Collective Cooperative Intelligence

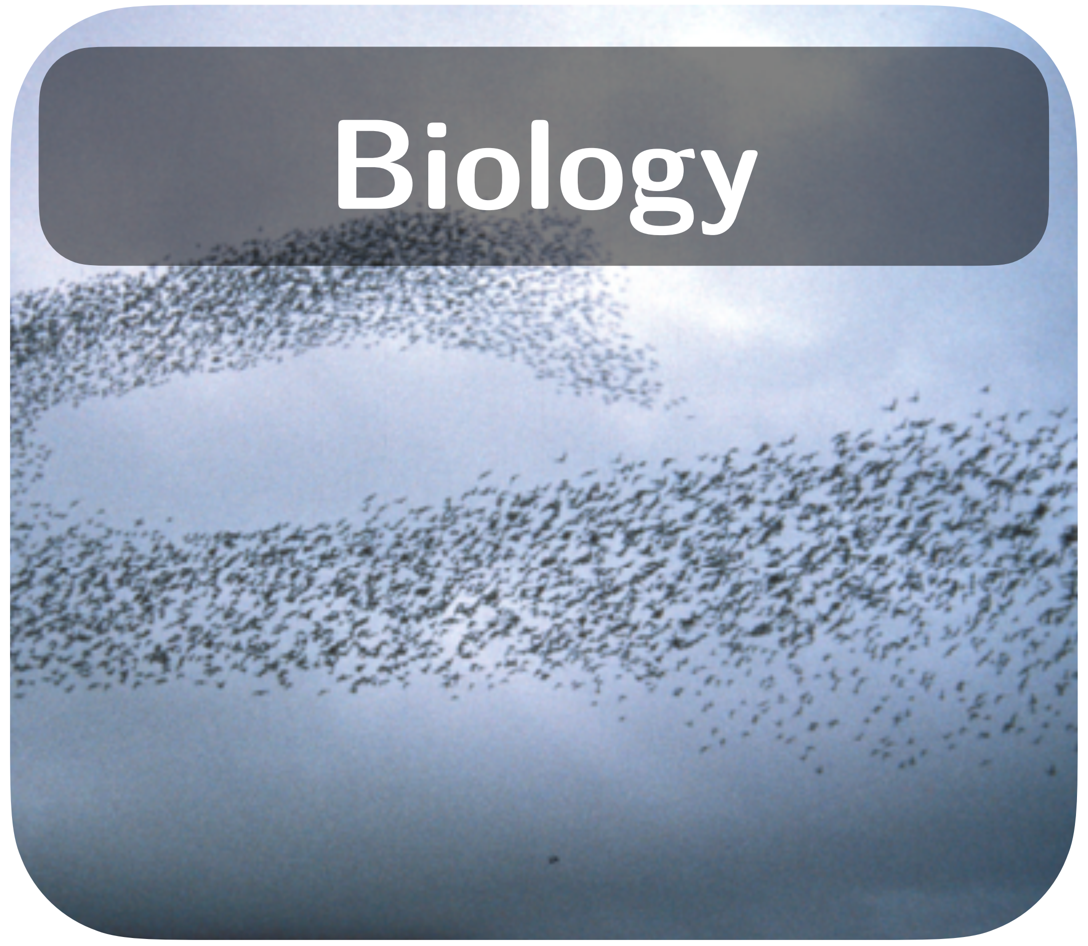
Wolfram Barfuss

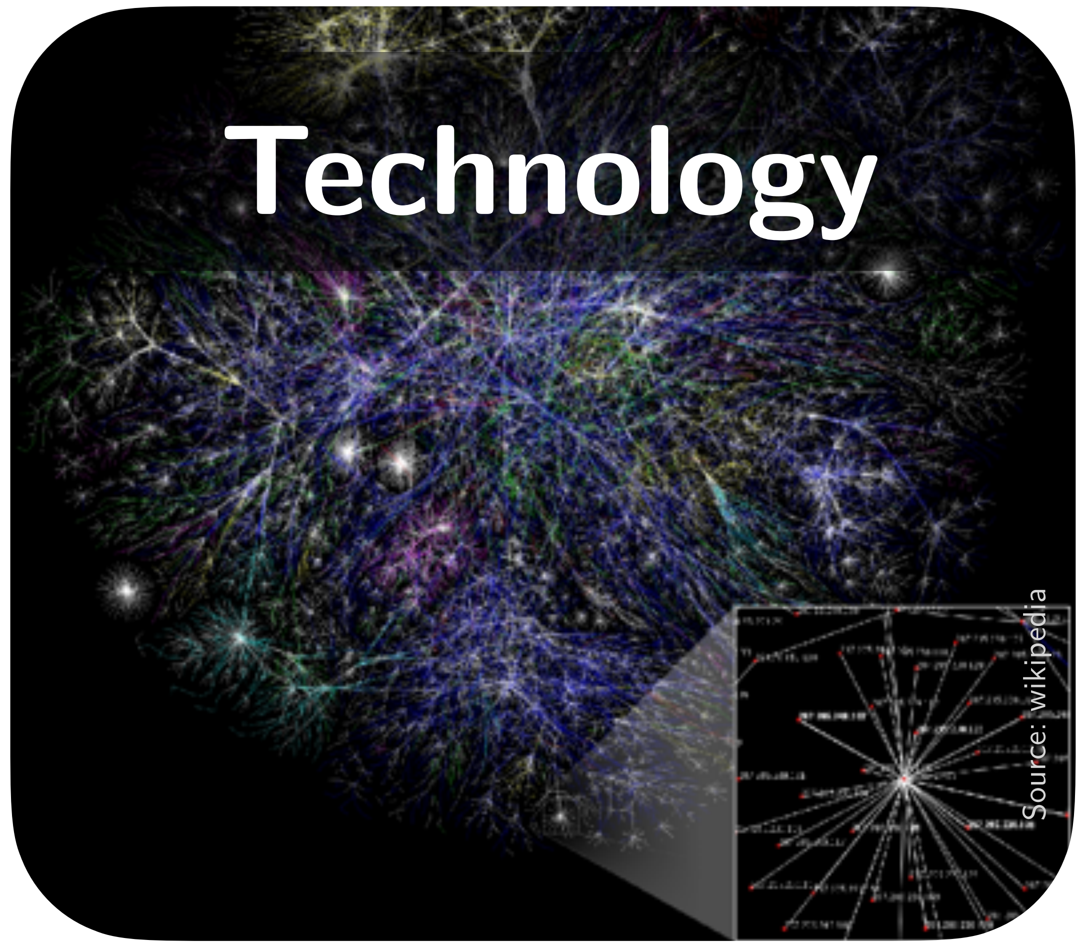
New Directions in Cooperative AI Seminar
May 19, 2022

Challenge

Collective Cooperation

- many individuals acting in the common interest
… even if their incentives are not aligned





What makes intelligent behavior cooperative?

How can we overcome limitations to cooperation?

How can AI help?

Levin (2020) *Collective Cooperation: From ecological communities to global governance and back*

Collective Cooperative Intelligence

Building Bridges between Complex Systems and Multiagent Machine Learning

Bridging communities is **important**

Bridging communities is **neglected**

Bridging communities is **tractable**

Important

Social dilemmas in classic game theory

Mutual cooperation
social optimal strategy ↘

Agent 2

Agent 1

	Cooperate	Defect
C	1, 1	-2, 3
D	3, -2	0, 0

⬅ Defection dominant strategy

Prisoner's Dilemma
Public Good Games

➡️ Outside authorities required

➡️ Cooperation can be stable in repeated interactions

⚡ Challenges for the classic game-theoretic approach

- Hyperrational agents
- Equilibrium selection problem
- Lack of a dynamic theory

https://plato.stanford.edu/entries/game-evolutionary/

Artificial Intelligence

Machine learning

🔨 focus on a single intelligent individual situated in complex environments

🎯 improve on multi-agent cooperation

arXiv > cs > arXiv:2012.08630

Search...

Help | Advanced

Computer Science > Artificial Intelligence

[Submitted on 15 Dec 2020]

Open Problems in Cooperative AI

Allan Dafoe, Edward Hughes, Yoram Bachrach, Tantum Collins, Kevin R. McKee, Joel Z. Leibo, Kate Larson, Thore Graepel

