Cooperative AI

Kate Larson

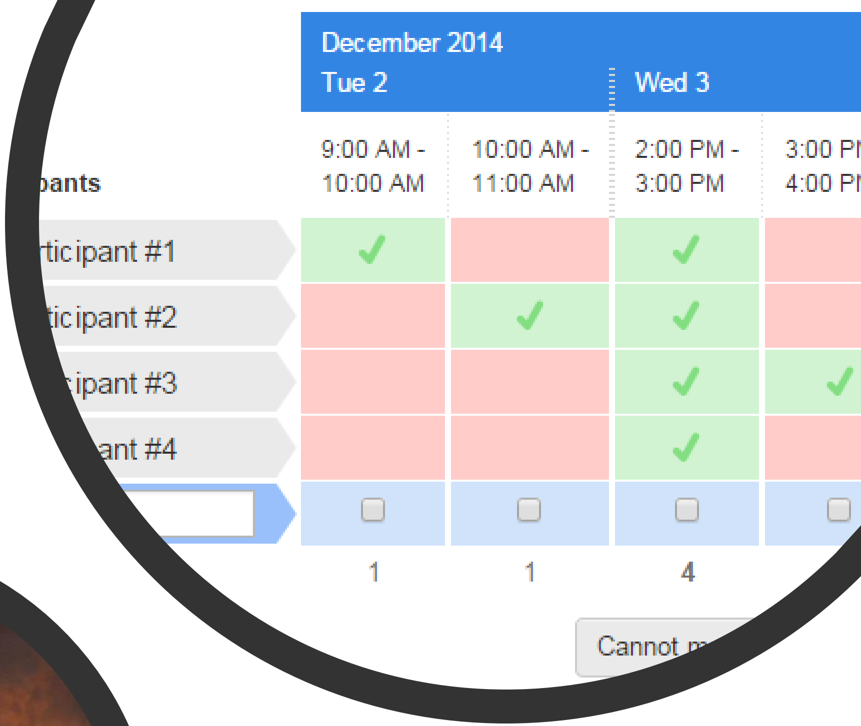
Cheriton School of Computer Science

University of Waterloo

UNIVERSITY OF WATERLOO

Problems of cooperation are ubiquitous and important

These are situations where agents have opportunities to improve their joint welfare but where it is not easy for them to do so.



Cooperation is Key

Arguably, the success of humans is rooted in our ability to cooperate.

Since machines powered by AI are playing an ever-greater role in our lives, it will be important to equip them with the capabilities necessary to **cooperate** and **foster cooperation.**

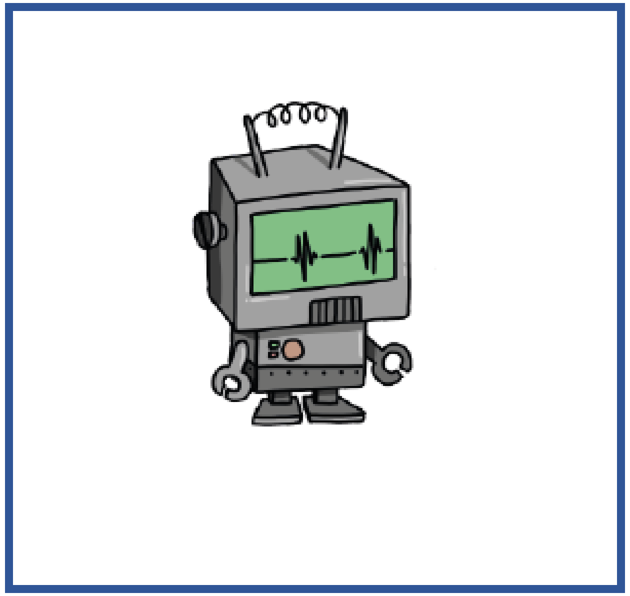
This requires **social understanding** and **cooperative intelligence.**

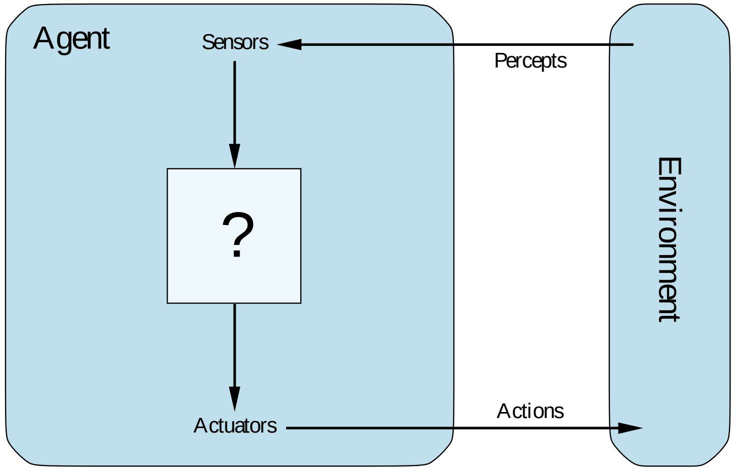
Cooperative AI: machines must learn to find common ground

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

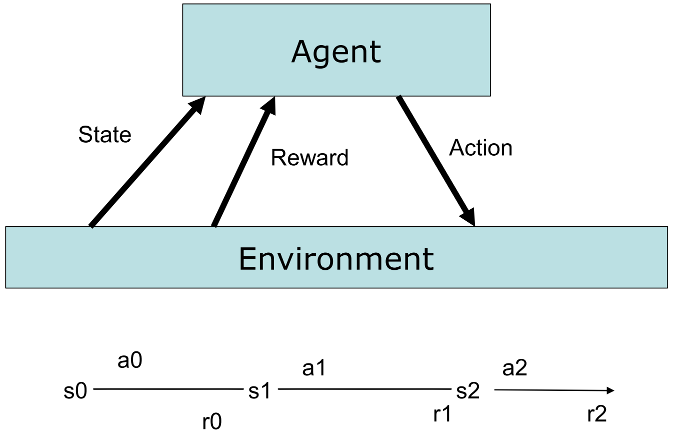
To help humanity solve fundamental problems of cooperation, scientists need to reconceive artificial intelligence as deeply social.

Historically AI has been steeped in “*methodological individualism*”





Russell&Norvig
AIMA



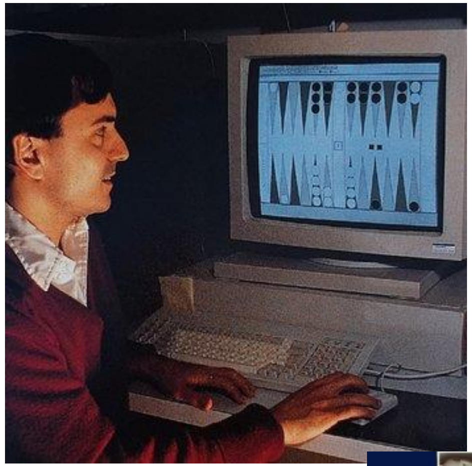
This is a sensible starting point.
An AI agent needs to understand the environment and how to interact with it first.

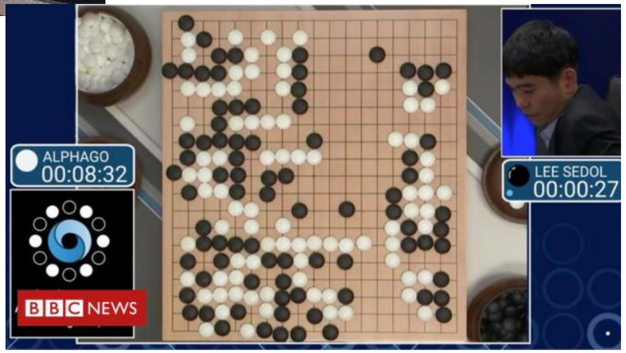
Cooperation is not just having multiple agents

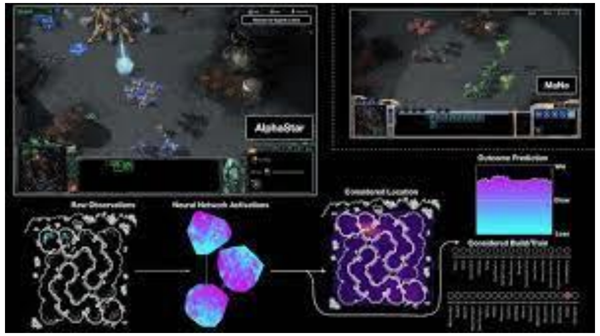
AI has seen significant progress in multi-agent settings

- Backgammon (e.g. TD-Gammon)
- Checkers (e.g. Chinook)
- Chess (e.g. DeepBlue)
- Go (e.g. AlphaGo)
- Poker (e.g.Pluribus)
- Starcraft (e.g. AlphaStar)
- Diplomacy
- …

But these, by and large, are games of conflict, not cooperation.





