes associated with the **linearity** of optimal strategies: $g_{i,h}^{a,\star}(\tau_{i,h+}) \stackrel{?}{=} G_{i,h}^{\star} \tau_{i,h+}$
- **Assumptions** are required and similar **hardness** results hold:
 - **Assumption 2:** **Control-useless** action is not **used**

Assumption 2: $\forall i \in [n], h \in [H]$, if action $a_{i,h}$ does not influence s_{h+1} , then $\forall h' > h, a_{i,h} \notin \tau_{h'-}, a_{i,h} \notin \tau_{h'+}$.

- Without it, two issues may arise: 1) **non-linearity** of $g_{1:h}^{a,\star}$
 - $g_{i,h}^{a,\star}$ may use **control-useless** $a_{i,h}$ to **encode** $o_{i,h}$ and **share** it with others



Hardness and Assumptions for LQG Setting

- **Tractability** is now oftentimes associated with the **linearity** of optimal strategies: $g_{i,h}^{a,\star}(\tau_{i,h+}) \stackrel{?}{=} G_{i,h}^{\star} \tau_{i,h+}$
- **Assumptions** are required and similar **hardness** results hold:
 - **Assumption 2:** **Control-useless** action is not **used**

Assumption 2: $\forall i \in [n], h \in [H]$, if action $a_{i,h}$ does not influence s_{h+1} , then $\forall h' > h, a_{i,h} \notin \tau_{h'-}, a_{i,h} \notin \tau_{h'+}$.

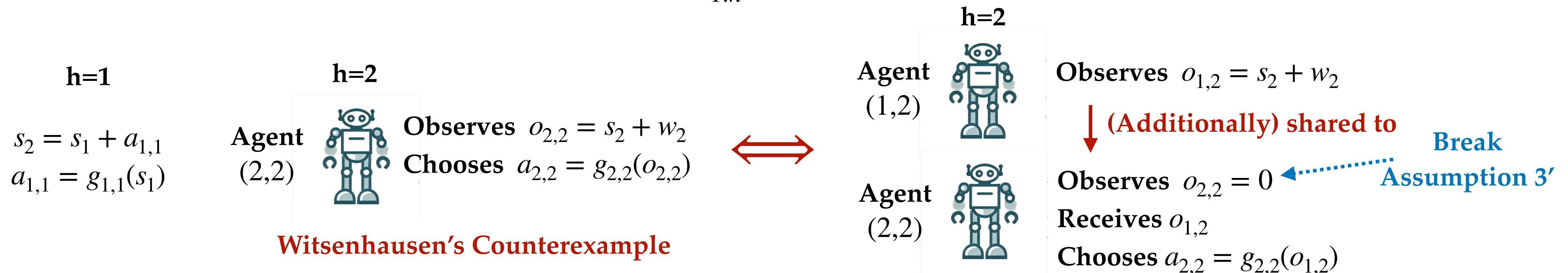
- Without it, two issues may arise: 2) **non-existence** of $(g_{1:h}^{a,\star}, g_{1:h}^{m,\star})$
 - $g_{i,h}^{a,\star}$ may use **control-useless** $a_{i,h}$ to **encode** $o_{i,h}$ and **share** it with others
 - $a_{i,h}$ is **control-useless** $\Rightarrow$ **Scaling** $a_{i,h}$, e.g., $a'_{i,h} = a_{i,h}/2$
 - 1) State **does not change**; 2) $a'_{i,h}$ is as **informative** as $a_{i,h}$; 3) Control cost becomes **smaller**
 - One can scale $a_{i,h}$ **arbitrarily small** $\Rightarrow$ **non-existence**

Hardness and Assumptions for LQG Setting

- **Tractability** is now oftentimes associated with the **linearity** of optimal strategies: $g_{i,h}^{a,\star}(\tau_{i,h+}) \stackrel{?}{=} G_{i,h}^{\star} \tau_{i,h+}$
- **Assumptions** are required and similar **hardness** results hold:
 - **Assumption 3'**: Other agent's observation channels can **observe control-relevant** state components

Assumption 3': $\forall i \in [n], h \in [H], \text{rank}(B_{i,h}) = \text{rank}(E_{-i,h+1}B_{i,h})$.