nature

Explore content

nature > comment > art

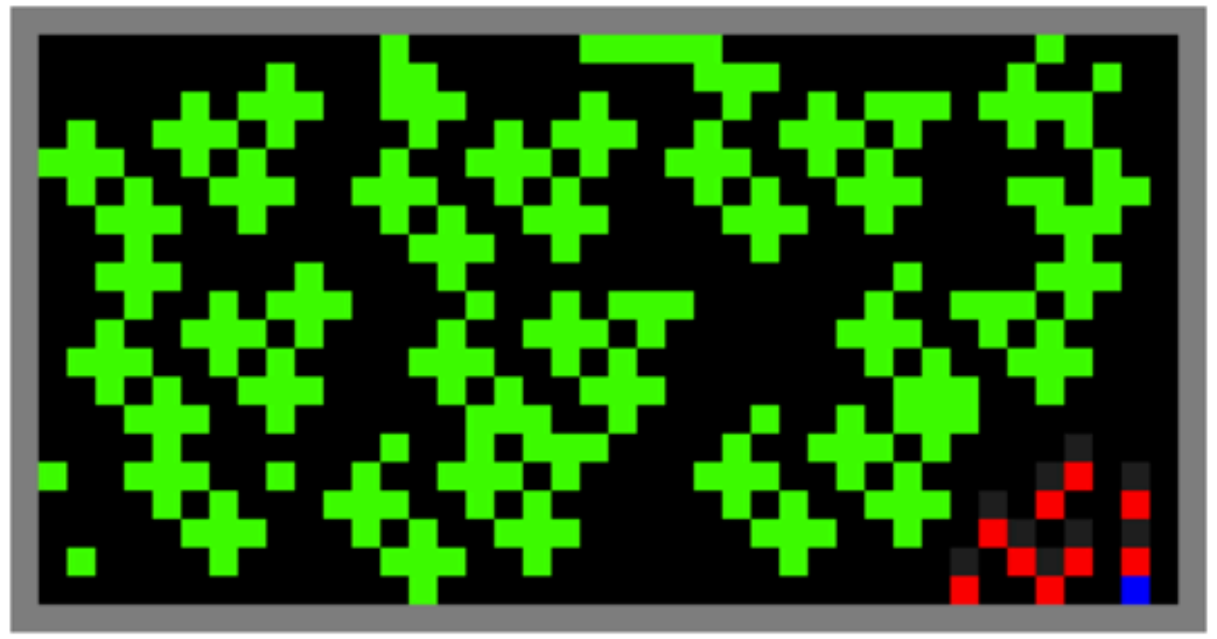
COMMENT | 04 May 2021

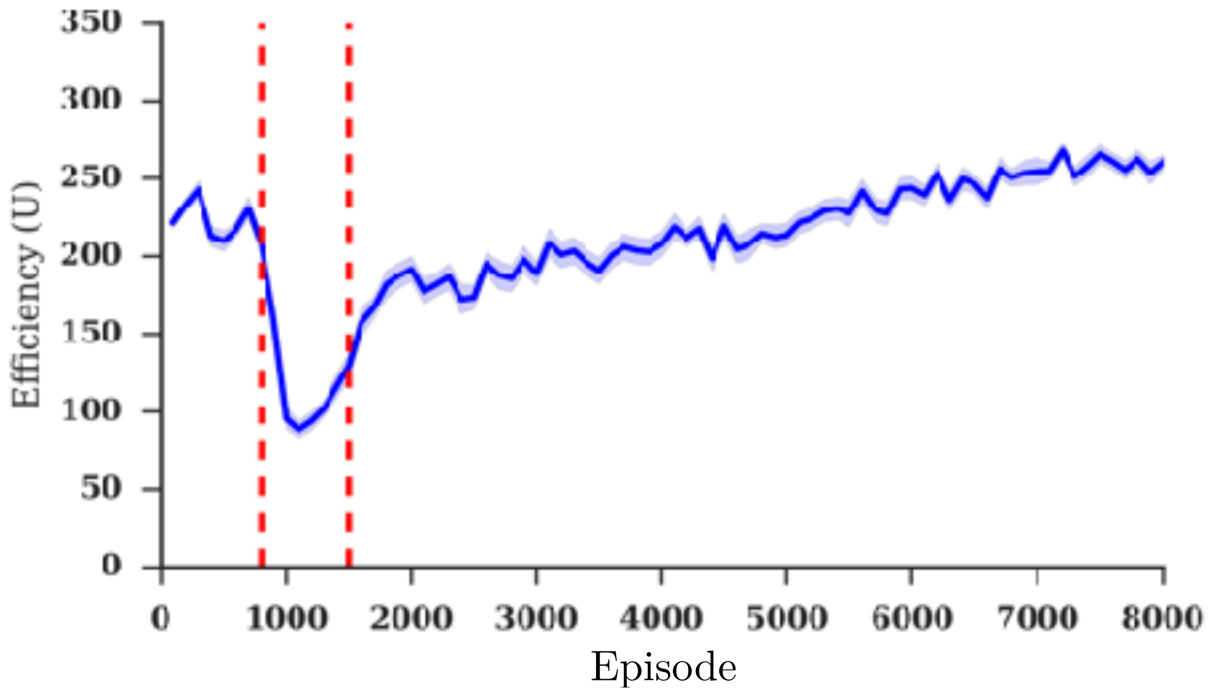
Cooperative AI: machines must learn to find common ground

To help humanity solve fundamental problems of cooperation, scientists need to reconceive artificial intelligence as deeply social.

Allan Dafoe, Yoram Bachrach, Gillian Hadfield, Eric Horvitz, Kate Larson & Thore Graepel

➡️ use deep multi-agent reinforcement learning when standard methods of non-cooperative game theory can no longer be used because of environmental complexity





⚡Challenge for the AI/ML approach

- Limited understanding & explainability
- Resource intensive

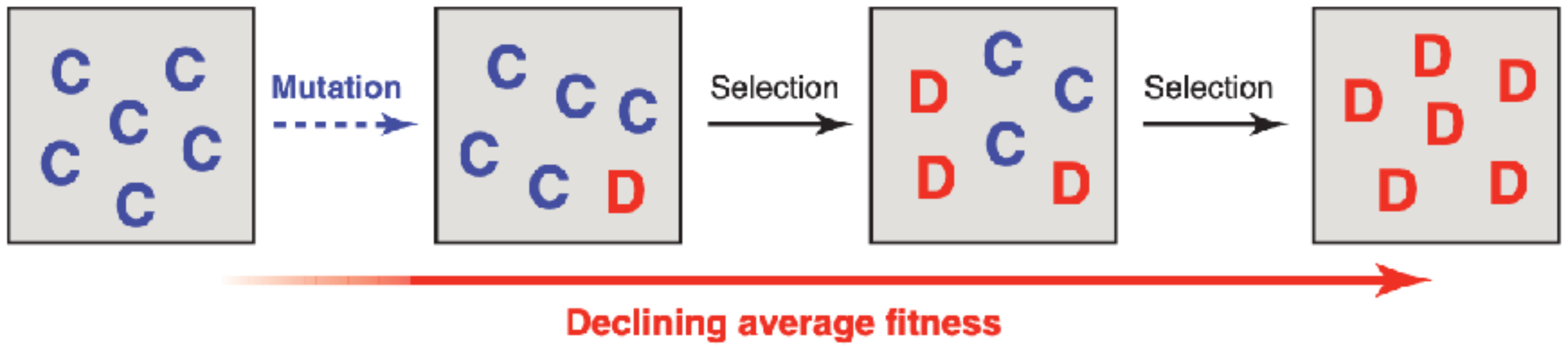
Perolat et al. (2017) *A multi-agent reinforcement learning model of common-pool resource appropriation*

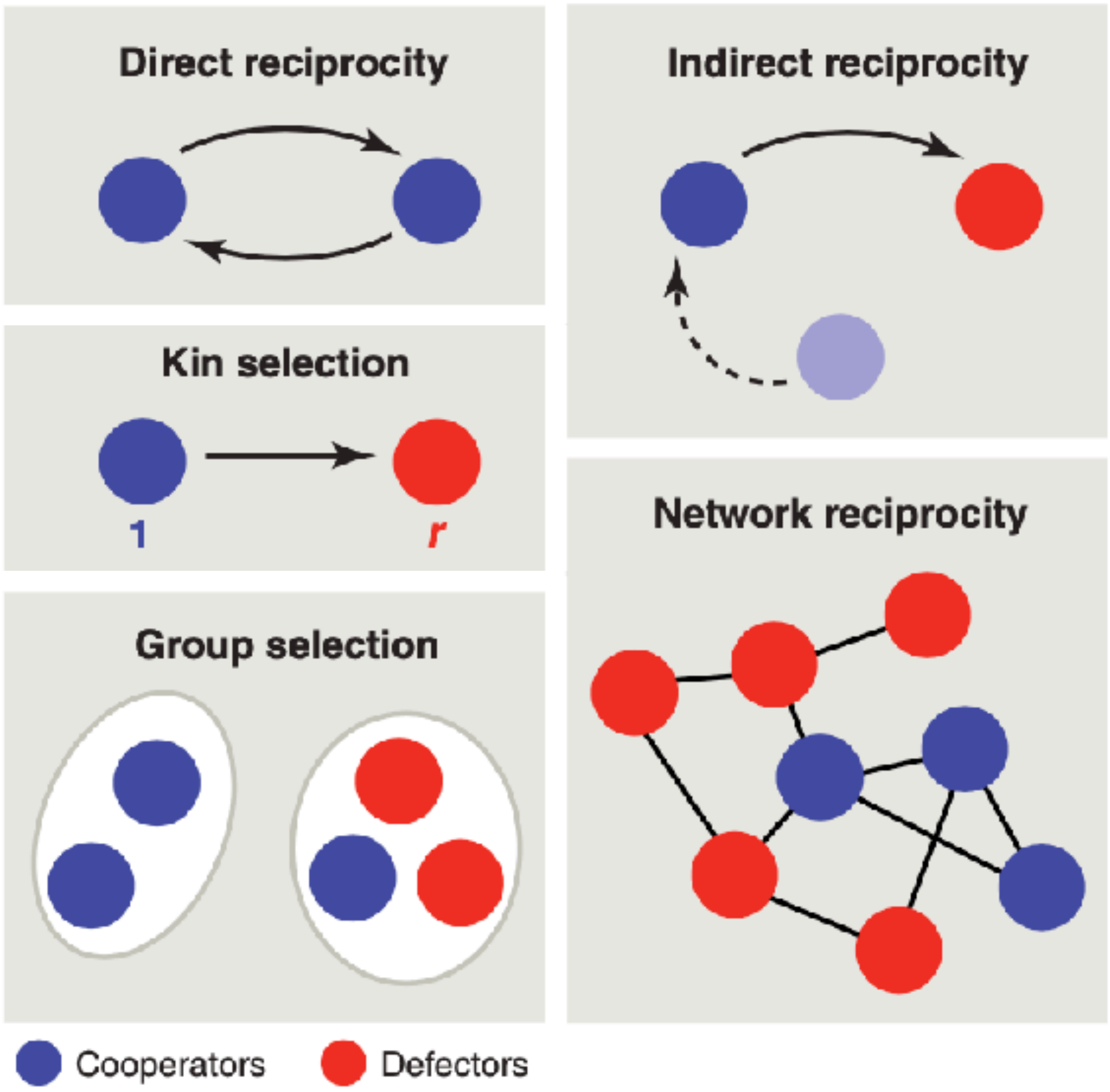
Complex Systems Theories

Evolutionary game theory and complex social networks

🔨 explains the evolution of collective cooperation from simple individuals & interactions

🎯 understand the emergence of cooperation





⚡Challenges for complex systems approaches

- Individual rationality is extremely bounded
- Can say little about individual cognition
- Traditionally focuses on static, non-changing physical environments

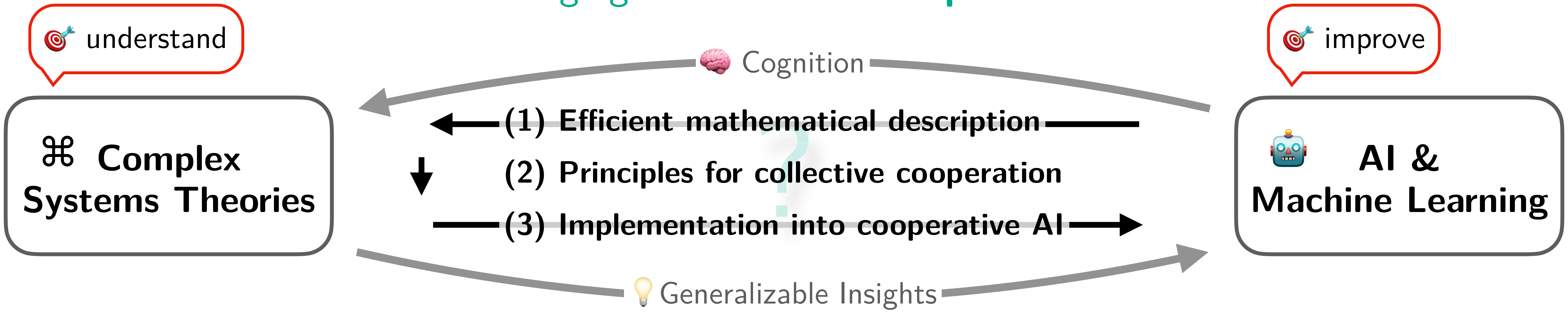
Nowak (2006) *Five rules of the evolution of cooperation*

Han (2022) *Understanding Emergent Behaviours in Multi-Agent Systems with Evolutionary Game Theory*