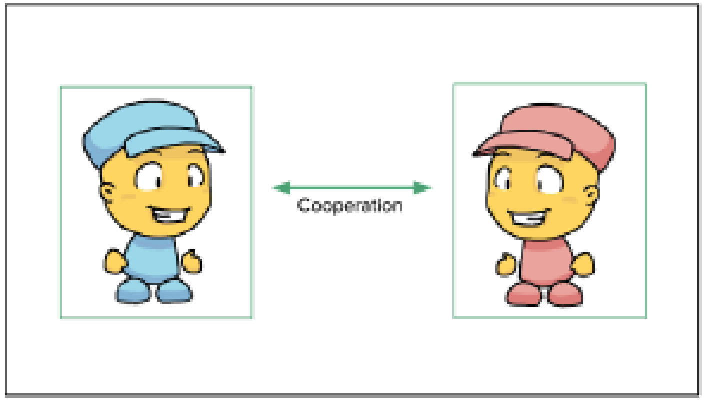


Cooperative AI

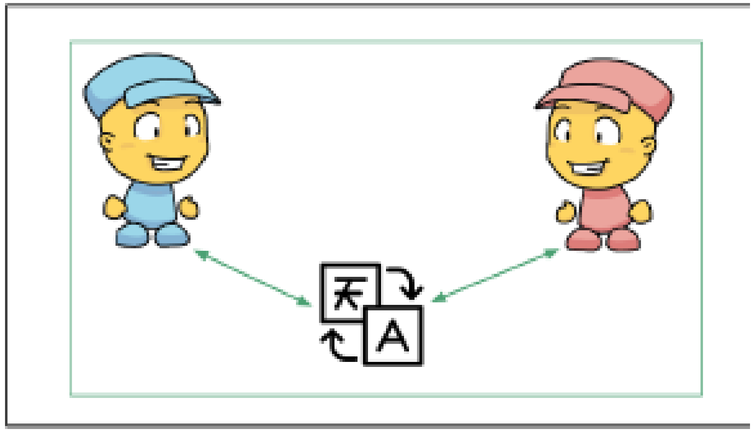
Cooperative AI

AI Research trying to help humans and machines find ways to improve their joint welfare.

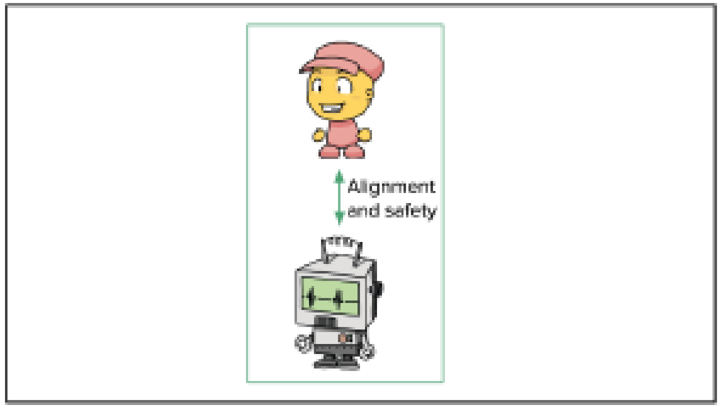
Different Types of Cooperation



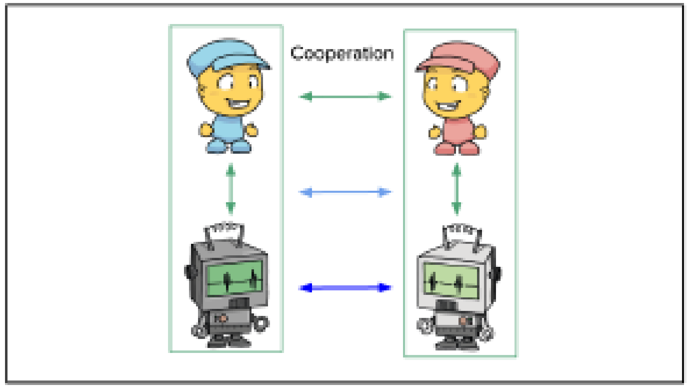
A: Human-Human Cooperation



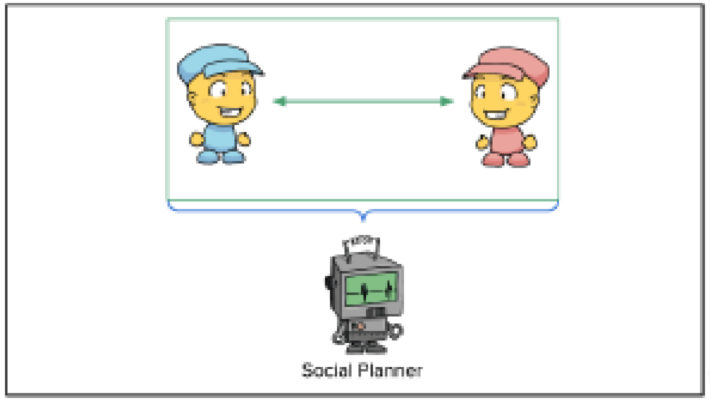
B: Cooperative Tools



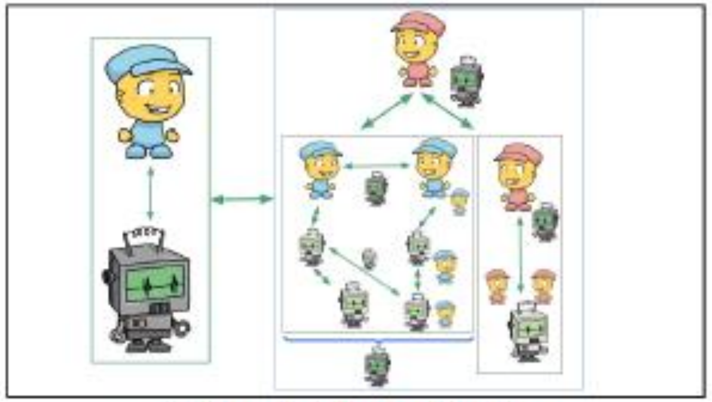
C: Alignment and Safety



D: {Human-AI}-{Human-AI} Cooperation



E: The Planner Perspective



F: Organizations and Society

Dafoe et al, 2021

To support cooperative AI we require

Understanding

The ability to take into account the consequences of actions, to predict others' behaviours, and the implications of another's beliefs and preferences

Communication

The ability to explicitly and credibly share information with others relevant to understanding behaviour, intentions, and preferences

Commitment

The ability to make credible promises when needed for cooperation.

Institutions

Social infrastructure – such as shared beliefs or rules – that reinforces understanding, communication and commitment.

Example - Autonomous Vehicles

There are numerous cooperative opportunities for AVs and other drivers (be they human or other AVs)

- AVs need to **understand** other drivers and road-users
- AVs need to be able to **communicate** with others
- AVs need to be able to make **commitments**
- Populations of drivers might be made better off by new **institutions** or **rules**

To support cooperative AI we require

Understanding

The ability to take into account the consequences of actions, to predict others' behaviours, and the implications of another's beliefs and preferences

Communication

The ability to explicitly and credibly share information with others relevant to understanding behaviour, intentions, and preferences

Commitment

The ability to make credible promises when needed for cooperation.

Institutions

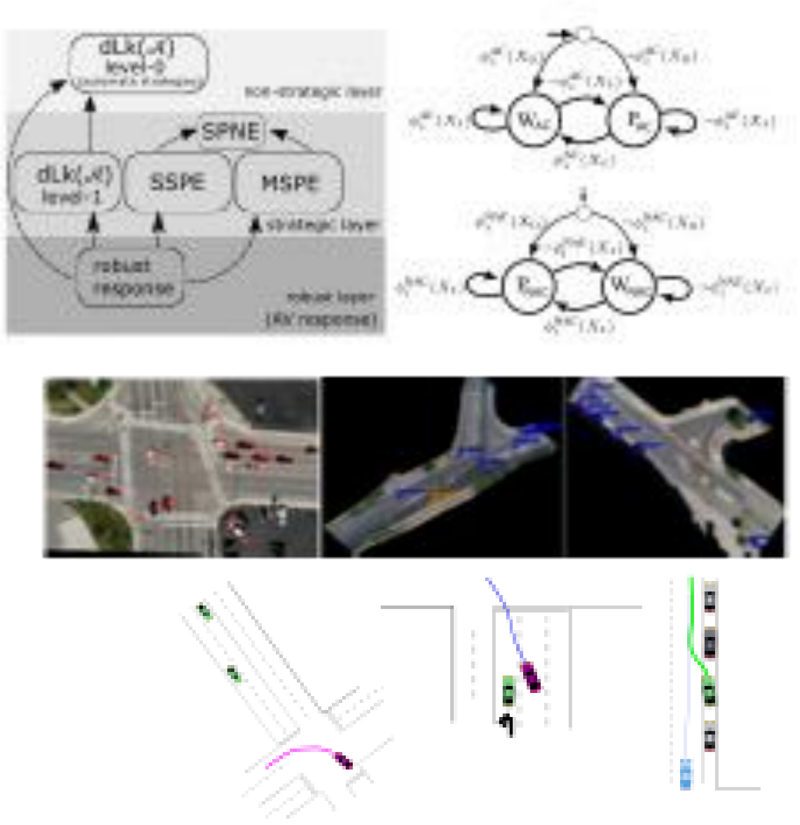
Social infrastructure – such as shared beliefs or rules – that reinforces understanding, communication and commitment.

Understanding

The ability to take into account the consequences of actions, to predict others' behaviours, and the implications of another's beliefs and preferences.

Possible directions forward

- Richer game theoretic models
- Preference elicitation and modelling
- Representation learning
- Inverse reinforcement learning
- Advances in computational theory of mind
- …



A. Sarkar et al

Game-theoretic models to support AVs

[A Sarkar, K. Larson, K Czarnacki, AAAI 2022, NeurIPS Workshop on Cooperative AI, 2021, AAMAS 2023]

Research Question: How should an AV safely handle other road users who show complex and varied behaviors?

Approach: There has been a shift from "predict-and-plan" approaches for driving behavior modelling to strategic models of non-zero sum games between road users and AVs.

