Symmetric Machine Theory of Mind

ICML 2022



Melanie Sclar, Graham Neubig, Yonatan Bisk

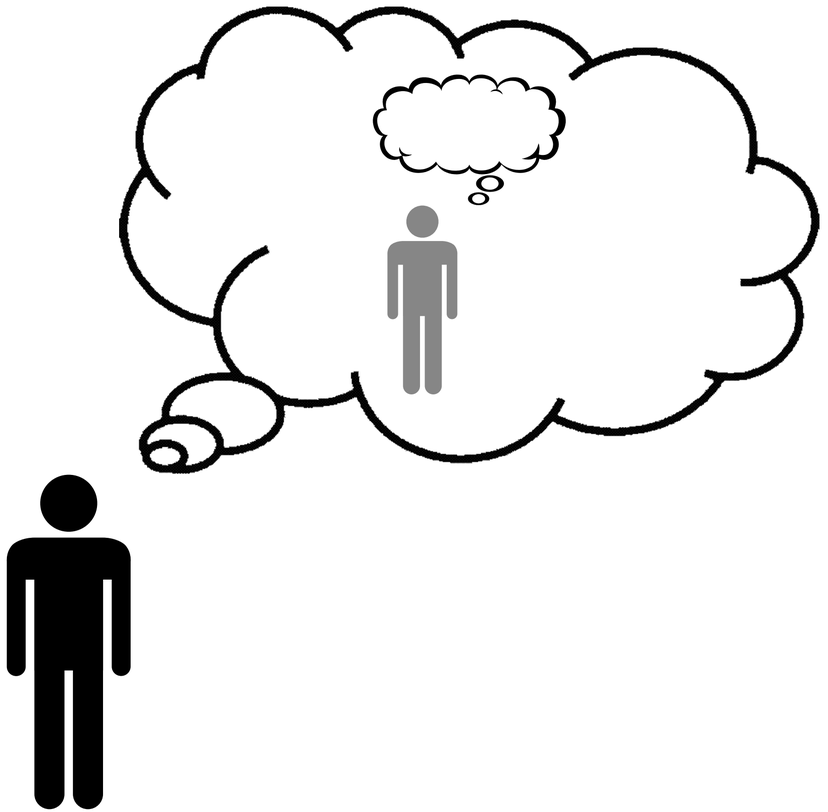
University of Washington

Carnegie Mellon University

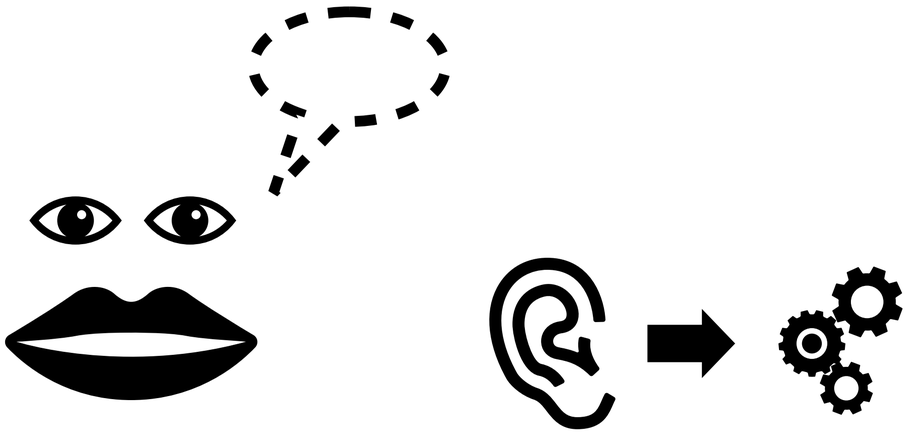
Carnegie Mellon University

Theory of Mind

The ability to understand others' mental states and act upon them

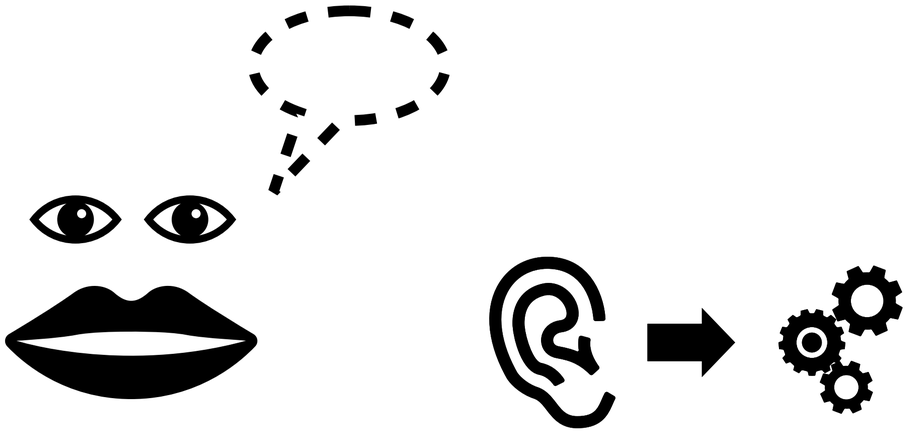


Most of Machine Theory of Mind prior art **is** *asymmetric*