- **Relaxed Assumption 3** before (Other agent's observations are **non-degenerate**, i.e., $E_{-i,h+1}$ is full-rank)
- Without it, the optimal (control) strategy $g_{1:h}^{a,\star}$ can be **non-linear**



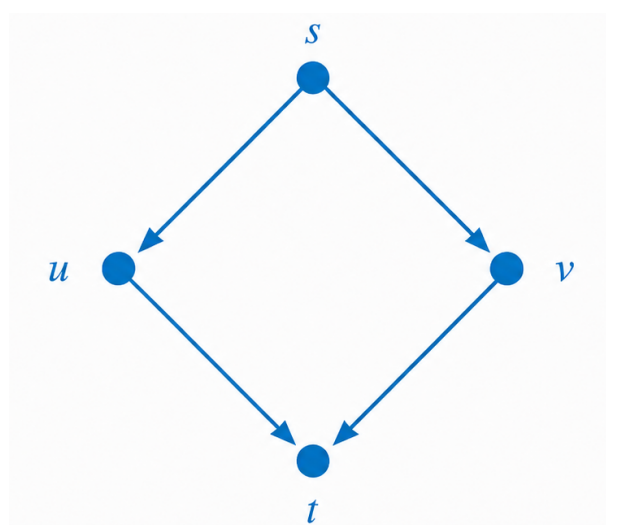
Existing Results for Dec-LQG

- Similar as before, we can reformulate $\mathcal{L}$ as a **Dec-LQG** problem $\mathcal{D}$
 - Under **Assumptions 1, 2, and 3'**: $\mathcal{D}$ has a **QC** information structure
- Existing approaches to solving (QC) **Dec-LQG** problems:
 - **Quasi-classical** IS [Ho and Chu, '72]: $g_{1:h}^\star$ is **linear**, $\mathcal{D}$ can be reformulate as a **convex** program
 - It may not be **efficiently computed**, as the program can be **large** [Lamperski and Lessard, '15]
 - **One-step-delay** sharing (thus **QC**) [Yoshikawa, '74]: **dynamic programming** can be derived
 - **Multi-tree** (thus **QC**) IS: [Nayyar and Lessard, '14]:

The **sufficient statistics** and **optimal strategy form** can be derived with

1) decoupled state (i.e., $s = (s_1, \dots, s_n)$); 2) **multi-tree** graph information structure

**Multi-Tree: No
diamond in graph**



Existing Results for Dec-LQG

- **Quasi-classical** (due to sparsity and delays) IS: [Lamperski and Lessard, '15] can solve it by **dynamic programming / Riccati Equations** with
 - 1) decoupled state (i.e., $s = (s_1, \dots, s_n)$); 2) **state-feedback** (agent i can observe s_i)
- **Common-information-based** framework [Mahajan and Nayyar, '15]: **Fix** the **private-information part** of the strategy, i.e., **fix** $\mathcal{F}_{i,h}$ in $g_{i,h}(c_h, p_{i,h}) = \mathcal{E}_{i,h} c_h + \mathcal{F}_{i,h} p_{i,h}$ it reduced to a **centralized** LQG
 - **Sufficient statistics** and **dynamic program** can be derived
 - Can solve **non-classical** IS and **best-in-(linear-)class** strategies
 - However, **joint** optimization of $(\mathcal{E}_{i,h}, \mathcal{F}_{i,h})_{i \in [n], h \in [H]}$ can be **non-convex** [Mahajan and Nayyar, '15]

Extension of Our Pipeline to LQG Setting

- Under Assumptions 1, 2, and 3': **Expand** $\mathcal{D}$ as $\tilde{\mathcal{D}}$, with **sQC** information structure, and satisfies **SI-CIB**
 - $\tilde{\mathcal{D}}$ has **SI-CIB** [Gupta et al., '14]: $\mathbb{P}(s_h, p_h | c_h)$ is **Gaussian** with mean Θ_h , covariance Σ_h
 - $\Theta_h = \Psi_h^1(\Theta_{h-1}, z_h)$, $\Sigma_h = \Psi_h^2(\Sigma_h)$, where (Ψ_h^1, Ψ_h^2) are **independent** of strategies and Ψ_h^1 linear
 - $\Rightarrow \exists g_{1:h}^\star$ based on conditional mean Θ_h : $g_{i,h}^\star(\tau_{i,h}) = f_{i,h}(\Theta_h, p_{i,h})$
 - $\tilde{\mathcal{D}}$ is **sQC** $\Rightarrow \exists g_{1:h}^\star$ **linear**: $g_{i,h}^\star(\tau_{i,h}) = G_{i,h}^\star \tau_{i,h}$ [Ho and Chu, '72]
- } the **same** $g_{i,h}^\star$?