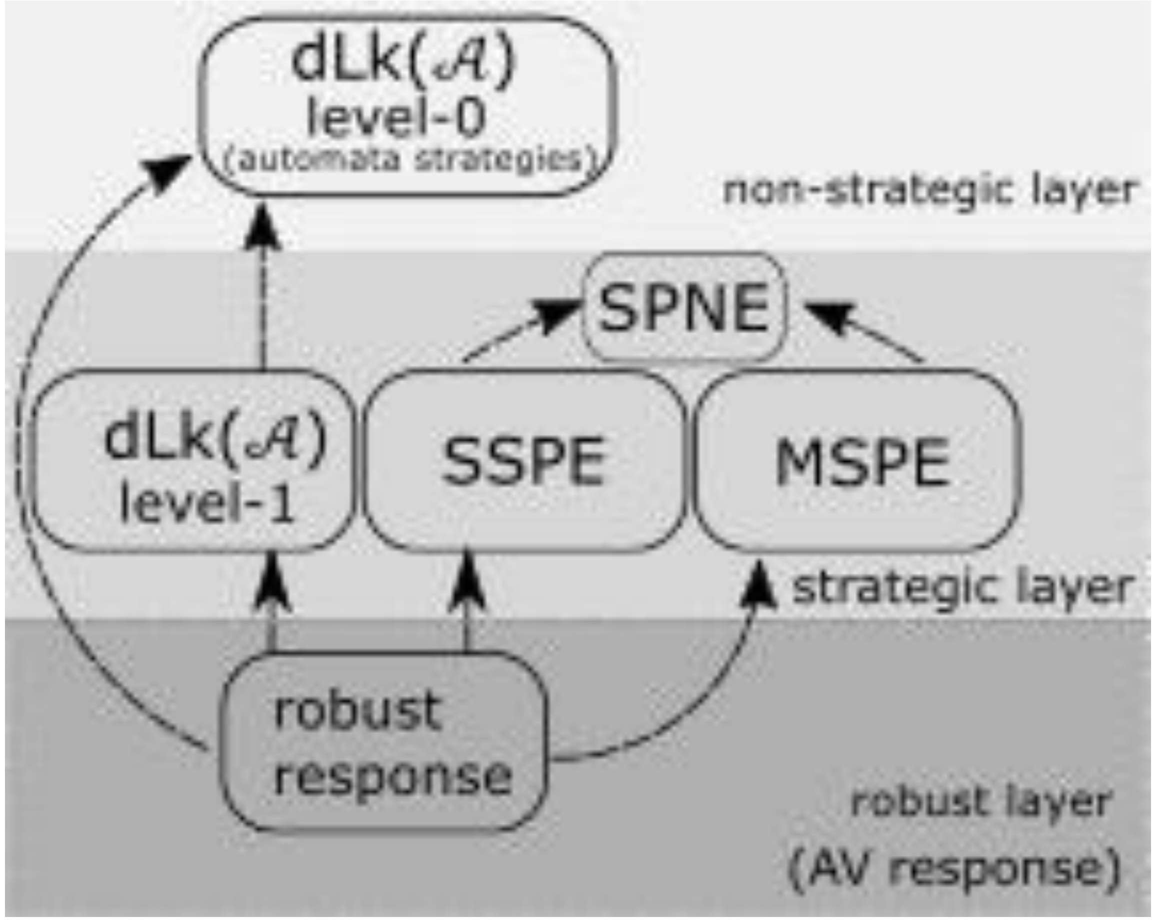
Challenge: (Human) driving behavior is diverse.

- Need to both model the diversity of human driving behavior as well as plan a response from the perspective of the AV



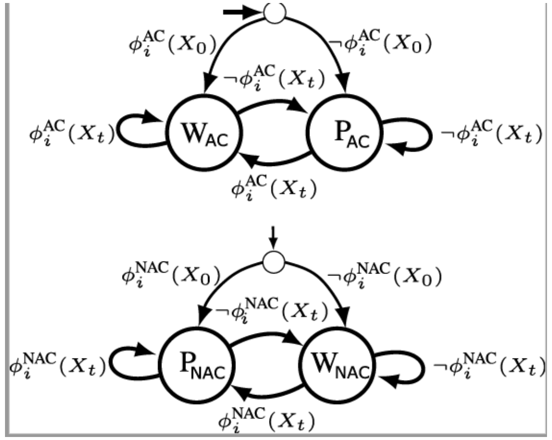
Game-theoretic models to support AVs

[A Sarkar, K. Larson, K Czarnacki, AAAI 2022, NeurIPS Workshop on Cooperative AI, 2021, AAMAS 2023]



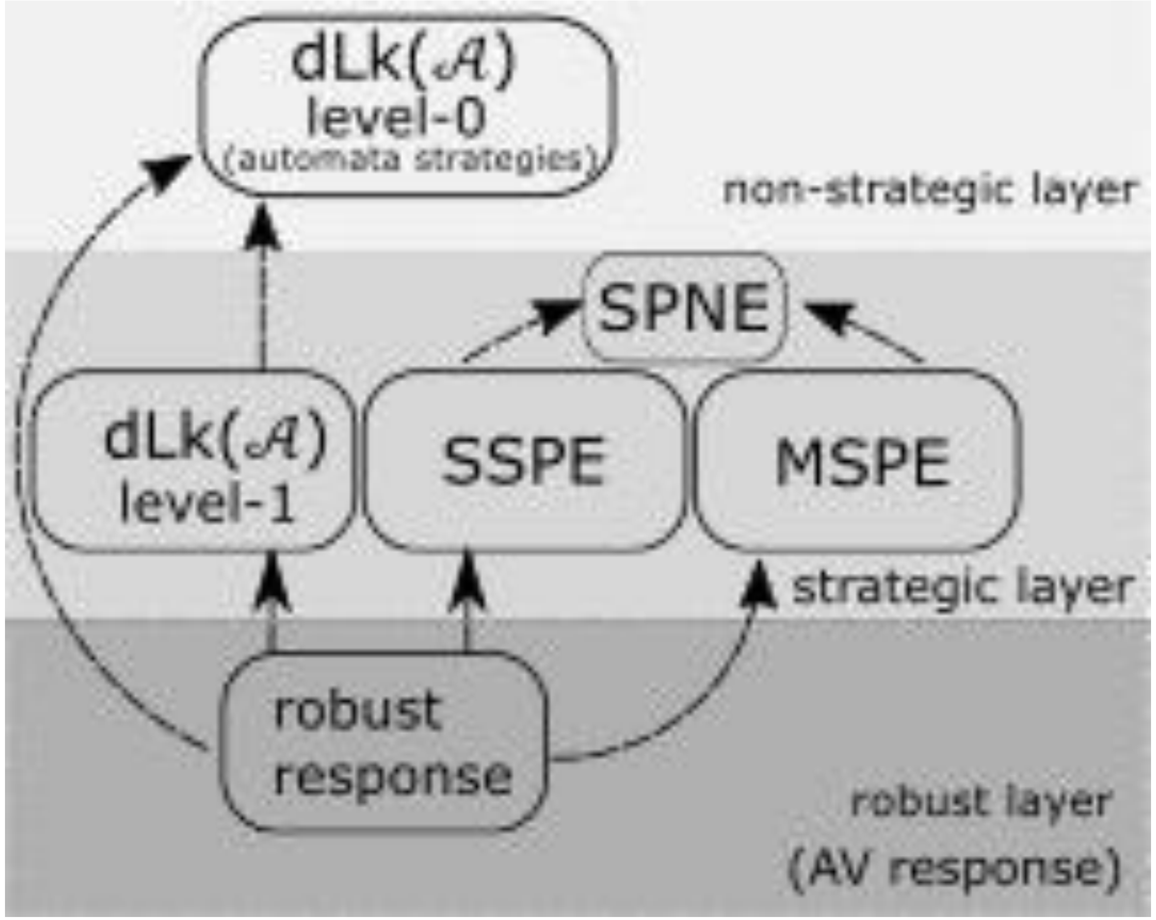
Generalized dynamic cognitive hierarchy models

- Non-strategic level: Agents (drivers) do not reason about others
  - We use automata strategies



Game-theoretic models to support AVs

[A Sarkar, K. Larson, K Czarnacki, AAAI 2022, NeurIPS Workshop on Cooperative AI, 2021, AAMAS 2023]