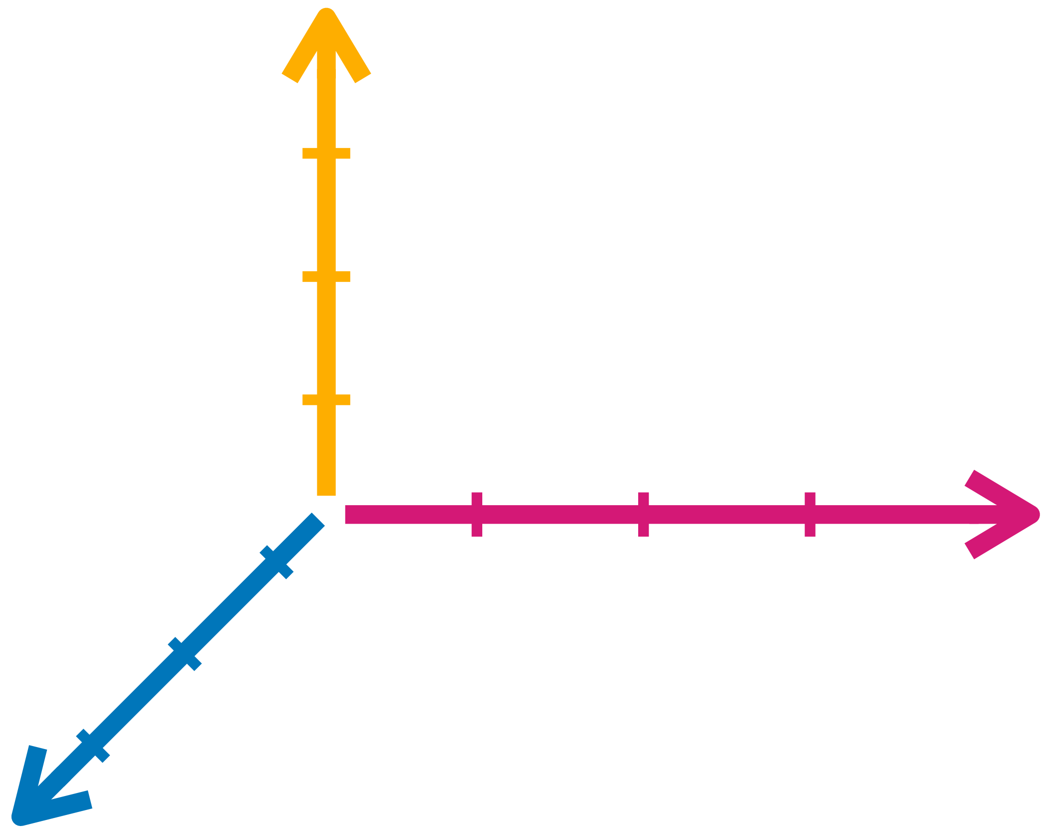
Collective Cooperative Intelligence

Building Bridges between Complex Systems and Multiagent Machine Learning

Bridging communities is **important**

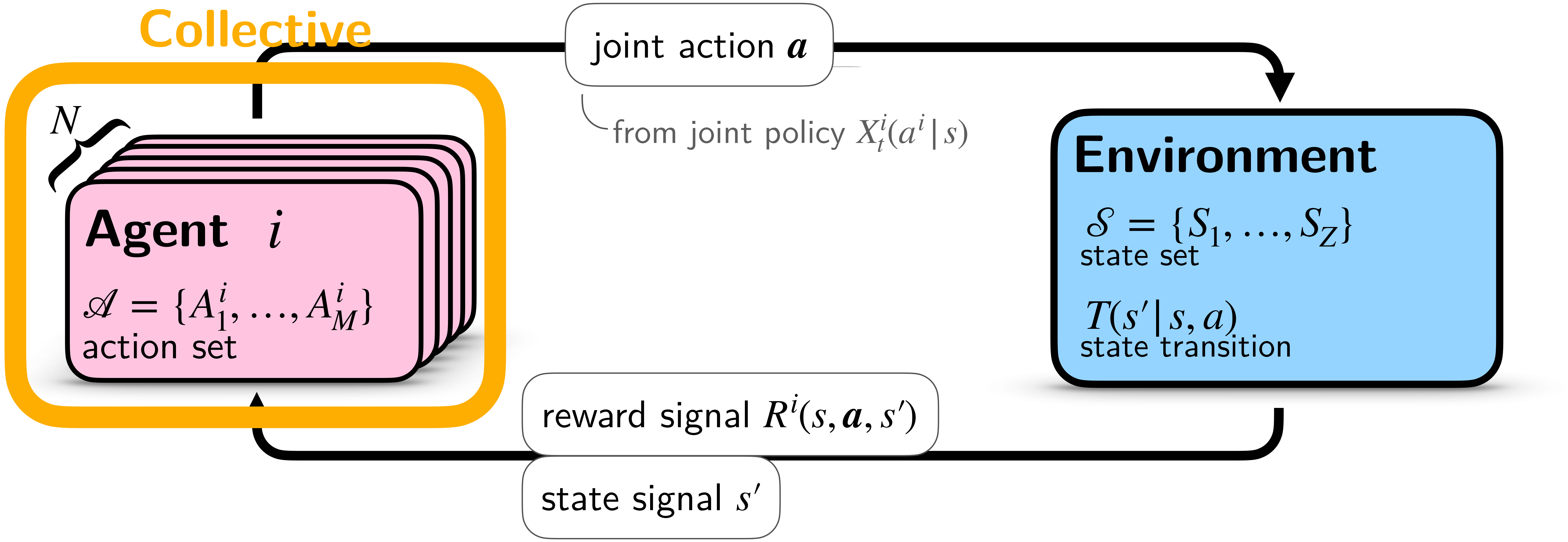


Bridging communities is **neglected**



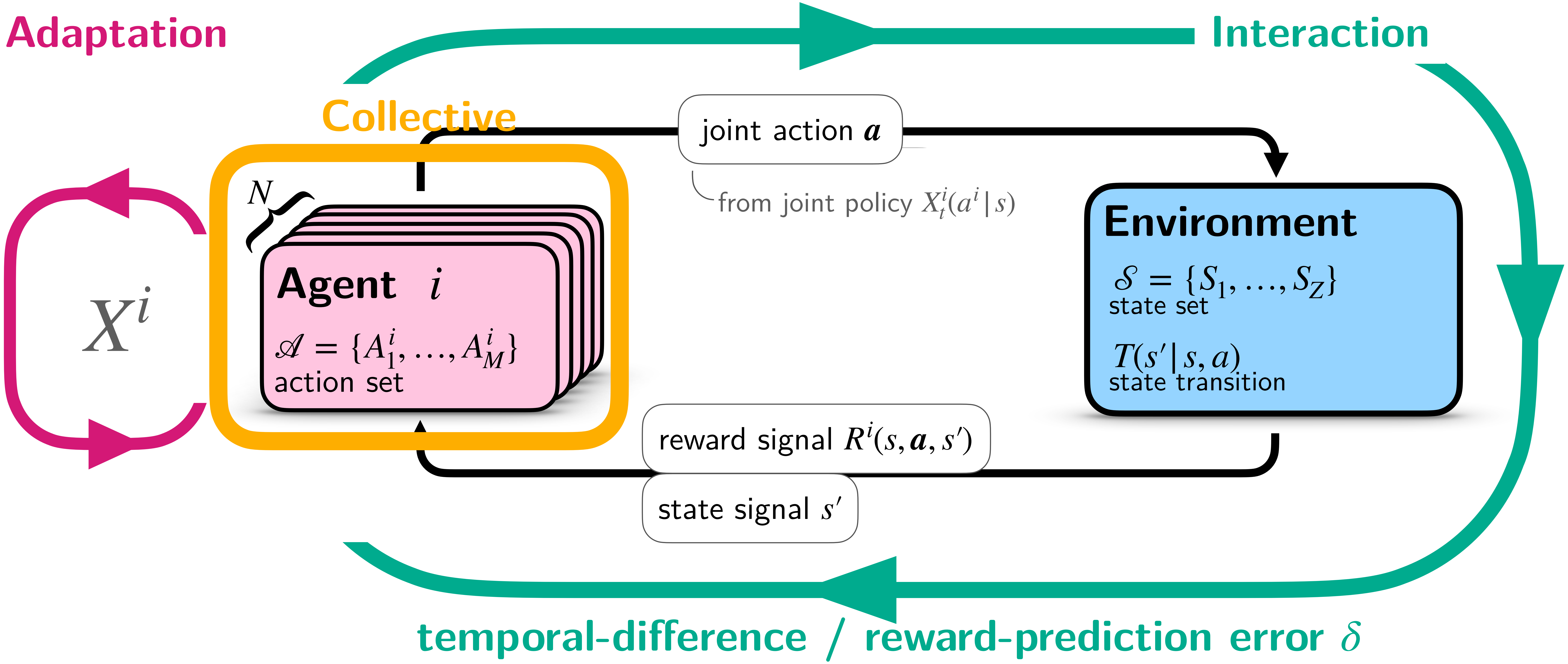
Neglected

Multiagent-Environment Interface



Temporal-Difference Reinforcement Learning

as a general principle to link behavior to the environment

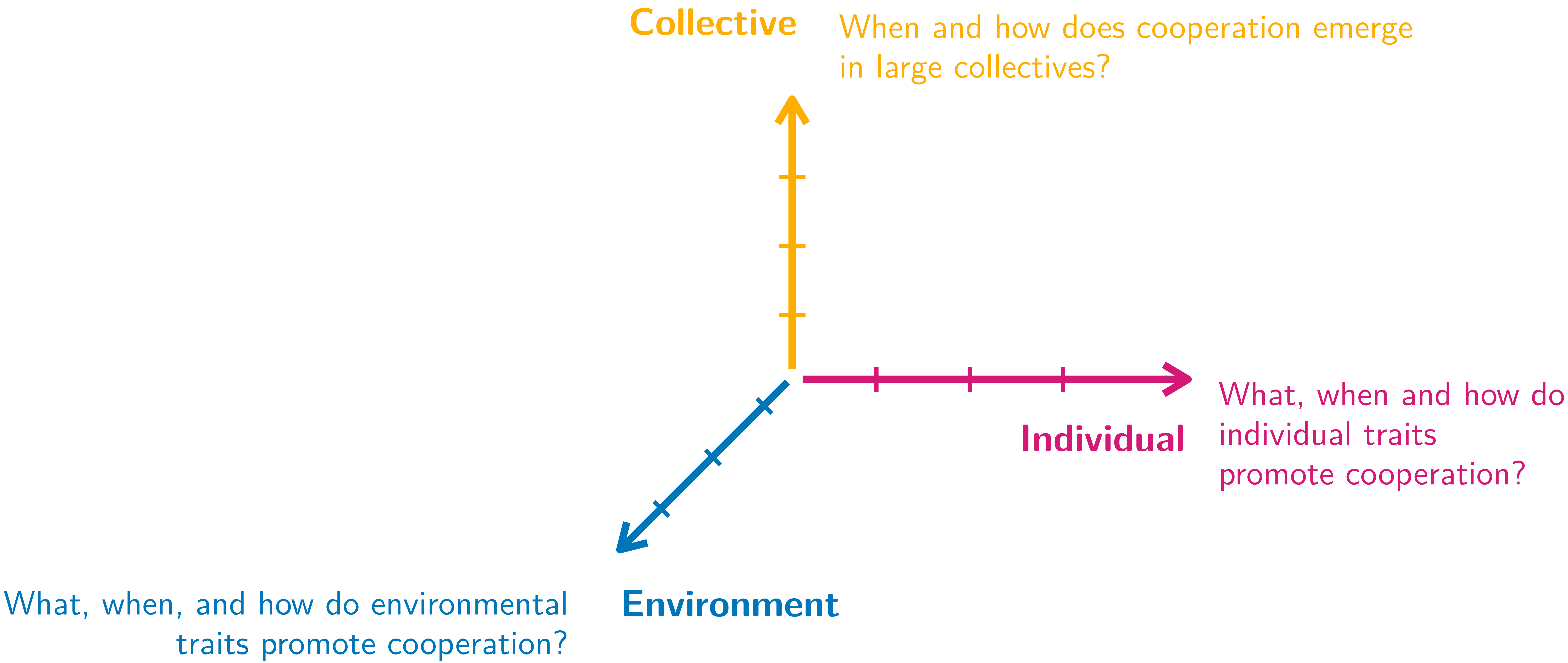


tells agent how to adapt

Sutton & Barto (2018) Reinforcement Learning

Schultz, Dayan & Montaque (1997) A Neural Substrate of Prediction and Reward

Multiagent-environment systems
from a complex systems perspective

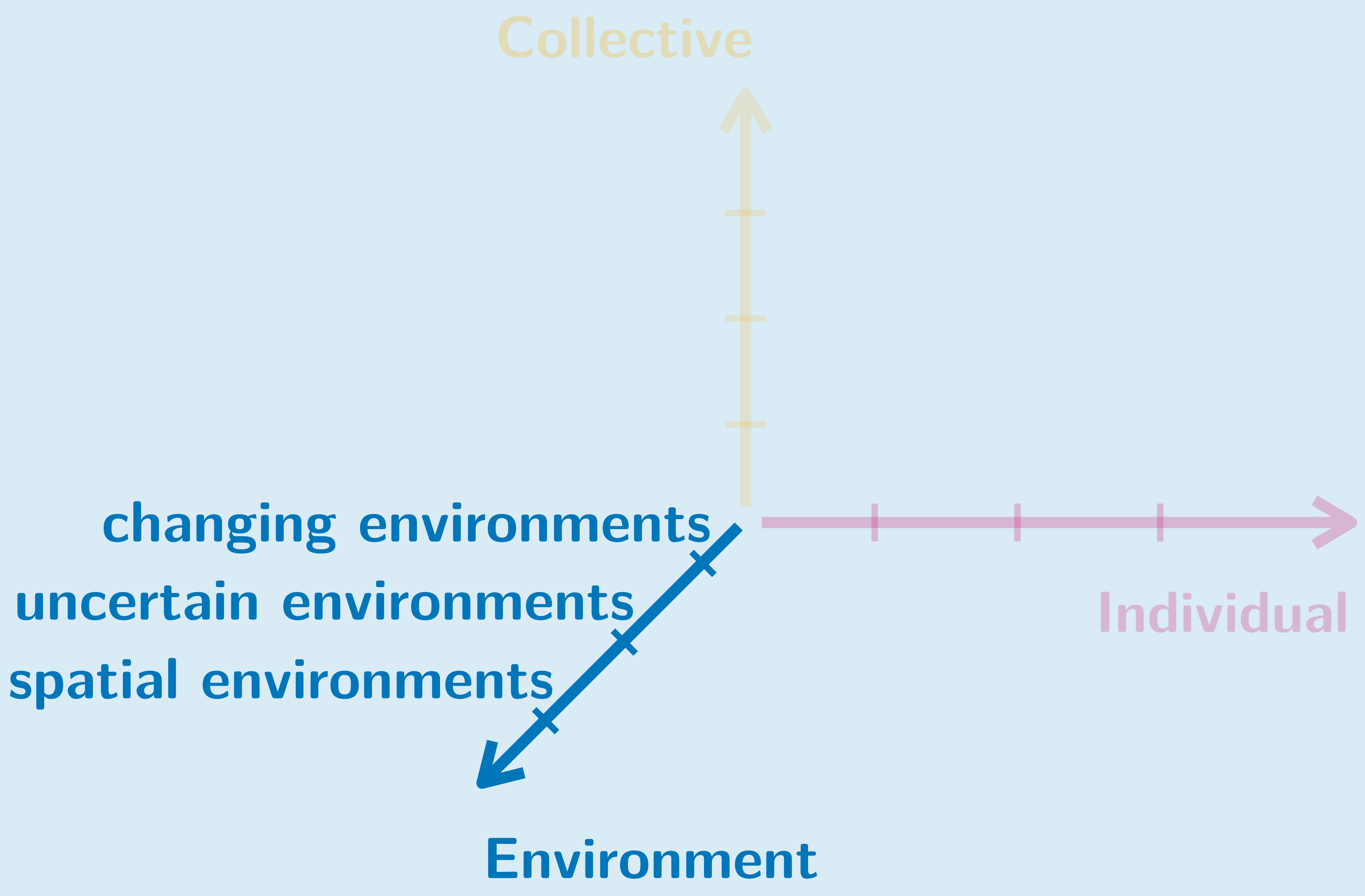


Environment

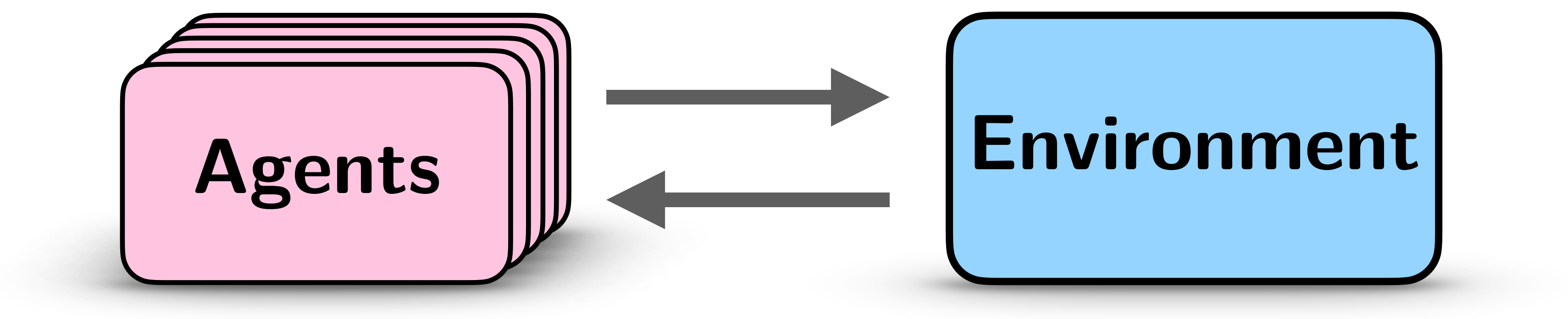
What, when, and how do environmental traits promote cooperation?

- ML ➡️ large, changing & uncertain environments
- CS ➡️ social environment, less on a dynamic physical environment