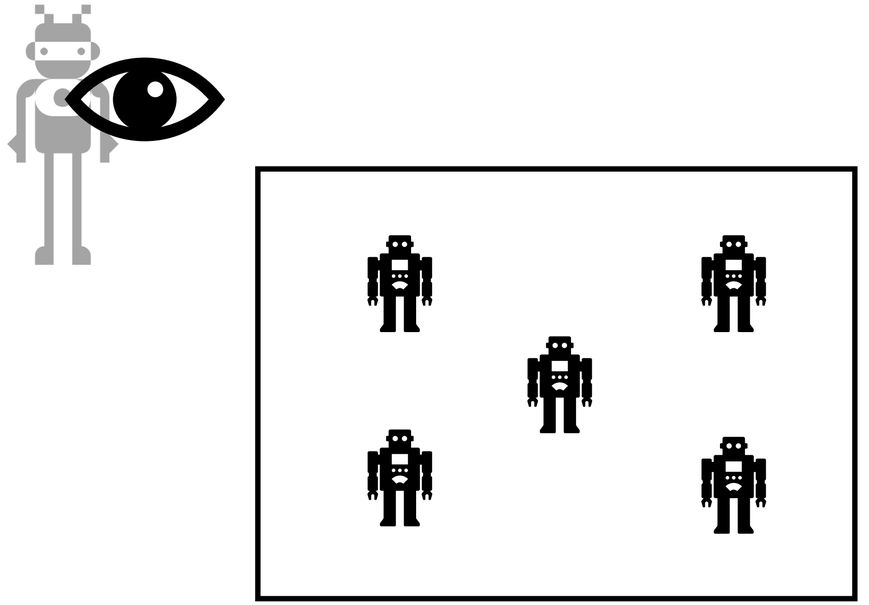


Speaker-Listener

Most of Machine Theory of Mind prior art **is** *asymmetric*



Speaker-Listener



Passive Observer

Symmetric **Machine Theory of Mind environments**

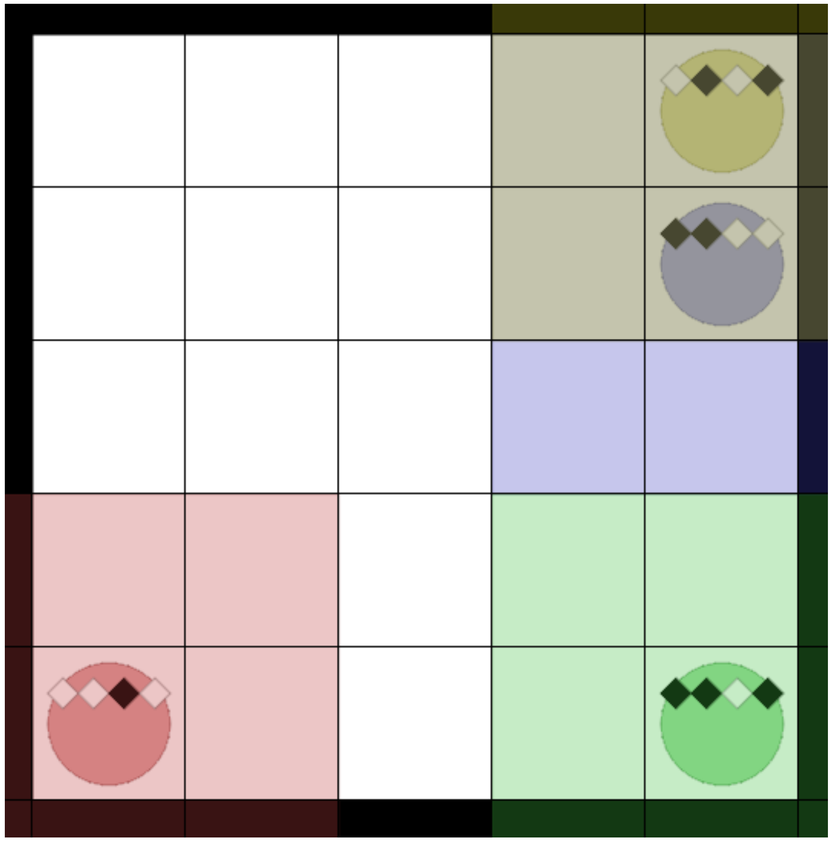
- Imperfect Information
- Theory of Mind is required to perform the task successfully (*information-seeking behavior*)
- All agents have equal abilities: same action space, observability