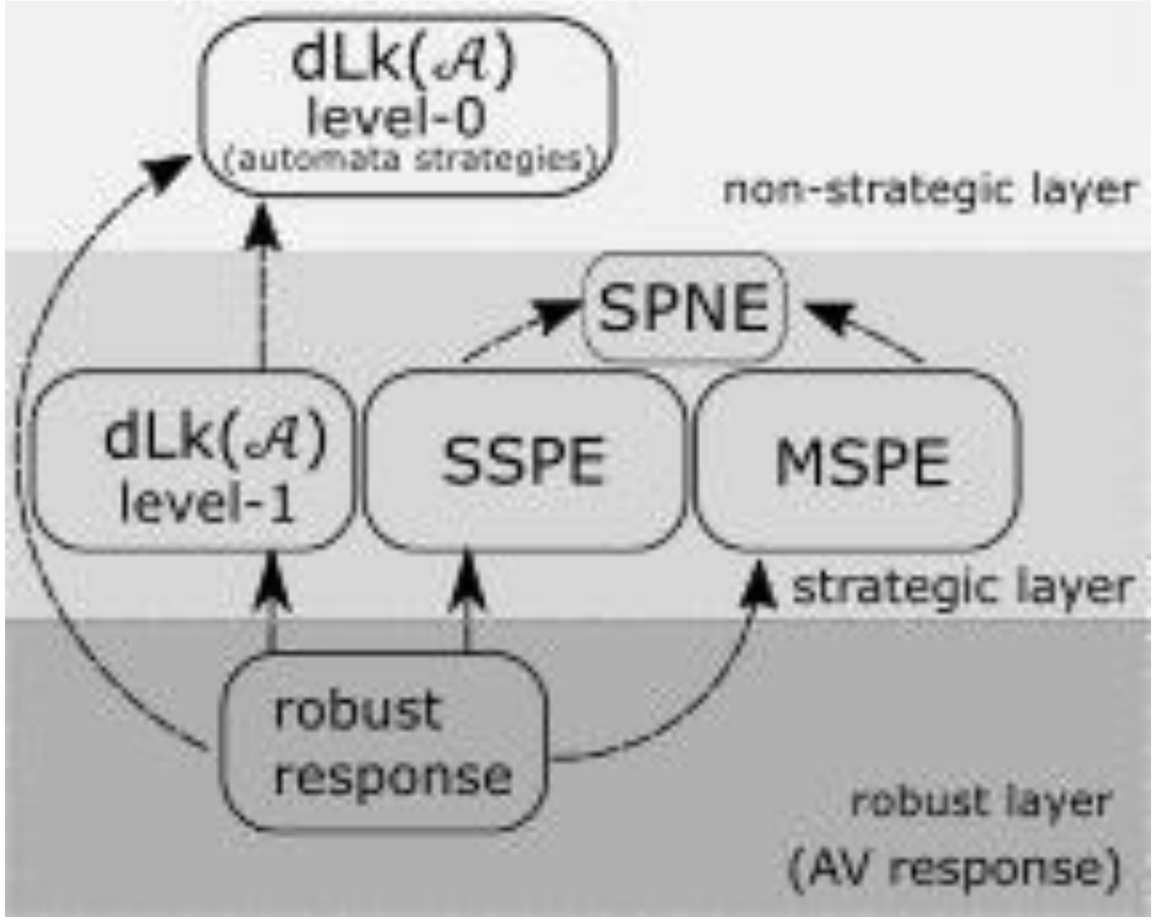


Generalized dynamic cognitive hierarchy models

- Non-strategic level
- Strategic level: Agents (drivers) reason about others on the road
    - dLk(level 1): dynamic quantal level-k model
    - Safety satisficing perfect equilibria (SSPE)
        - Select actions “close” to a NE as long as actions lead to outcomes what are above some safety aspiration threshold
    - Maneuver satisficing perfect equilibria (MSPE)
        - Select actions “close” to a NE as long as actions lead to outcomes that are above some maneuver aspiration threshold

Game-theoretic models to support AVs

[A Sarkar, K. Larson, K Czarnacki, AAAI 2022, NeurIPS Workshop on Cooperative AI, 2021, AAMAS 2023]

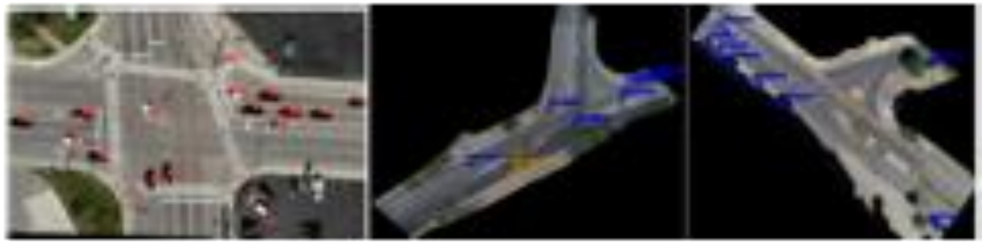


Generalized dynamic cognitive hierarchy models

- Non-strategic level
- Strategic level: Agents (drivers) reason about others on the road
  - dLk(level 1): dynamic quantal level-k model
  - Safety satisficing perfect equilibria (SSPE)
    - Select actions “close” to a NE as long as actions lead to outcomes what are above some safety aspiration threshold
  - Maneuver satisficing perfect equilibria (MSPE)
    - Select actions “close” to a NE as long as actions lead to outcomes that are above some maneuver aspiration threshold
- Robust layer: AV planning
  - Provides the ability to reason about heterogeneous populations of reasoners including strategic, non-strategic, and those following different models within each layer.

Game-theoretic models to support AVs

[A Sarkar, K. Larson, K Czarnacki, AAAI 2022, NeurIPS Workshop on Cooperative AI 2021, AAMAS 2023]



(a) Snapshot of naturalistic datasets (WMA and inD)

(b) Simulation of critical scenarios: intersection clearance, merge before intersection, parking pullout.

Evaluation:

- Evaluation on naturalistic data sets and simulations of critical scenarios

Findings

- Models matched human driving behaviour well compared to alternative models from literature
- For behaviour planning, robust response to heterogeneous behaviour models is both effective and stable across populations of drivers with different levels of risk tolerance

To support cooperative AI we require

Understanding

The ability to take into account the consequences of actions, to predict others' behaviours, and the implications of another's beliefs and preferences

Communication

The ability to explicitly and credibly share information with others relevant to understanding behaviour, intentions, and preferences

Commitment

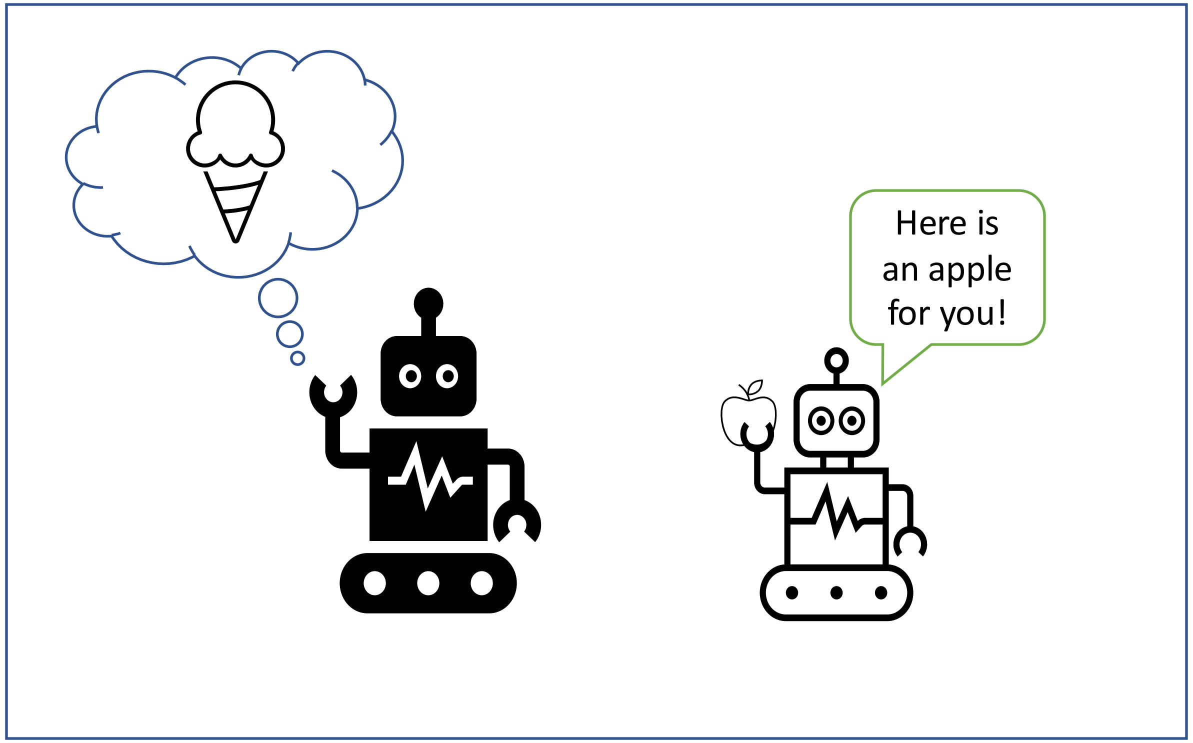
The ability to make credible promises when needed for cooperation.

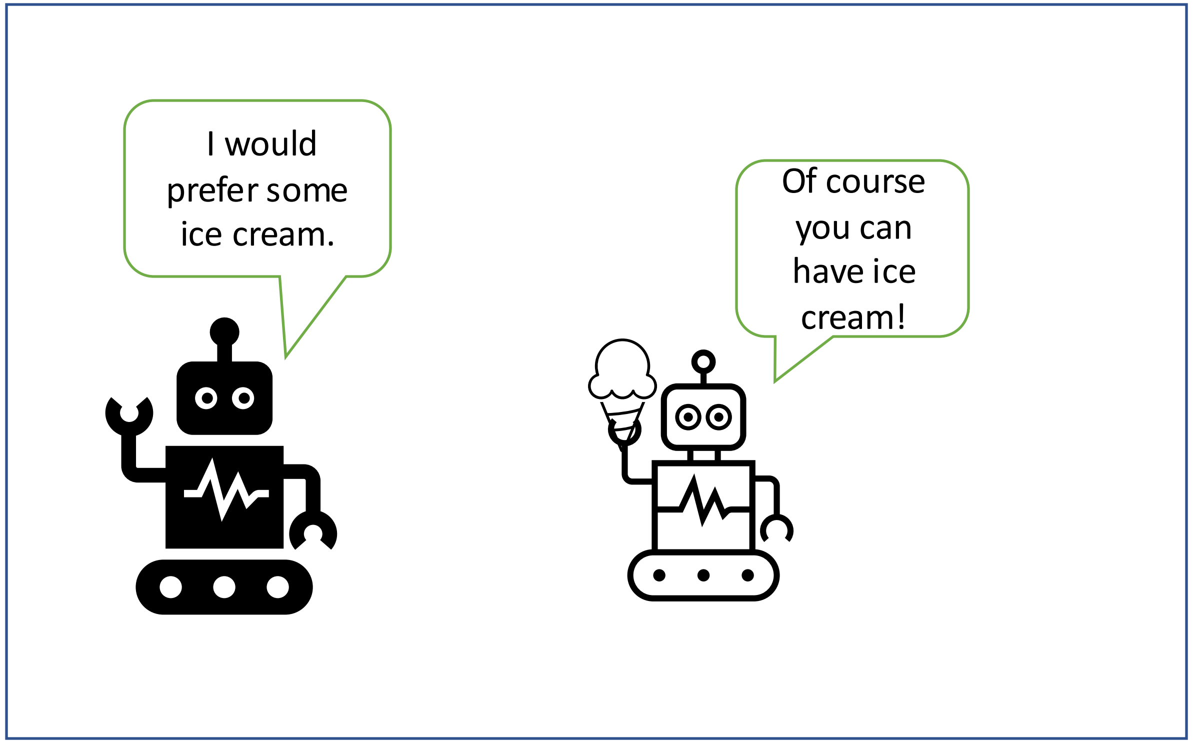
Institutions

Social infrastructure – such as shared beliefs or rules – that reinforces understanding, communication and commitment.