- ❓ How to bring environments to a complex-systems treatment and provide insights for cooperative AI?



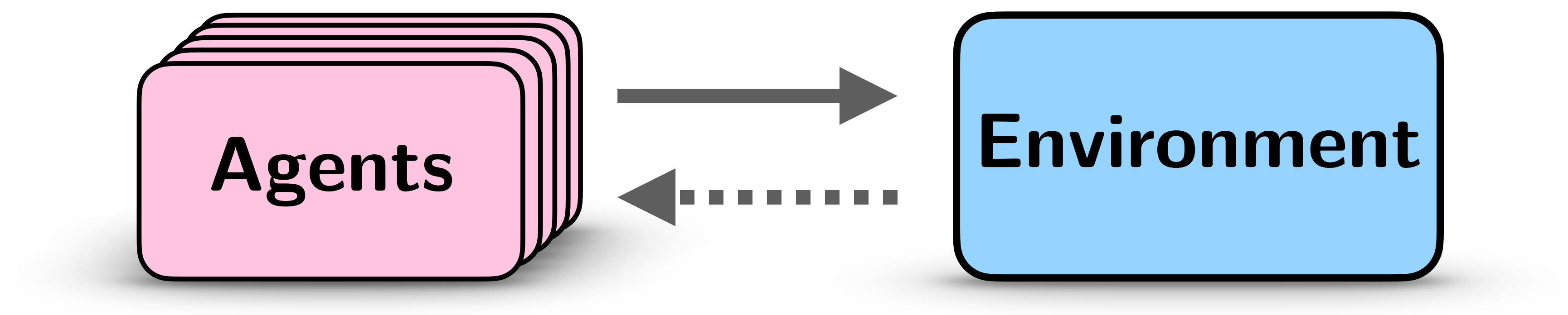
Changing Environments



- Social dilemmas are often embedded in dynamically changing environments
- The multiagent-environment framework integrates an environment into multi-agent learning
- How can we mathematically capture the environmental embedding in multi-agent learning efficiently?
- When and how do changing environments promote or hinder collective cooperation?

Leibo et al. (2017) *Multi-agent Reinforcement Learning in Sequential Social Dilemmas*

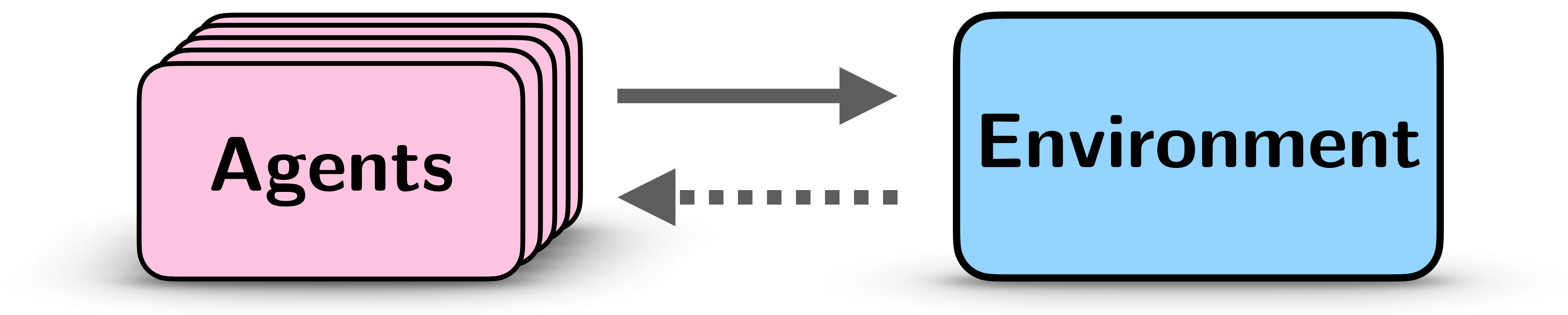
Uncertain Environments

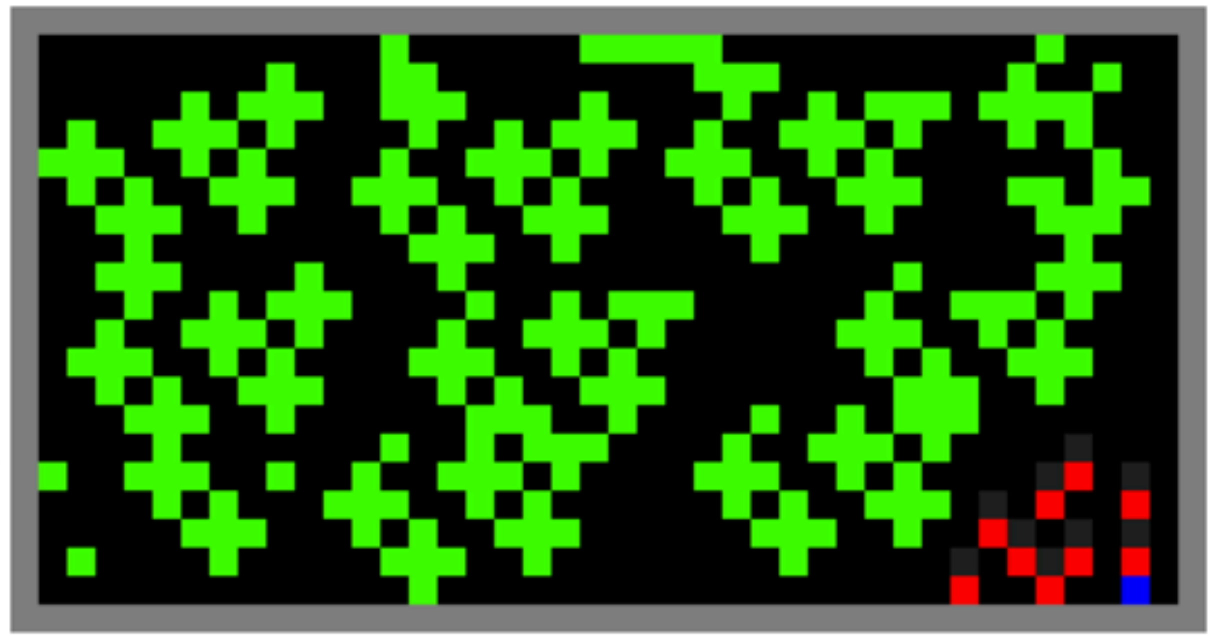


- Observations of the environmental states are often noisy and incomplete

- How can we mathematically capture environmental uncertainty efficiently?
- When and how do uncertain environments promote or hinder collective cooperation?