Theorem: There exists an optimal strategy $g_{1:h}^\star$ of $\tilde{\mathcal{D}}$ that is **linear** in $(\Theta_h, p_{i,h})$:

$$g_{i,h}^\star(c_h, p_{i,h}) = \mathcal{E}_{i,h} \Theta_h + \mathcal{F}_{i,h} p_{i,h}.$$

- Combining the insights from [Ho and Chu, '72] and [Mahajan and Nayyar, '15]
- Solving $g_{1:H}^\star$ in $\tilde{\mathcal{D}} \Rightarrow$ Finding optimal $(\mathcal{E}_{i,h}^\star, \mathcal{F}_{i,h}^\star)_{i \in [n], h \in [H]}$

Extension of Our Pipeline to LQG Setting

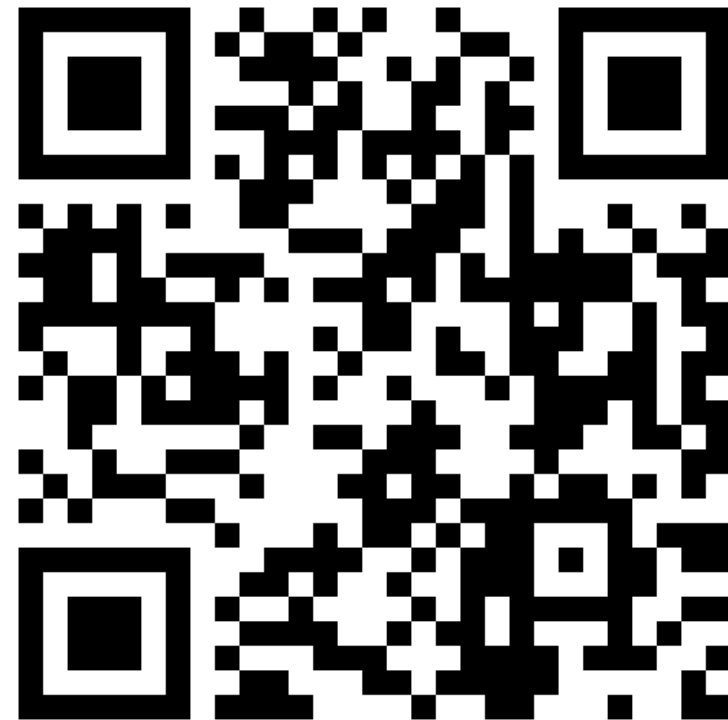
- How to compute the Θ_h, Σ_h ? Recall that $\Theta_h = \Psi_h^1(\Theta_{h-1}, z_h)$, $\Sigma_h = \Psi_h^2(\Sigma_h)$
 - Under **SI-CIB** condition, Ψ_h^1, Ψ_h^2 should be **independent** of $\{\mathcal{F}_{i,h}^\star\}_{i \in [n], h \in [H]}$
 - Indeed, under **Assumptions 2** and **3'**, we can reduce the **sQC** $\tilde{\mathcal{D}}$ to another centralized LQG problem with system dynamics **independent** of $\{\mathcal{F}_{i,h}^\star\}_{i \in [n], h \in [H]}$
 - Ψ_h^1, Ψ_h^2 can be computed by **Kalman Filter** via **forward Riccati Equations**
- How to compute the optimal $(\mathcal{E}_{i,h}^\star, \mathcal{F}_{i,h}^\star)_{i \in [n], h \in [H]}$?

Theorem: Let $\tilde{\mathcal{D}}$ be a Dec-LQG with **SI-CIB** and **sQC** information structure. Then, the **common-information-based** value function can be written as $V_h(\Theta_h) = \Theta_h^\top P_h \Theta_h + C_h$, where C_h is a constant that does not depend on Θ_h . Moreover, $P_h, C_h, \mathcal{E}_{i,h}^\star, \mathcal{F}_{i,h}^\star$ can be computed by solving **backward Riccati Equations**.

Takeaways

- **Learning-to-communicate (LTC)** problem in multi-agent RL was **formulated** and **solved** through the lens of **information structures (ISs)** in (decentralized) stochastic control
- Structural **conditions** were developed, under which **quasi-classical IS** can be preserved, and planning/learning in LTCs can be done **without computationally intractable oracles**
 - Key: **(strictly) quasi-classical** LTC $\mathcal{L}$ can be connected to the **strategy-independent common-information-based belief (SI-CIB)** condition of the reformulated Dec-POMDP $\mathcal{D}'_{\mathcal{L}}$
- We extended the relationship between **sQC IS** and **SI-CIB** condition to **general** Dec-POMDPs
- We also extended LTC to LQG settings, which can be solved by **Kalman Filter** and **Riccati Equations** under the **QC IS** and similar **structural assumptions** above
- Overall, LTC can/should be investigated with more **decentralized stochastic control** principles :)

Thanks!