Communication

The ability to explicitly and credibly share information with others relevant to understanding behaviour, intentions, and preferences.





Communication

Where we are

- Learning through imitation/demonstrations (i.e. having a "teacher" in the system)
- Communication equilibria in game theory
- Emergence of simple communication in multiagent systems
- Large language models (e.g. GPT-3, BART)
- …

Where we might go

- Automating negotiations in complex open domains
- Moving from language models (P(text)) to *intentful models* (P(text|intent))
- Emergence of complex language from scratch

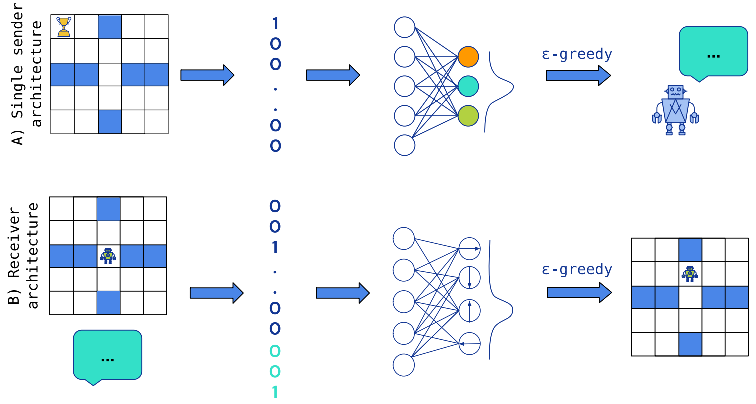
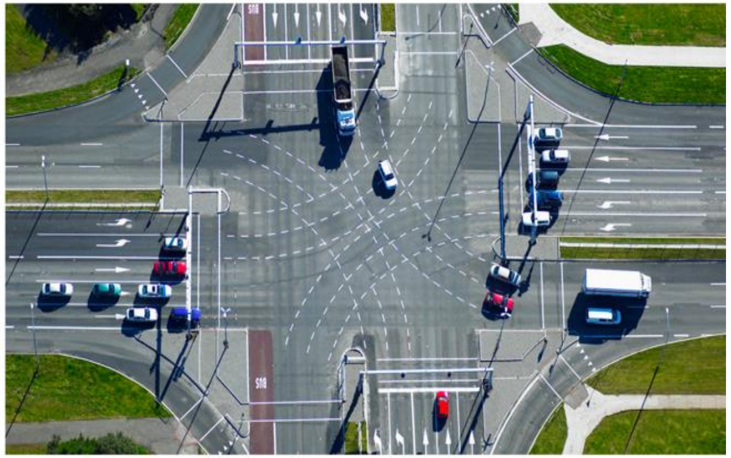


Figure 1: Both the sender and the receiver see the gridworld environment, yet only the sender sees the goal location (A). It selects a message action (a single symbol) based on the one-hot encoding of the goal location. The receiver selects a navigation action based on the multi-hot input vector that encodes its own location and the message (B).

I. Kajic et al



Autonomous vehicles might negotiate with each other for right of way. PHOTO_CONCEPTS/ISTOCKPHOTO

How artificial intelligence could negotiate better deals for humans

To support cooperative AI we require

Understanding

The ability to take into account the consequences of actions, to predict others' behaviours, and the implications of another's beliefs and preferences

Communication

The ability to explicitly and credibly share information with others relevant to understanding behaviour, intentions, and preferences

Commitment

The ability to make credible promises when needed for cooperation.

Institutions

Social infrastructure – such as shared beliefs or rules – that reinforces understanding, communication and commitment.

Commitment

The ability to make credible promises when needed for cooperation.

Where we are

- Trust and reputation systems
- Privacy preserving ML
- Smart contracts and distributed ledgers (blockchain)
- Assistants to track commitments

Where we might go

- Automated auditing of agent behaviour
- Automated reasoning about effects of commitments
- Novel commitment devices
- …

DOI:10.1145/3448248

The pursuit of responsible AI raises the ante on both the trustworthy computing and formal methods communities.

BY JEANNETTE M. WING

Trustworthy AI

Trusted AI and the Contribution of Trust Modeling in Multiagent Systems

Blue Sky Ideas Track

Robin Cohen, Mike Schaekermann, Sihao Liu, Michael Cormier
Computer Science; University of Waterloo; Waterloo, Canada

To support cooperative AI we require

Understanding

The ability to take into account the consequences of actions, to predict others' behaviours, and the implications of another's beliefs and preferences

Communication

The ability to explicitly and credibly share information with others relevant to understanding behaviour, intentions, and preferences

Commitment

The ability to make credible promises when needed for cooperation.

Institutions

Social infrastructure – such as shared beliefs or rules – that reinforces understanding, communication and commitment.

Institutions

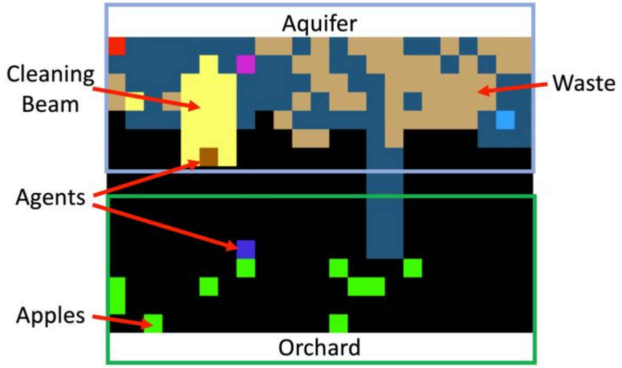
Social infrastructure – such as shared beliefs or rules – that reinforces understanding, communication and commitment.

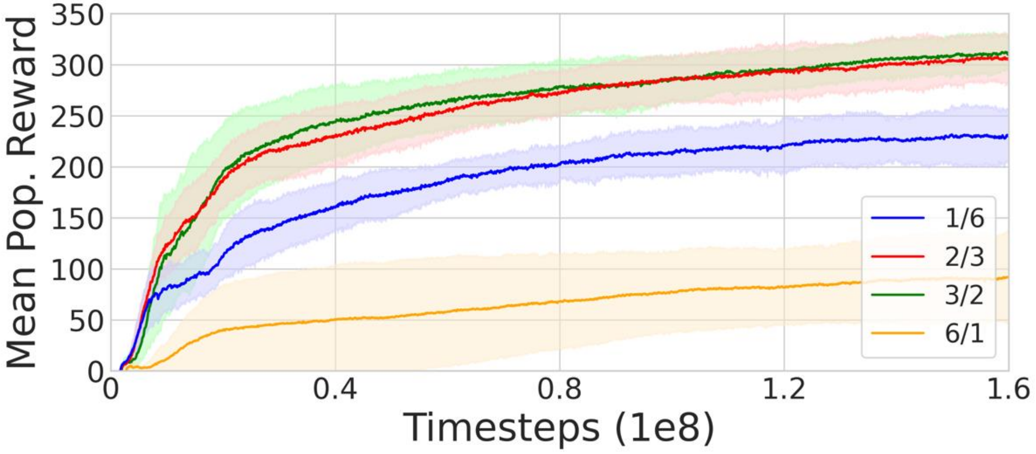
Institutional structures can take many forms

- Informal norms like holding a door open for someone
- Formalized institutions like rules that describe voting processes for elections

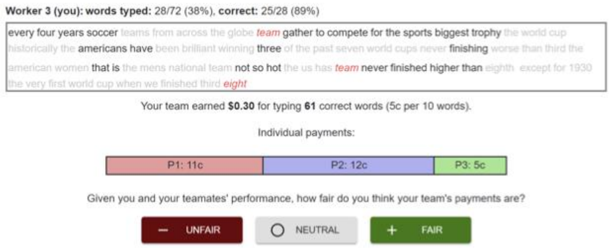
Institutional Structures

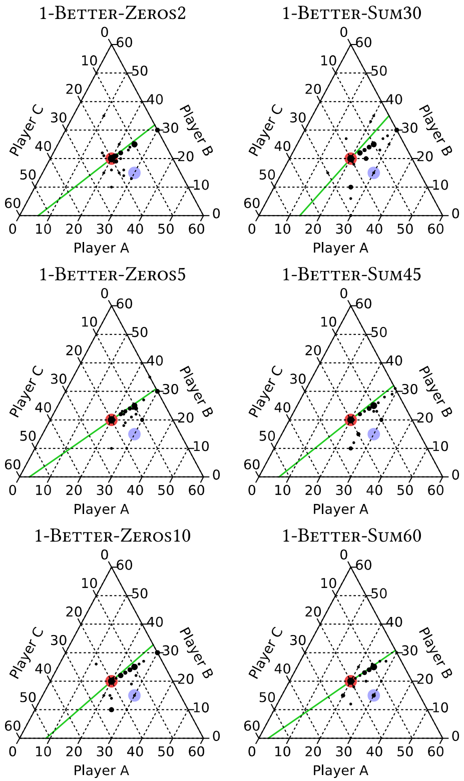
Teams as a way of promoting cooperation [Radke, Larson, and Brecht, IJCAI 2022, AAMAS 2023, IJCAI 2023]