Spatial Environments





- Social dilemmas are often spatially embedded, leading to an explosion of the state space.

- How can we mathematically capture spatial embeddings efficiently?
- When and how does spatial structure promote or hinder collective cooperation?

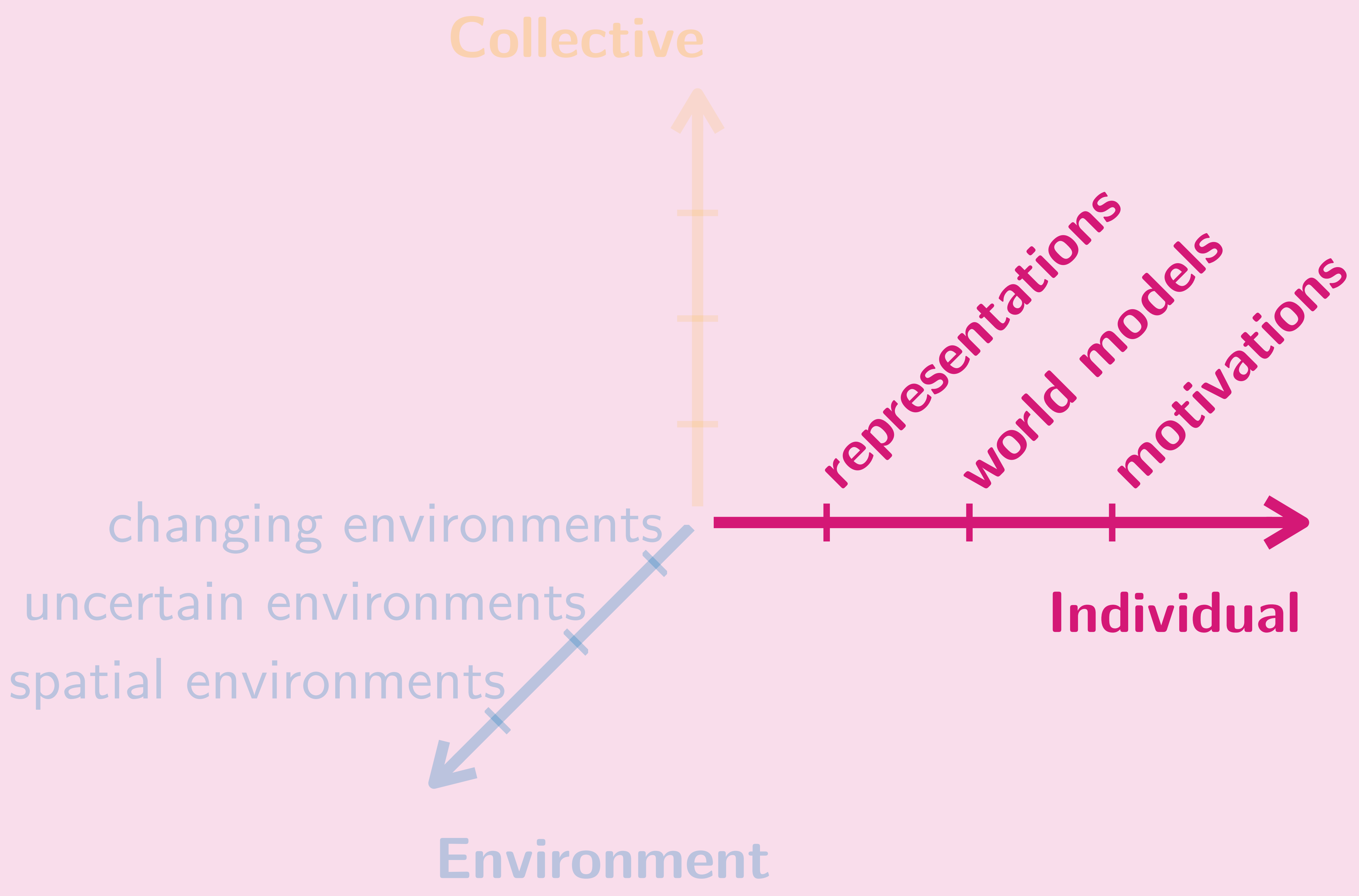
Perolat et al. (2017) *A multi-agent reinforcement learning model of common-pool resource appropriation*

Individuals

What, when, and how do individual traits promote cooperation?

- ML ➡️ increasingly sophisticated individuals
- CS ➡️ deliberately simple individuals

- ❓ How to bring **individual cognition** to a complex-systems treatment and provide insights for cooperative AI?



Intrinsic Representations

- Agents may have different representations of how the world **is**
- Learning efficient representations is critical for intelligent behavior

- How to mathematically capture individual representation learning efficiently?
- When and how do intrinsic representations promote or hinder collective cooperation?

Intrinsic Models of the World

- Agents may have different models of how the world **works**
- Learning and using intrinsic world models efficiently is critical for intelligent behavior

- How to mathematically capture intrinsic world models efficiently?
- When and how do intrinsic world models promote or hinder collective cooperation?

Intrinsic Motivations

- Intrinsic motivation guide learning without relying on externally supplied rewards for improved exploration, control, and homeostasis
- Pro-sociality / other-regarding preferences obviously promote cooperation

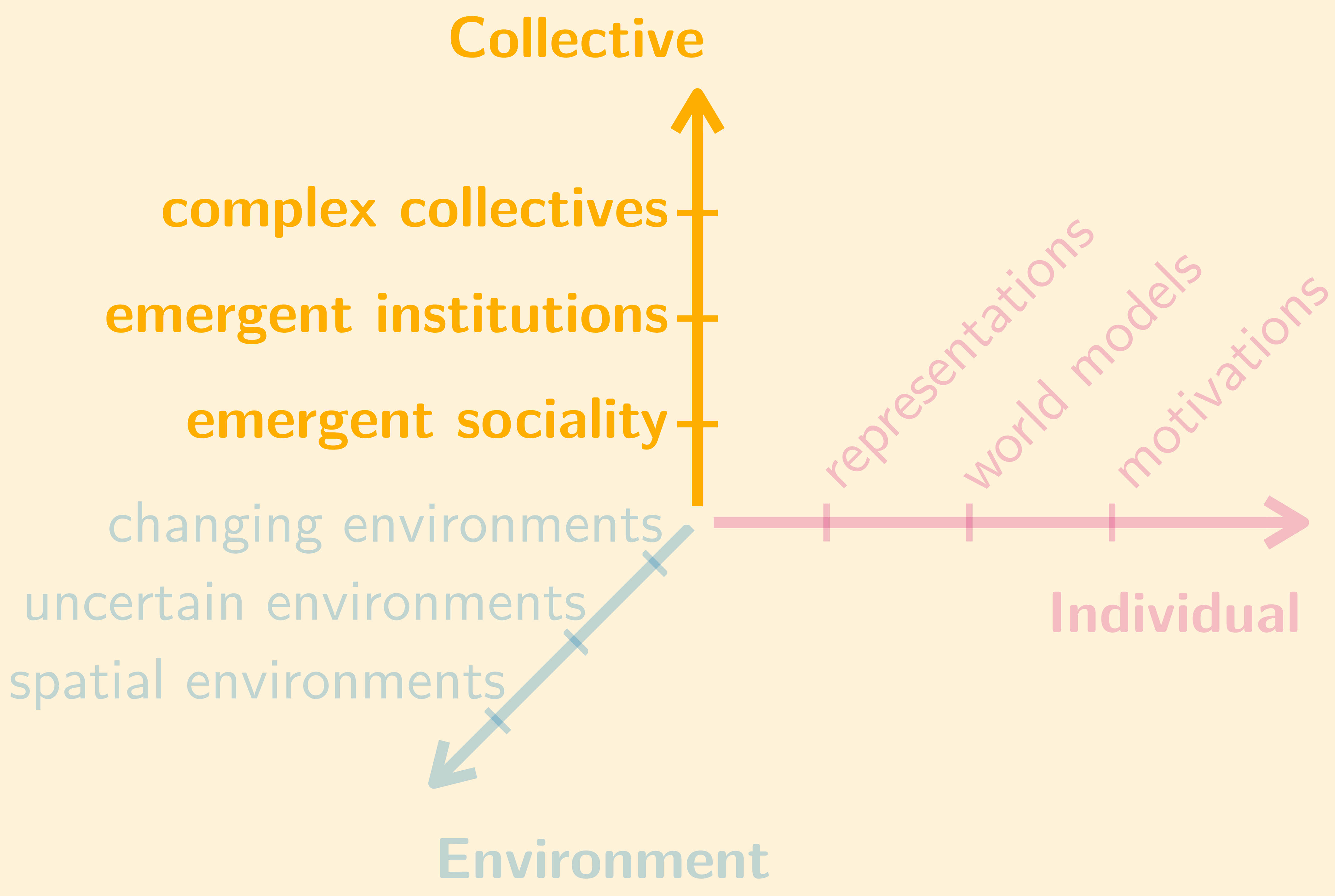
- When and how do other common intrinsic motivation types promote or hinder collective cooperation?
  - Being curious (Novelty/Surprise/Curiosity)
  - Being cautious or a risk-taker (Risk preference / Distributional RL)
  - Being controlling (Empowerment & Agency)
  - Being predictive (Low model prediction error)

Jaques et al. (2019) *Social Influence as Intrinsic Motivation for Multi-Agent Deep Reinforcement Learning*
Hughes et al. (2018) *Inequity aversion improves cooperation in intertemporal social dilemmas*

Collective

When and how does cooperation emerge in large collectives?

- ML ➡️ individual behavior in complex environments
- CS ➡️ emergent collective behavior
- ❓ How to bring the **collective level** to machine learning and provide insights for cooperative AI?