Symmetric Machine Theory of Mind is richer and more realistic

Human interaction is usually multi-party, with no unique predefined roles or passive observers.

We Propose *SymmToM*, a novel multi-agent Symmetric Machine Theory of Mind environment

SymmToM is simple, yet proves challenging even for models tailored to the task.
This makes it compelling to use to evaluate future models!



🔴🔵🟡🟢 = agents
shaded = hearing range
◇ = unknown information
◆ = known information

- Perfect vision, imperfect hearing, up-down-left-right movement
- Communication through fixed set of symbols
- **Gaining reward efficiently requires theory of mind: reward is given for sharing and receiving new knowledge**
- Initial information of each agent is public

Tailoring RL models for SymmToM to prove its difficulty

- We modify MADDPG —a well-known multi-agent actor-critic framework— to include a matrix (one per agent) reflecting perceived knowledge
- Not necessarily a reflection true knowledge, but rather what information pieces each agent knows or estimates others know
- How do we reason through interactions an agent did not witness?

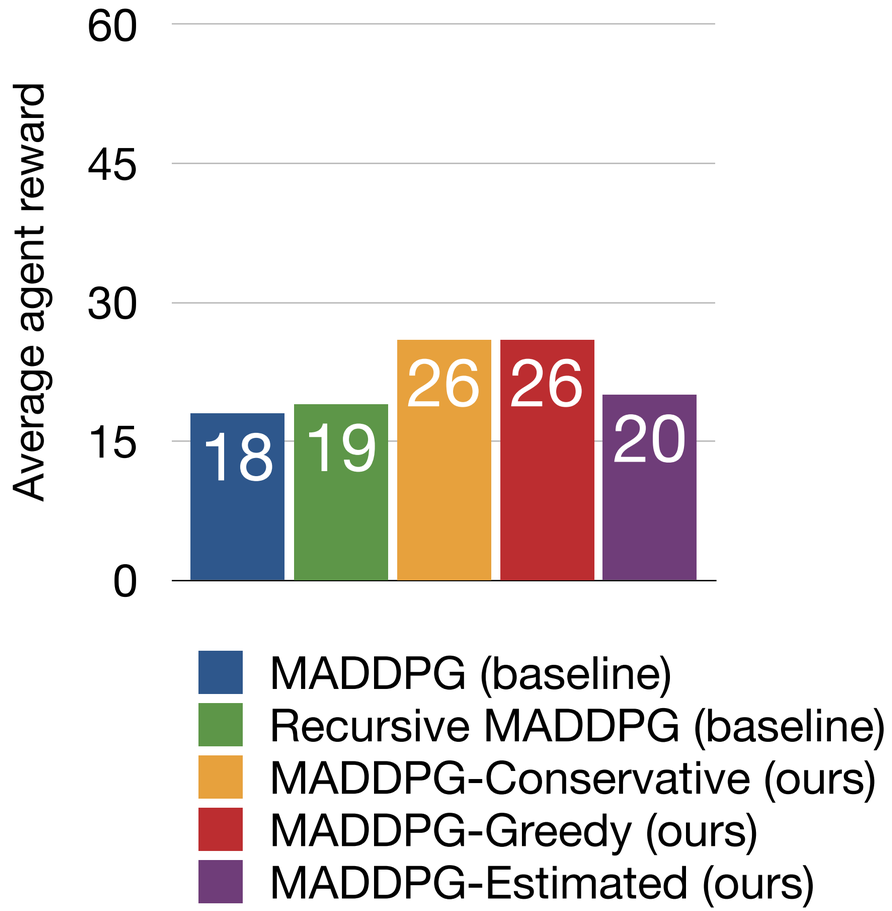
Estimating unseen interactions

- ***Conservative***: only what you personally witnessed is known
- ***Greedy***: assume everyone plays optimally, deduce unseen interactions from there
- ***Estimated***: assign a probability to each agent knowing each information piece

Results

Average reward per agent for the setting {3 agents, 6x6 grid, 6 information pieces}