What are effective ways of designing group rewards? [d'Eon, Larson, and Law, CSCW 2019, d'Eon and Larson, AAMAS 2020]

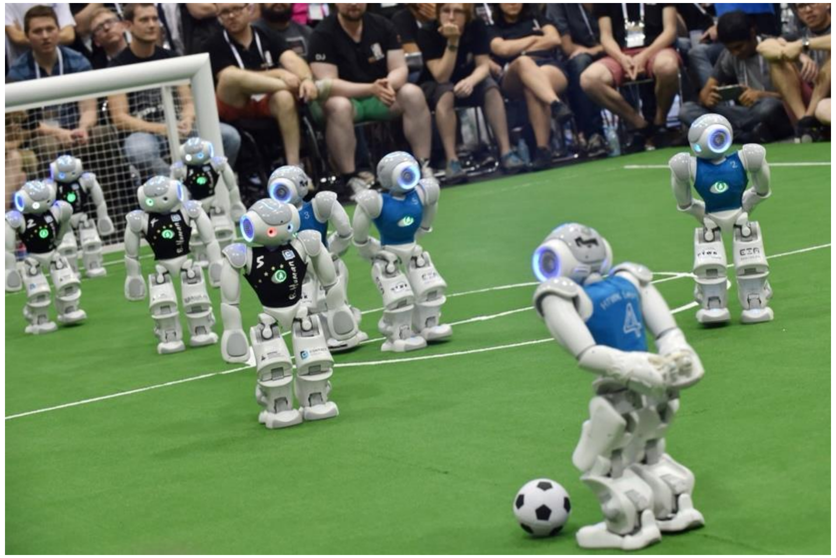




Towards a better understanding of teams in multiagent systems [Radke, Larson, and Brecht, AAAI 2022, AAMAS 2023, IJCAI 2023]

Base Environment

- Stochastic Game: $\mathcal{G} = \langle N, S, \{A\}_{i \in N}, \{R\}_{i \in N}, P, \gamma, \Sigma \rangle$
  - N: Set of all agents, initialized randomly
  - S: State space observable by all agents
  - $\{A\}_{i \in N}$: Joint action space for all agents (indexed by i)
  - $\{R\}_{i \in N}$: Joint reward space for all agents (indexed by i)
  - $P : S \times A \mapsto \Delta(S)$: Represents the transition function
  - γ : Discount factor
  - Σ : Represents the policy space of all agents

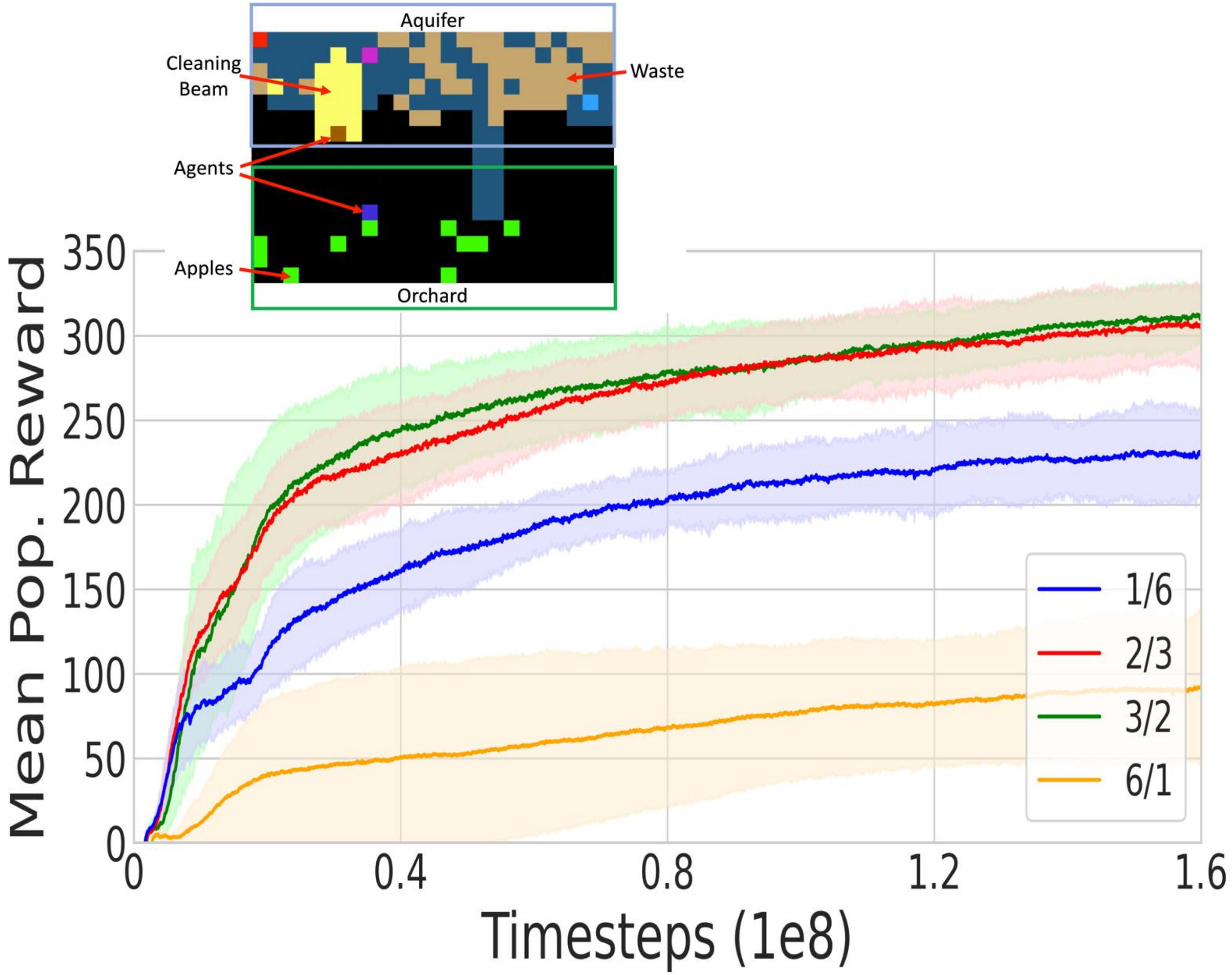


- **Predefined Teams** $\langle \mathcal{G}, \mathcal{T} \rangle$;

$$\mathcal{T} = \{T_i | T_i \subseteq N, \cup T = N, T_i \cap T_j = \emptyset \forall i, j\}$$

- Agents have **modified** reward functions

Towards a better understanding of teams in multiagent systems [Radke, Larson, and Brecht, AAAI 2022, AAMAS 2023, IJCAI 2023]



Smaller teams

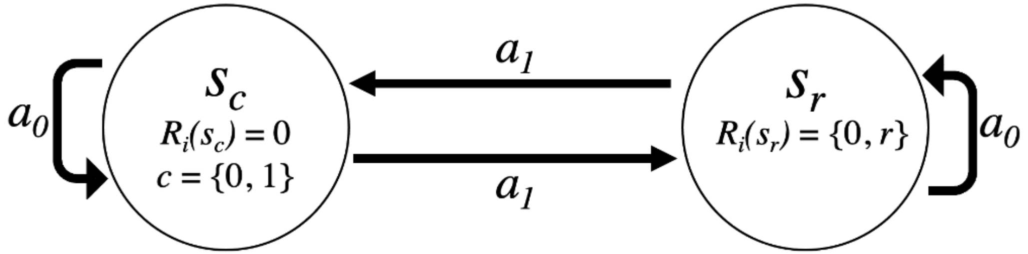
Fully cooperative agents

Fully independent agents

Emergence of specialization in policies

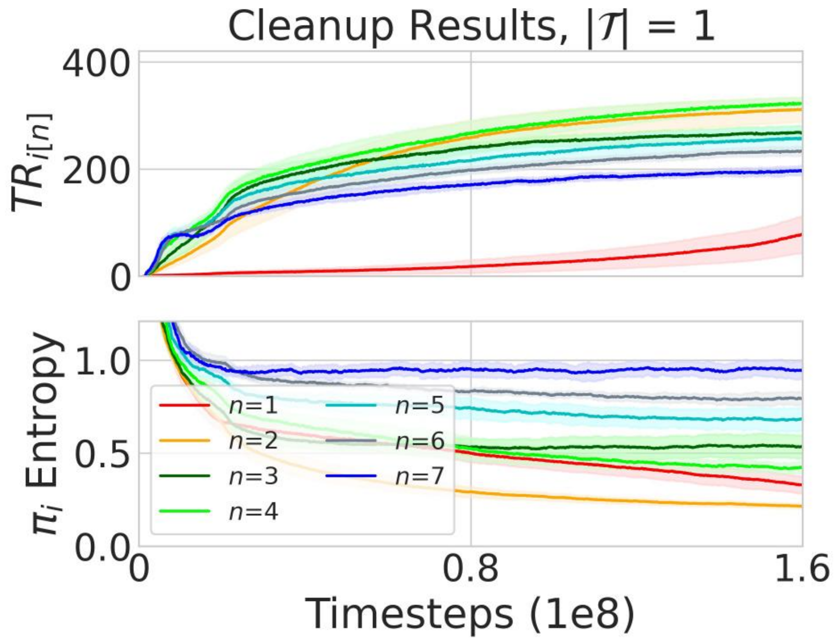
Towards a better understanding of teams in multiagent systems [Radke, Larson, and Brecht, AAAI 2022, AAMAS 2023, IJCAI 2023]