Ha & Tang (2021) *Collective Intelligence for Deep Learning: A Survey of Recent Developments*

Emergent Sociality

- Most biological agents are social - learn through imitation, bond with similar others (homophily)
- Many complex systems models model these traits directly
- Unclear when they are likely to change, break down, and how they respond to changing environments.

- When and how do such social interaction rules emerge, depending on environmental and individual traits?

Emergent Institutions

- Complex systems are often hierarchically structured and interlinked
  neurons $\curvearrowright$ animals $\curvearrowright$ groups (households, firms, states) $\curvearrowright$ societies $\curvearrowright$ norms & value systems
- When and how does a collective obtain higher-level agency?
- How can micro-agents induce higher-level cooperation?

Flack & Krakauer (2011) *Challenges for complexity measures: A perspective from social dynamics and collective social computation*

Complex Collectives

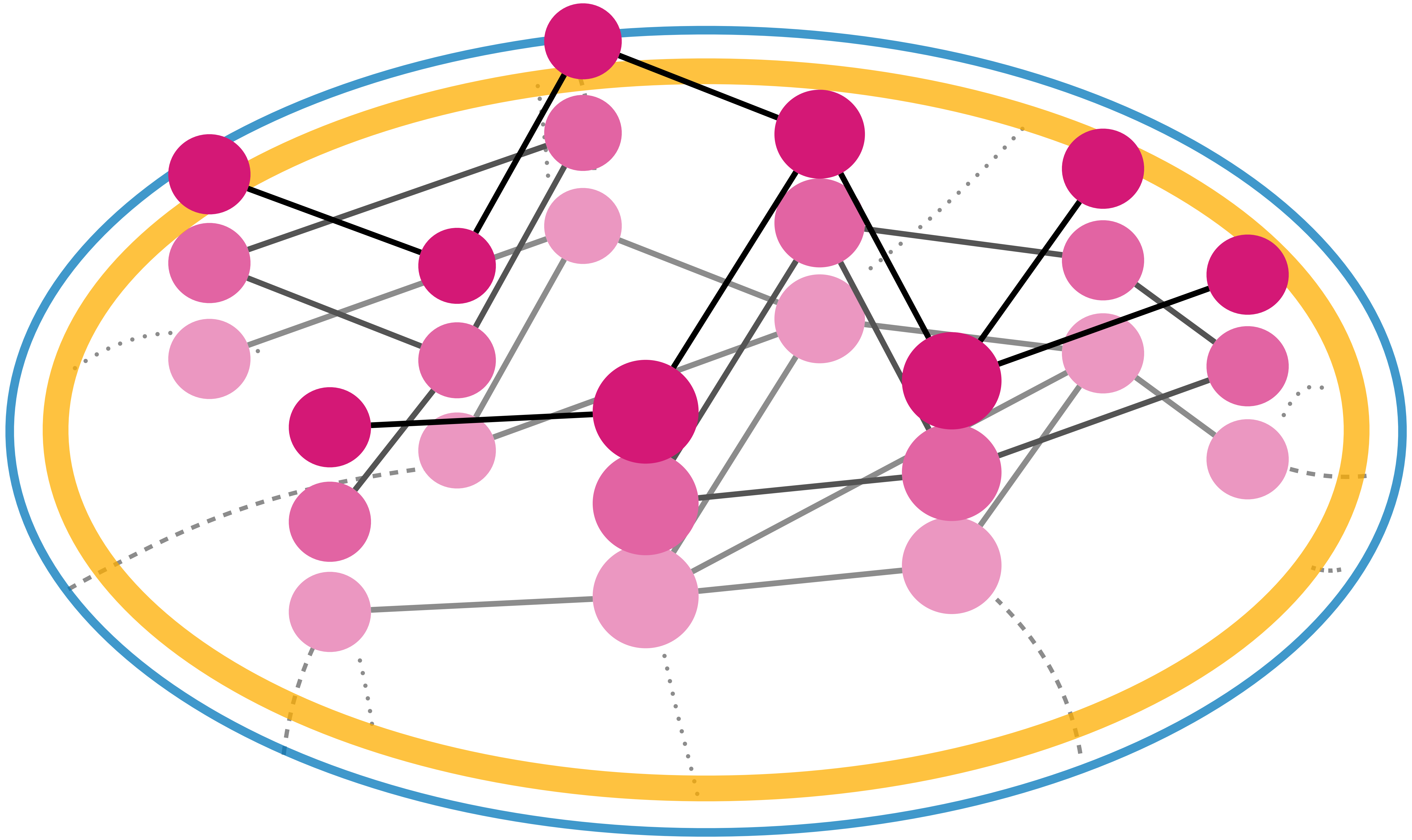
When and how does cooperation emerge in large, diverse & complex collectives?

➡️ Modeling multiagent-environment systems as a complex multi-layer network

Action-Observation Layer

Reward-Observation Layer

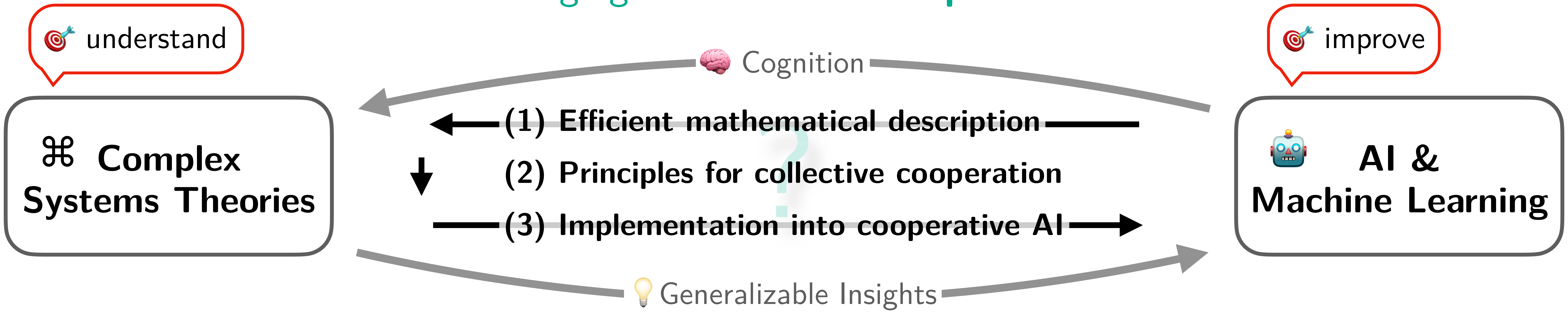
Reward Layer



Collective Cooperative Intelligence

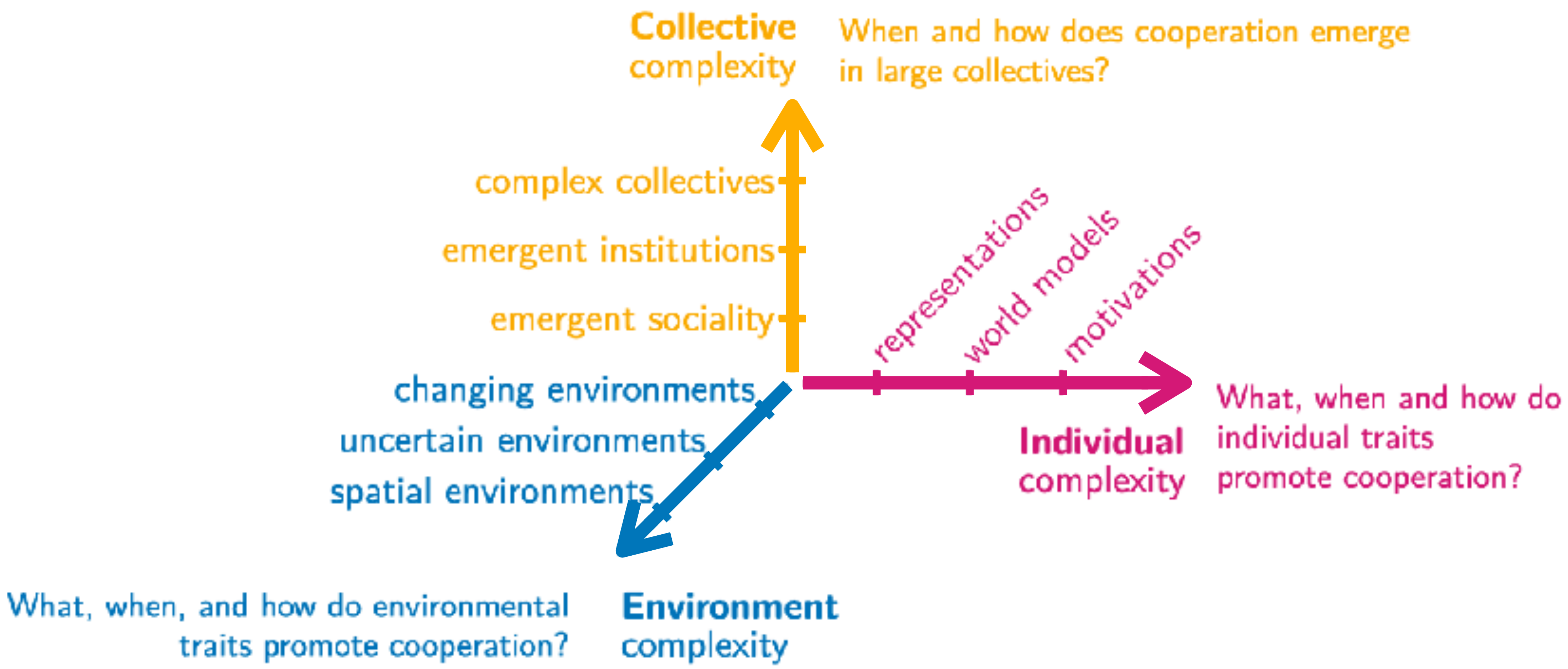
Building Bridges between Complex Systems and Multiagent Machine Learning