**[LYZ (α - β), '26] Principled Learning-to-Communicate with
Quasi-Classical Information Structures**



**[YZ, '26] Joint Communication-Control Strategy
Optimization with Partially Nested Information
Structures: The Linear-Quadratic Case**

Riccati Equations

Theorem: Let $\tilde{\mathcal{D}}$ be a Dec-LQG with SI-CIB and sQC information structure. Then, the value function can be written as $V_h(c_h) = V_h(\Theta_h) = \Theta_h^\top R_h \Theta_h + C_h$, where C_h is a constant that does not depend on Θ_h . Moreover, $R_h, C_h, \mathcal{E}_{i,h}^*, \mathcal{F}_{i,h}^*$ can be computed via **Riccati Equations**.

- From information evolution, we have $\Theta_{h+1} = K_{h+1}^1 \Theta_h + K_{h+1}^2 p_h + K_{h+1}^3 a_h + K_{h+1}^4 o_{h+1}$, then

$$R_h = (\hat{\mathcal{E}}_h^*)^\top Q_h^2 \hat{\mathcal{E}}_h^* + (L_h^1 \hat{\mathcal{E}}_h^* + L_h^2)^\top R_{h+1} (L_h^1 \hat{\mathcal{E}}_h^* + L_h^2) + \mathbb{I}_{s,h}^\top Q_h^1 \mathbb{I}_{s,h}, \quad \mathcal{F}_h^* \in \operatorname{argmin}_{\mathcal{F}_h = \operatorname{diag}(\mathcal{F}_{1,h}, \dots, \mathcal{F}_{n,h})} \operatorname{Tr}(\mathcal{F}_h \mathbb{I}_{p,h} \Sigma_h \mathbb{I}_{p,h}^\top \mathcal{F}_h^\top Q_h^2)$$

$$L_h^1 = K_{h+1}^3 + K_{h+1}^4 E_{h+1} B_h, \quad + \operatorname{Tr}((L_h^3 \mathcal{F}_h \mathbb{I}_{p,h} + L_h^4) \Sigma_h (L_h^3 \mathcal{F}_h \mathbb{I}_{p,h} + L_h^4)^\top R_{h+1}),$$

$$L_h^2 = K_{h+1}^1 + K_{h+1}^2 \mathbb{I}_{p,h} + K_{h+1}^4 E_{h+1} A_h \mathbb{I}_{s,h}, \quad \mathcal{F}_h^* =: \operatorname{diag}(\mathcal{F}_{1,h}^*, \dots, \mathcal{F}_{n,h}^*),$$

$$L_h^3 = K_{h+1}^3 + K_{h+1}^4 E_{h+1} B_h, \quad C_h = \operatorname{Tr}(\mathcal{F}_h^* \mathbb{I}_{p,h} \Sigma_h \mathbb{I}_{p,h}^\top (\mathcal{F}_h^*)^\top Q_h^2) + \operatorname{Tr}(\mathbb{I}_{s,h} \Sigma_h \mathbb{I}_{s,h}^\top Q_h^1)$$

$$L_h^4 = K_{h+1}^2 \mathbb{I}_{p,h} + K_{h+1}^4 E_{h+1} A_h \mathbb{I}_{s,h}, \quad + \operatorname{Tr}((L_h^3 \mathcal{F}_h^* \mathbb{I}_{p,h} + L_h^4) \Sigma_h (L_h^3 \mathcal{F}_h^* \mathbb{I}_{p,h} + L_h^4)^\top R_{h+1}) +$$

$$\hat{\mathcal{E}}_h^* = -(Q_h^2 + (L_h^1)^\top R_{h+1} L_h^1)^\dagger (L_h^1)^\top R_{h+1} L_h^2, \quad \operatorname{Tr}(K_{h+1}^4 \operatorname{diag}(\Sigma_{1,h+1}, \dots, \Sigma_{n,h+1}) (K_{h+1}^4)^\top R_{h+1}) +$$

$$\mathcal{E}_h^* = \hat{\mathcal{E}}_h^* - \operatorname{diag}(\mathcal{F}_{1,h}^*, \mathcal{F}_{2,h}^*, \dots, \mathcal{F}_{n,h}^*) \mathbb{I}_{p,h}. \quad \operatorname{Tr}(K_{h+1}^4 E_{h+1} \Sigma_{0,h} E_{h+1}^\top (K_{h+1}^4)^\top R_{h+1}) + C_{h+1}.$$
- $\mathbb{I}_{s,h}, \mathbb{I}_{p,h}$ are the **identity projection matrices** that project vectors (s_h, p_h) to the corresponding vectors s_h and p_h