Require reward-causing state-action pairs [Aronja-Medina et al, 2019]



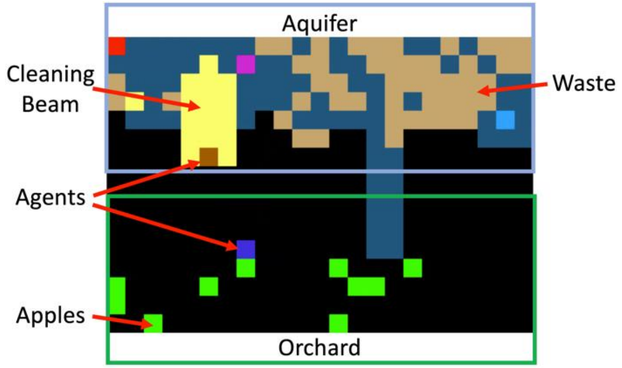
BUT

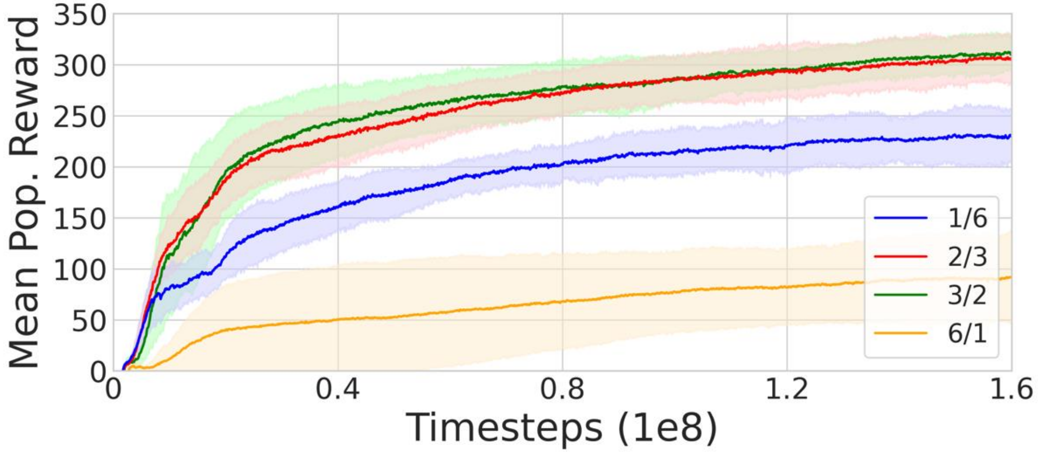
If team size becomes too large, we fall into an information sparsity scenario where credit assignment is challenging



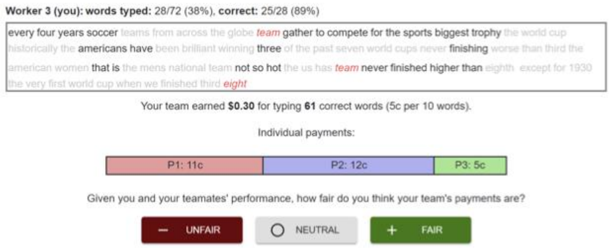
Institutional Structures

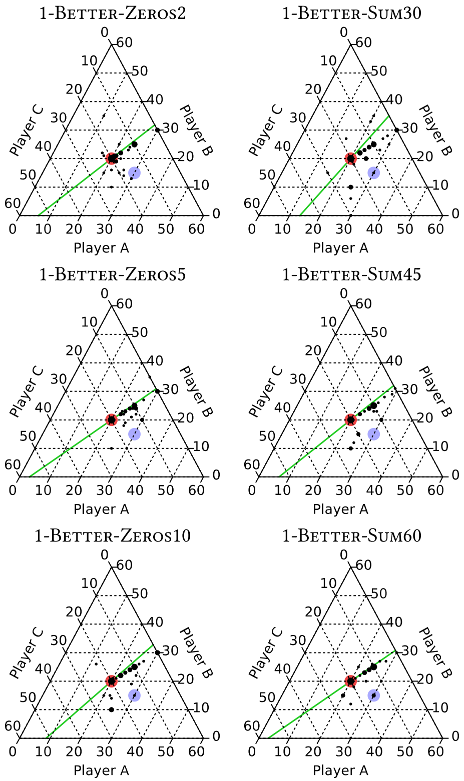
Teams as a way of promoting cooperation [Radke, Larson, and Brecht, IJCAI 2022, AAMAS 2023, IJCAI 2023]





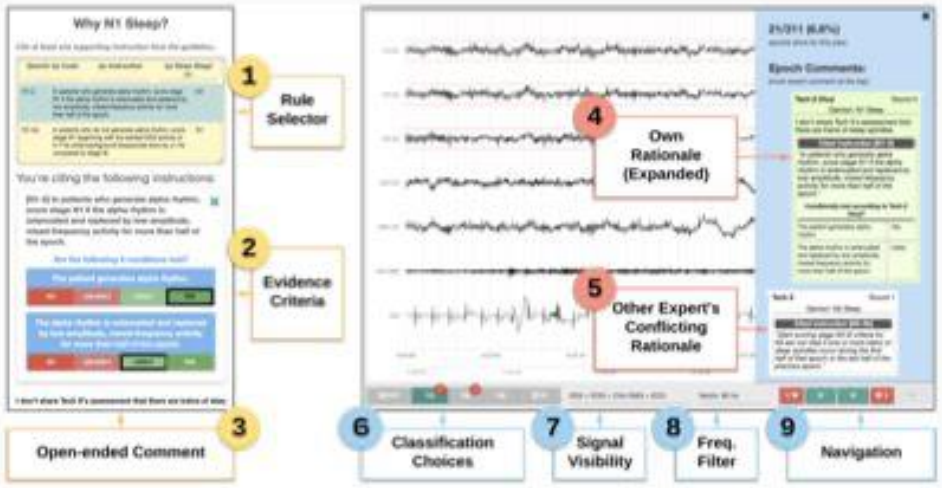
What are effective ways of designing group rewards? [d'Eon, Larson, and Law, CSCW 2019, d'Eon and Larson, AAMAS 2020]





Motivating Problem

With colleagues in HCI, we have been designing platforms to support collaborative work [CSCW18, CSCW19a, CSCW19b, CHI20]



(a) Rationale form. (b) Data view for disagreement case with expert discussion.

Fig. 1. Interface for structured adjudication of classification decisions in medical time series analysis.

[Schaekermann et al, CSCW 2019a]

How do you reward workers for their effort?

Supporting Collaborative Work Through Fair Reward Sharing [d'Eon, Goh, Larson, Law, CSCW2019]

We studied collaborative tasks and workers' perception of fair and unfair payments.

While workers were biased, they were perceptive of fair and unfair payments.

Fairness mattered.

Is There a Relationship Between the Shapley Value and Human Reward-Division? [d'Eon, Larson, AAMAS 2020]

Axioms of Fairness

Efficiency: $\sum_i v_i = f(N)$

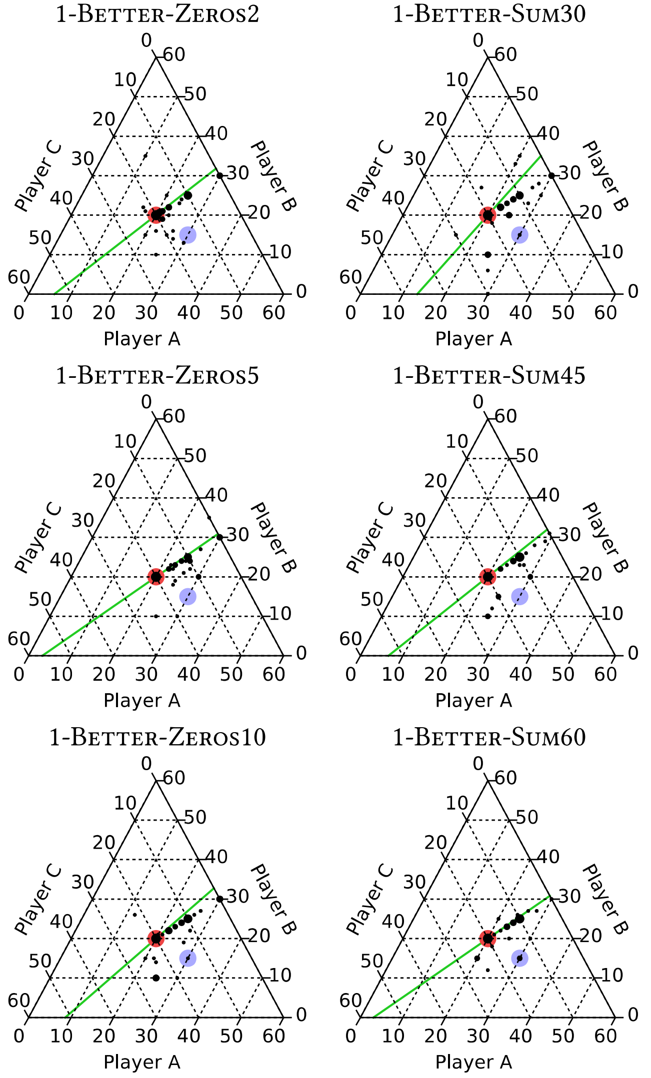
Symmetry: Equal players are rewarded equally.