Bridging communities is **important**

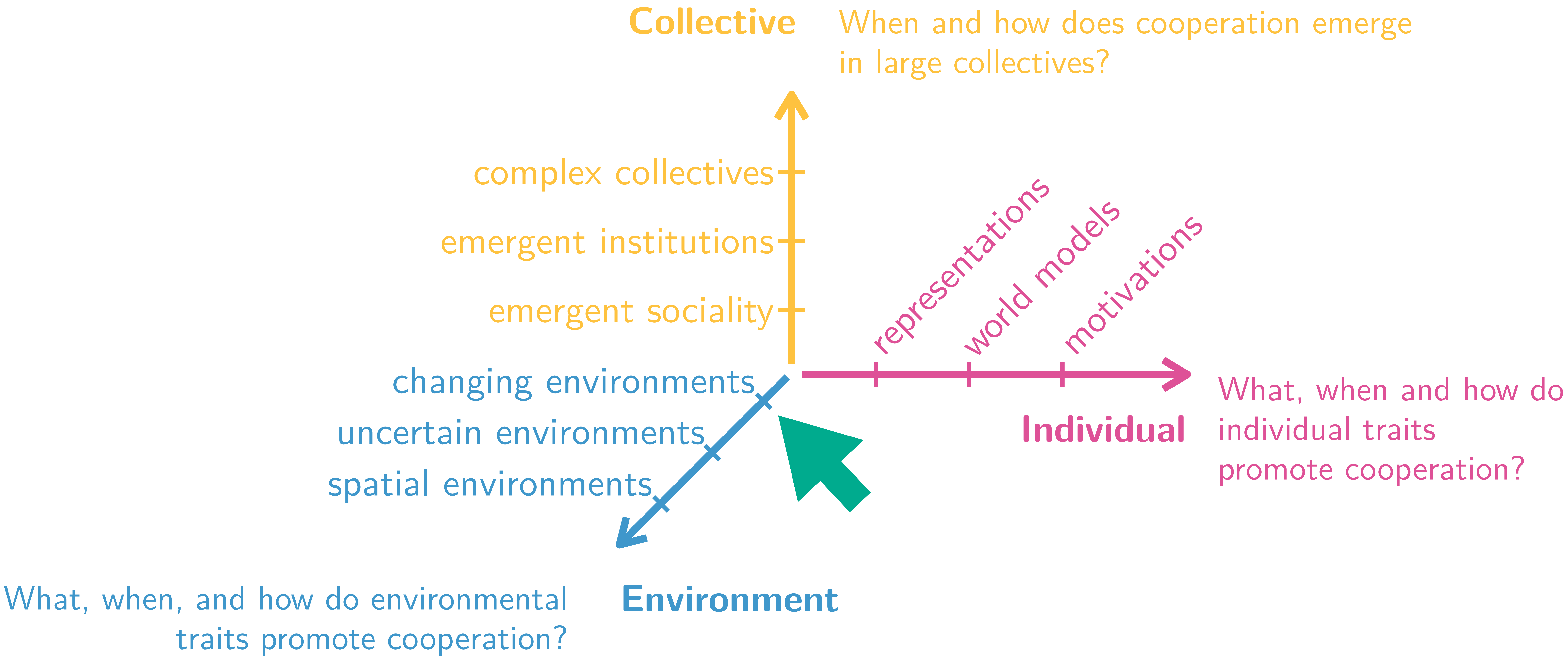


Bridging communities is **neglected**

Bridging communities is **tractable**



Tractable


Collective
When and how does cooperation emerge in large collectives?
complex collectives
emergent institutions
emergent sociality
representations
world models
motivations
What, when and how do individual traits promote cooperation?
Individual
changing environments
uncertain environments
spatial environments
What, when, and how do environmental traits promote cooperation?
Environment

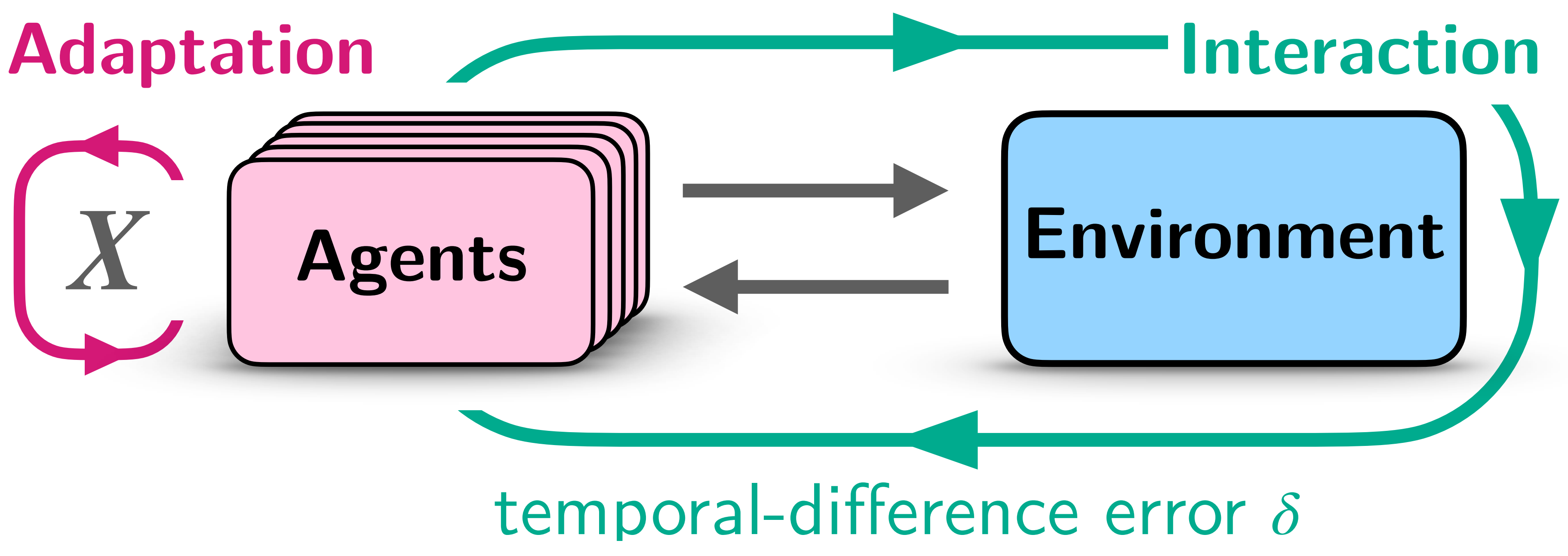
Collective Learning Dynamics

to understand multiagent learning in changing environments as a dynamical system

Existing dynamics focused on stateless games without changing environments
Hofbauer & Sigmund, 1998
Fudenberg & Levine, 1998
Sato & Crutchfield, 2003
Bloembergen et al., 2015

Separating the timescales of **interaction** and **adaptation**

Standard, stochastic, independent, model-free
Reinforcement learning **algorithms**



Deterministic (environmental-aware)
Reinforcement learning **equations**

$$X_{t+1} = \frac{1}{\bar{Z}} X_t \exp(\alpha \bar{\delta})$$

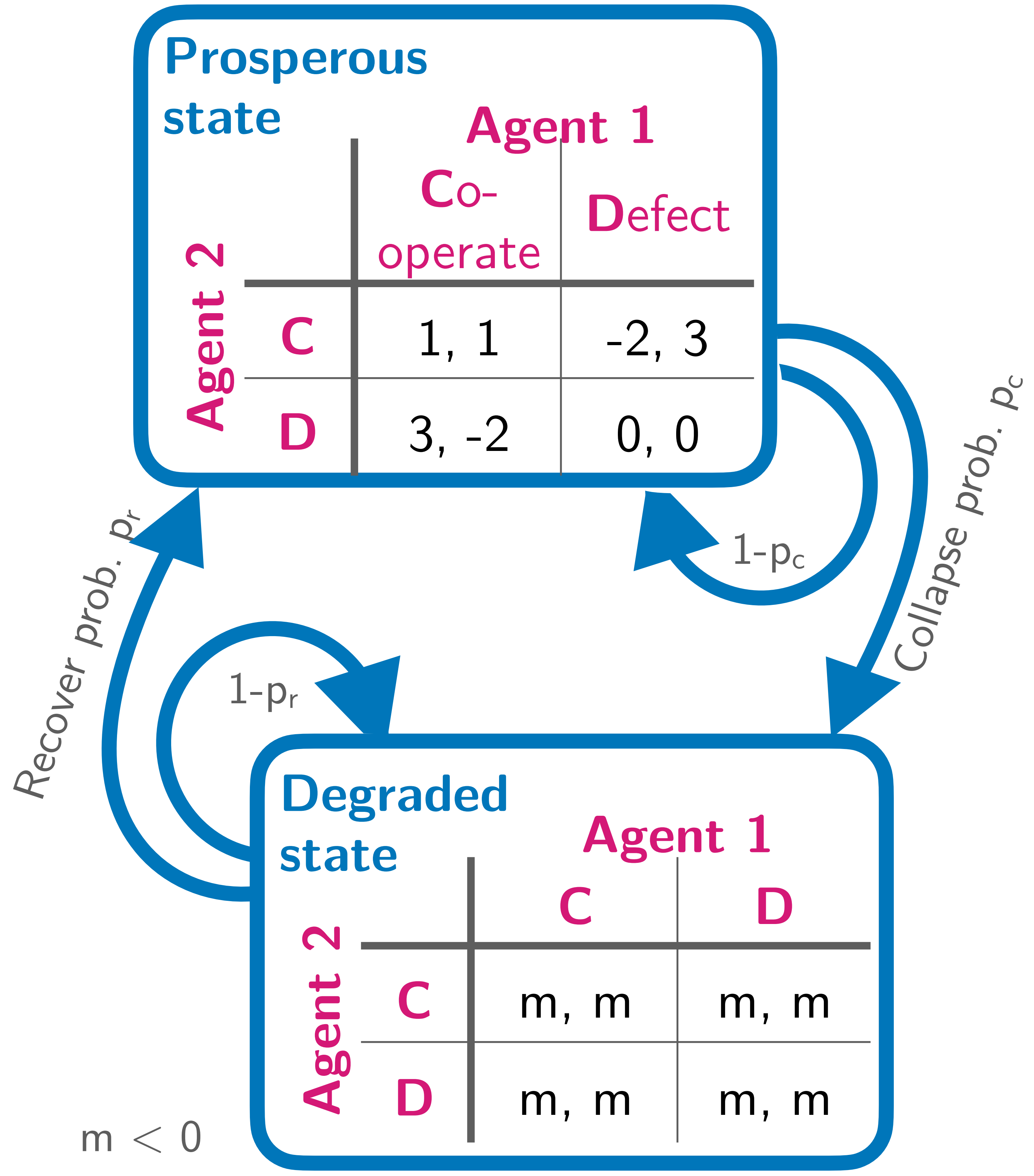
Learning rate (α), Policy-average ($\bar{\delta}$), Normalisation ($\bar{Z}$), TD-error (δ)