Null Players: A player who contributes nothing to any coalition should get no reward.

Additivity: If f and g are two games, then define a new game $(f+g)(C) = f(C) + g(C)$ for all C. Then $v_i(f+g) = v_i(f) + v_i(g)$.

Shapley Value

$$Sh_i(f) = \sum_{C \subseteq N \setminus i} \frac{|C|!(|N|-|C|-1)!}{|N|!}(f(C \cup \{i\}) - f(C))$$



Data-Driven Axiomatic Testing

Axioms of Fairness

Efficiency: $\sum_i v_i = f(N)$ ✓

Symmetry: Equal players are rewarded equally. ✓

Null Players: A player who contributes nothing to any coalition should get no reward. ✗

Additivity: If f and g are two games, then define a new game $(f+g)(C) = f(C) + g(C)$ for all C. Then $v_i(f+g) = v_i(f) + v_i(g)$. ✗

Shapley Value

$$Sh_i(f) = \sum_{C \subseteq N \setminus i} \frac{|C|!(|N| - |C| - 1)!}{|N|!} (f(C \cup \{i\}) - f(C))$$

Data-Driven Axiomatic Testing

Axioms of Fairness

Efficiency: $\sum_i v_i = f(N)$

Symmetry: Equal players are rewarded equally.

Null Players: A player who contributes nothing to any coalition should get no reward.

Additivity: If f and g are two games, then define a new game $(f+g)(C) = f(C) + g(C)$ for all C. Then $v_i(f+g) = v_i(f) + v_i(g)$.

Shapley Value

$$Sh_i(f) = \sum_{C \subseteq N \setminus i} \frac{|C|!(|N| - |C| - 1)!}{|N|!} (f(C \cup \{i\}) - f(C))$$

Young's [1985] alternative axiomatization of Shapley replaces null-player and additivity with a strong monotonicity property.

Relaxations of strong monotonicity include **local monotonicity** [Casajus and Huettner, 2013] and **coalitional monotonicity** [Young 1985].

Local Monotonicity:

At least 89% of our data was consistent.

Coalitional monotonicity:

At least 77% of our data was consistent (for games where coalitional monotonicity was defined).

Data-Driven Axiomatic Approaches

Process requires two key ingredients

Data:

- Controlled experiments allow for testing a particular axiom
- In-the-wild experiments may provide more representative reactions
- (Speculative) Possibly use LLMs to generate data [e.g. Horton, 2023]

Testing Axioms:

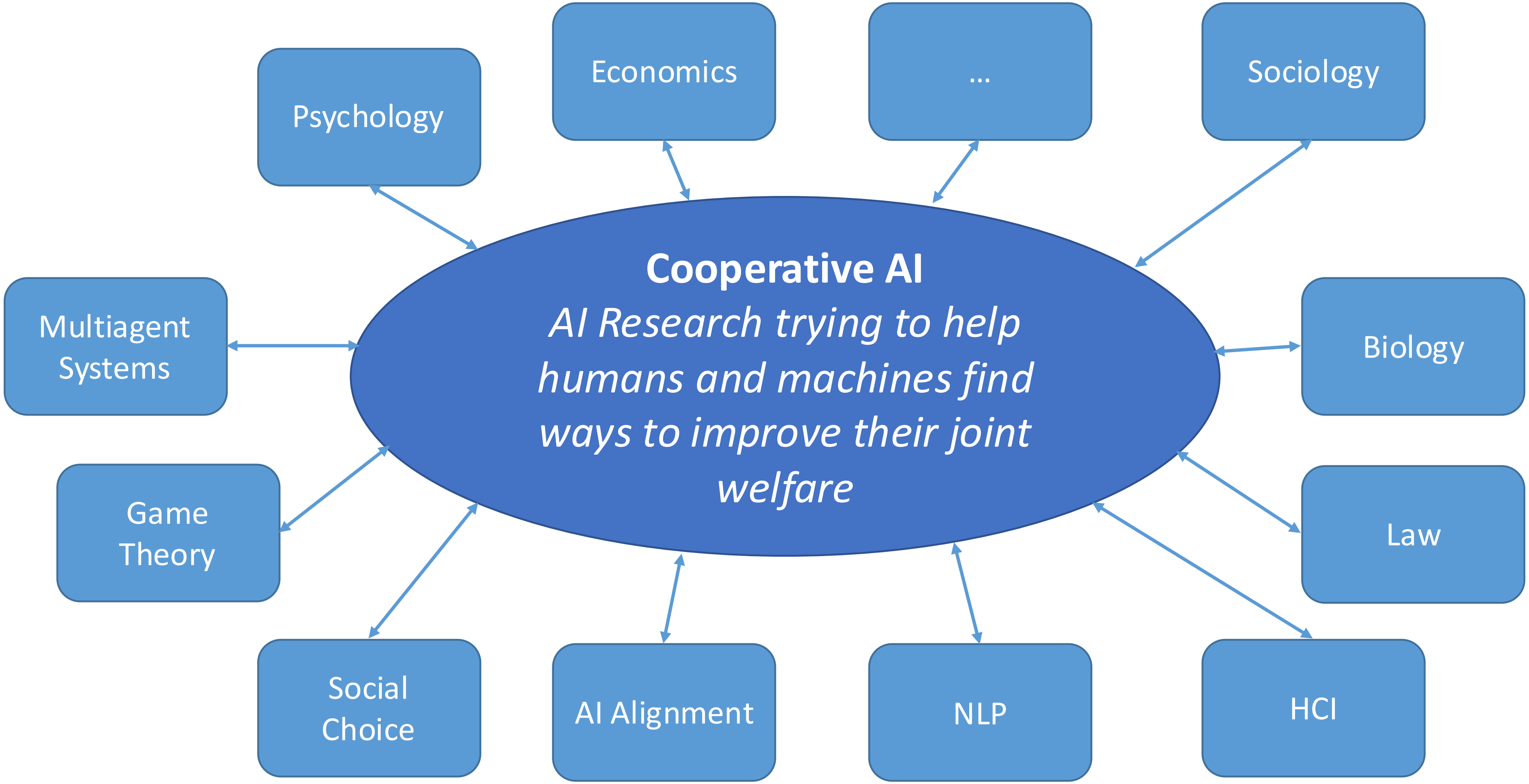
- Count violations of axioms
- Quantify how drastically an axiom has been violated
  - Development of rigorous tools for quantifying axiomatic breakdown

A possible approach for testing and refining institutional structures (i.e. rules for supporting collaborative and cooperative behaviours).

Cooperative AI

Cooperative AI
AI Research trying to help humans and machines find ways to improve their joint welfare.

Cooperative AI

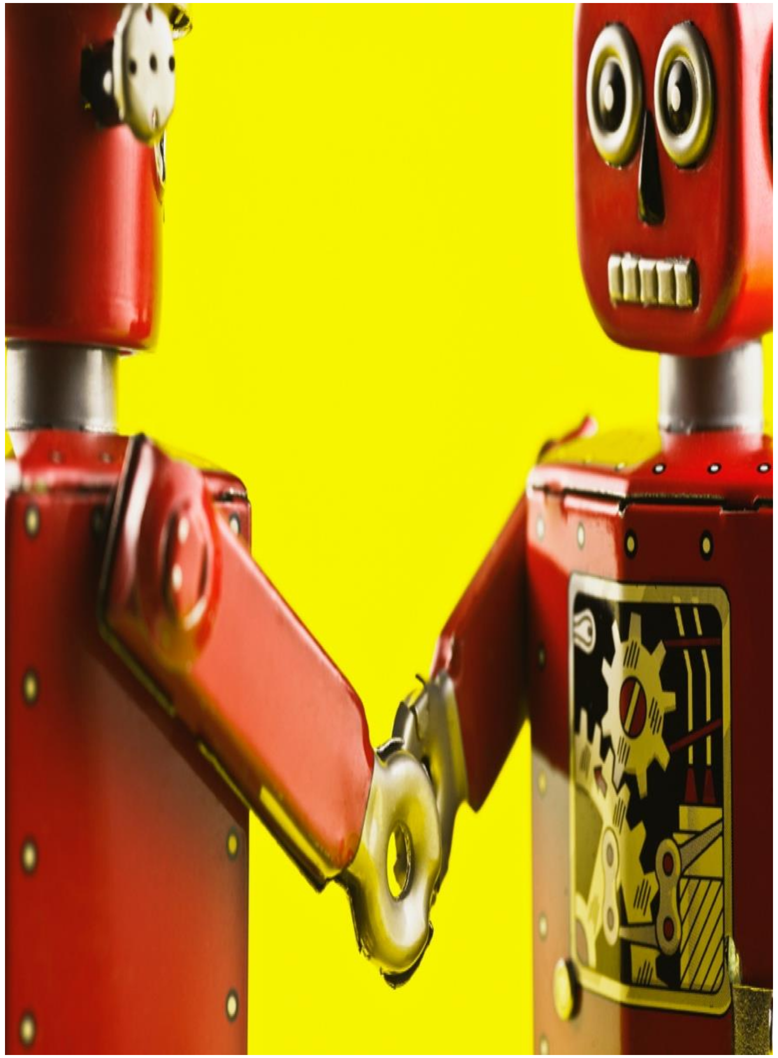


Cooperation should be at the centre of AI research

It is unlikely to emerge as a by-product of other kinds of AI research.

Research in this area is inherently inter-disciplinary and will require many different perspectives.

In general, we need to move from individual objectives to shared, poorly defined, ways humans solve social problems: creating language, norms and institutions.



Questions