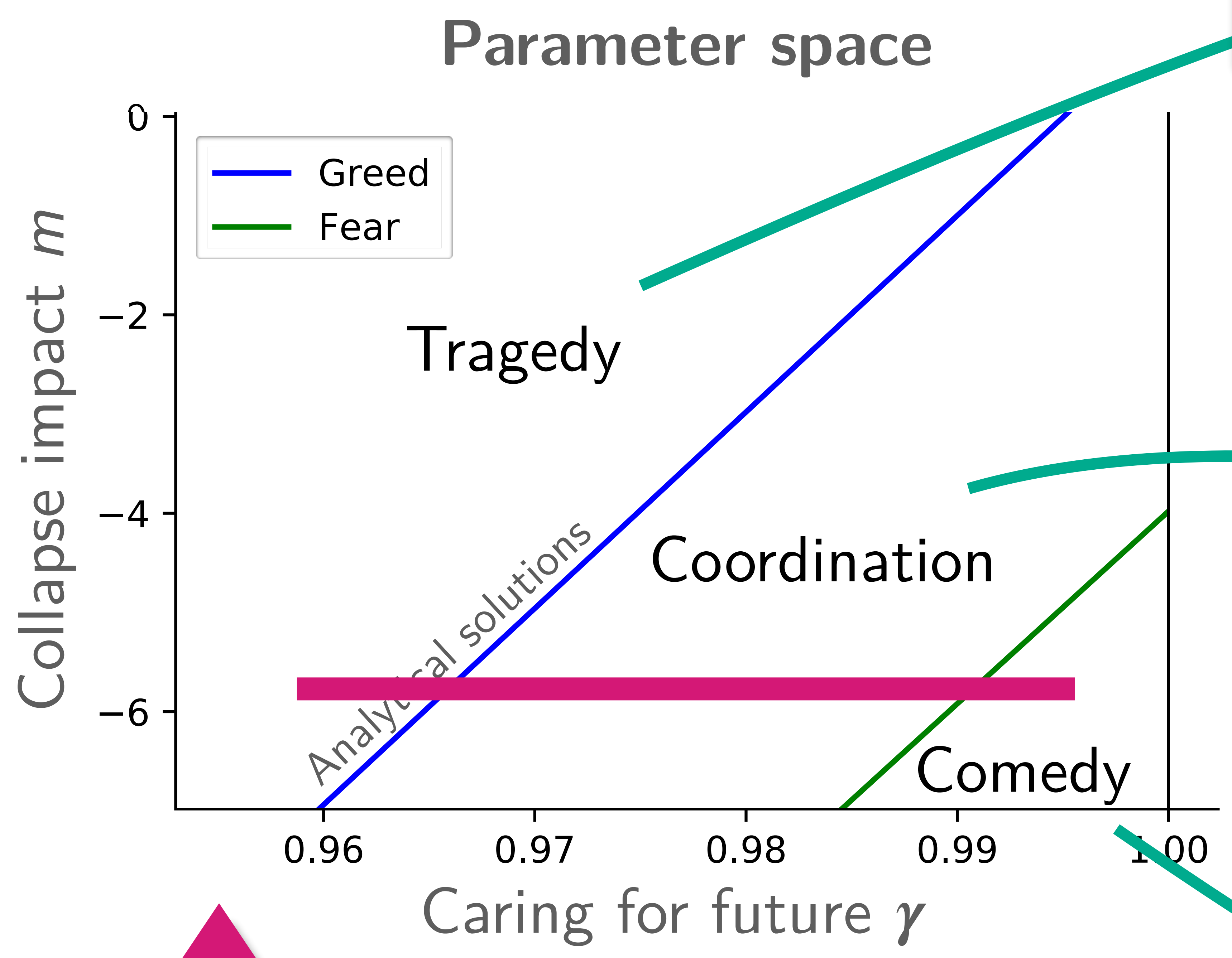
+ Analytical understanding
+ Deterministic reproducible
+ Computationally light
O Connected to scalable algorithms

An **ideal model** of multi-agent reinforcement learning
as if having a perfect intrinsic model of the current environment

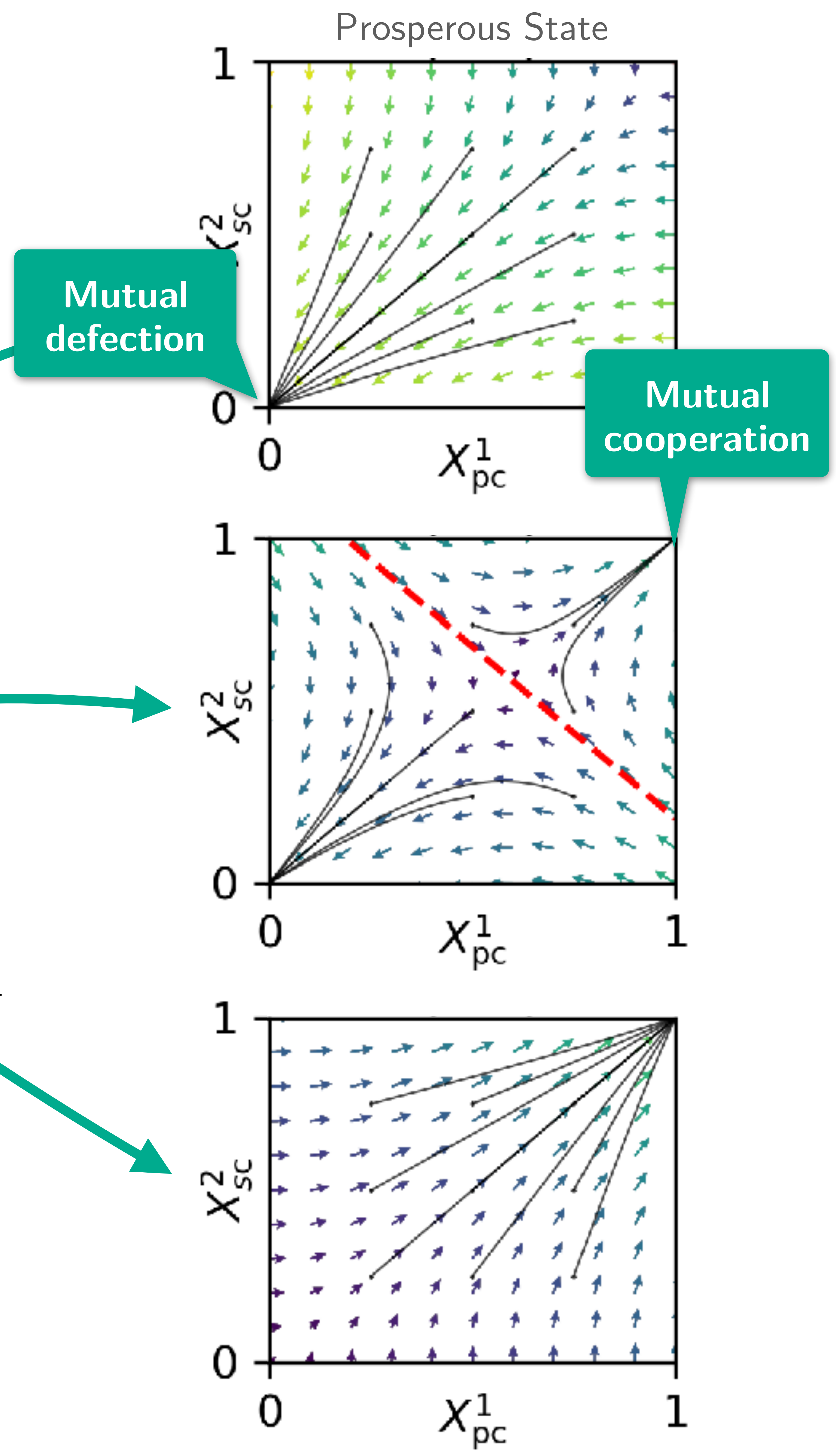
Barfuss et al. (2019) *Deterministic limit of temporal difference reinforcement learning for stochastic games*
Barfuss (2022) *Dynamical systems as a level of cognitive analysis of multi-agent learning*

Caring for the future alone! without social reciprocity
can turn tragedy of the commons into a comedy
in a dynamic environment



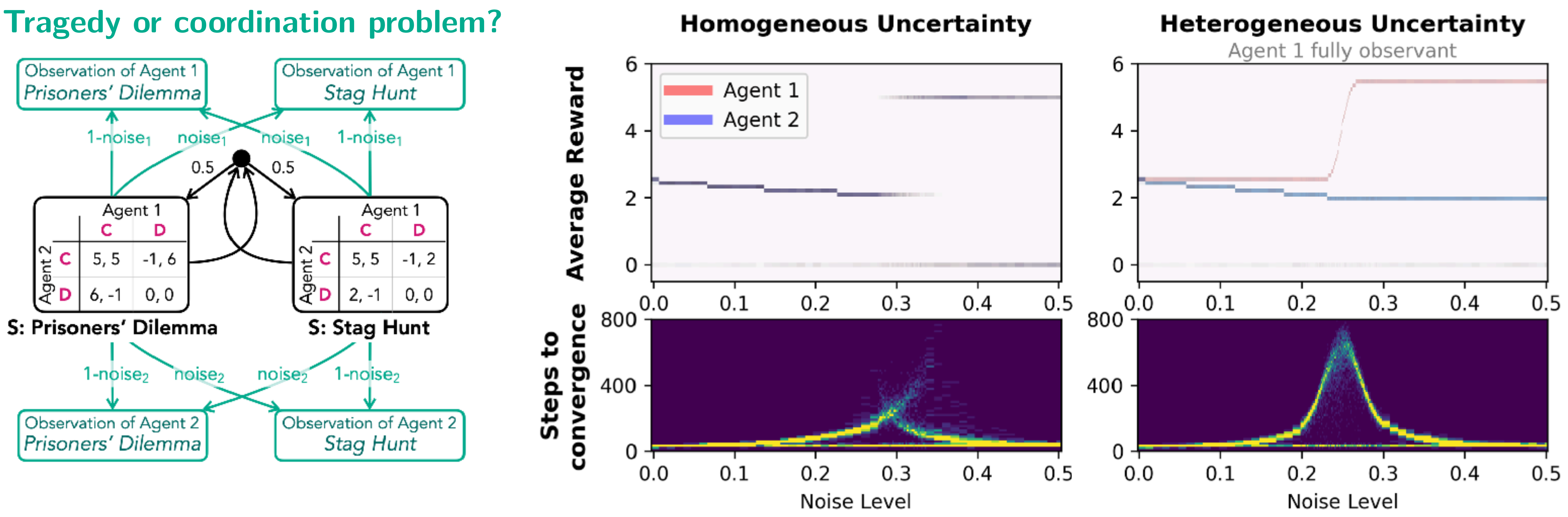


provides generalizable insight for cooperative AI



Uncertain Environments

- When and how does environmental uncertainty promote or hinder collective cooperation?
- ➡️ Partial observability can induce a critical transition towards cooperation



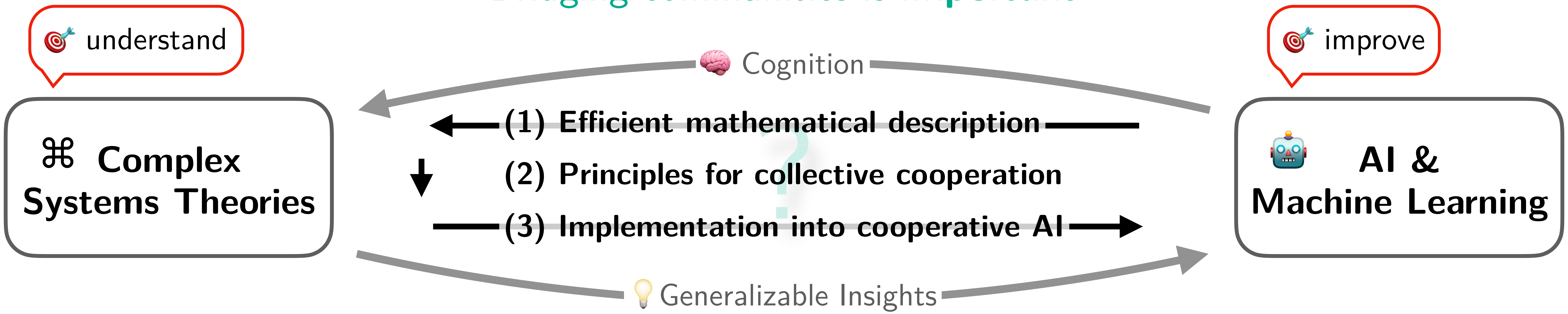
Barfuss & Mann (2022) *Modeling the effects of environmental and perceptual uncertainty using partially observable learning dynamics*

Conclusion

Collective Cooperative Intelligence

Building Bridges between Complex Systems and Multiagent Machine Learning

Bridging communities is **important**



Bridging communities is **neglected**

Thank you

Bridging communities is **tractable**