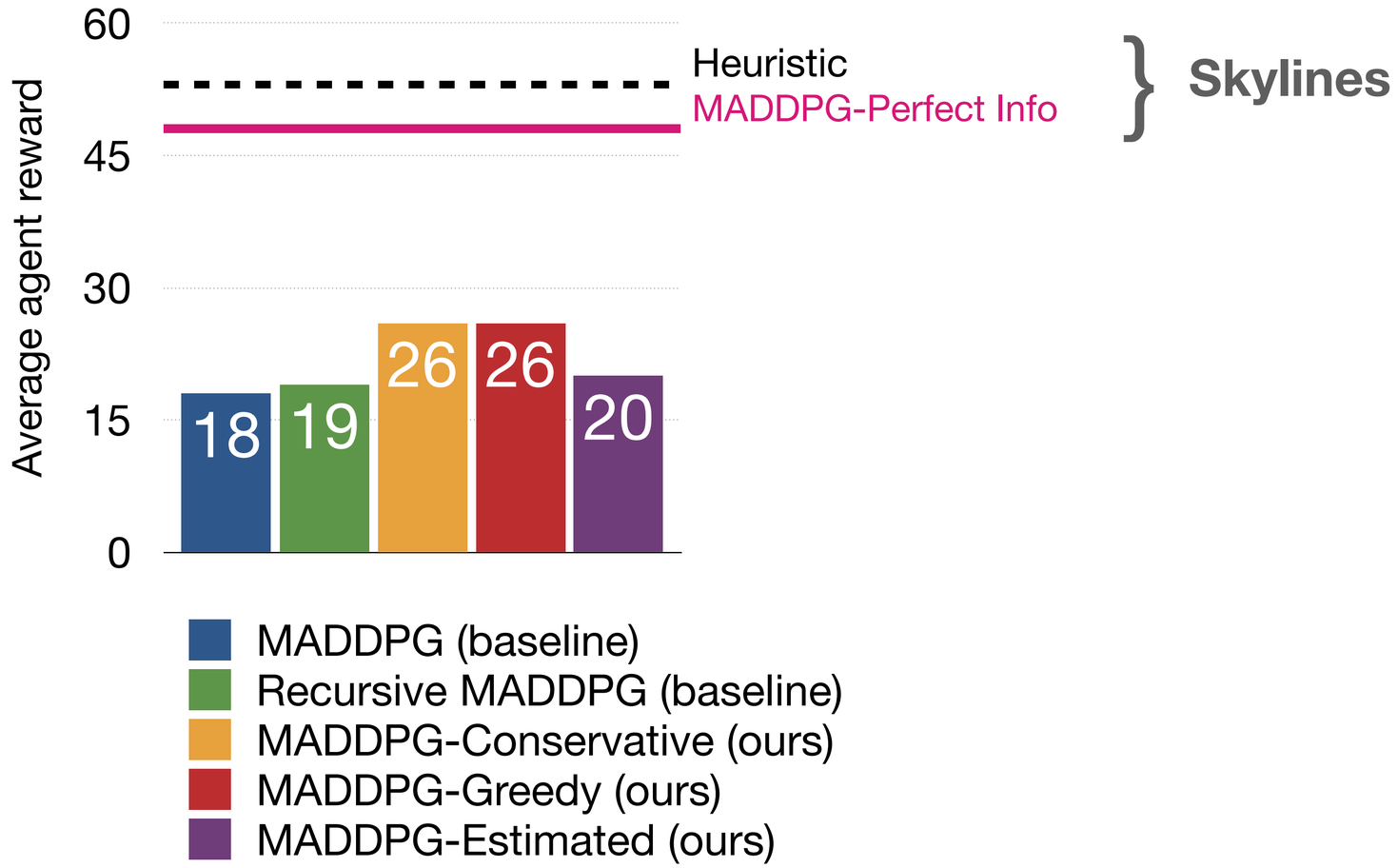
(one of the twelve settings analyzed!)



Results

Average reward per agent for the setting {3 agents, 6x6 grid, 6 information pieces}

(one of the twelve settings analyzed!)



Conclusions

- SymmToM is a simple environment that proves extremely hard even for well-known multi-agent RL models tailored to it
- Theory of Mind is a key phenomenon to model for successful multi-agent and human-agent interactions and often tested in asymmetric settings
- We invite everyone to use SymmToM to test new and more efficient approaches that solve this task, to move on to more complex and nuanced scenarios

Thank you!

Symmetric Machine Theory of Mind

Come chat at poster #228 (Hall E) on 20 Jul 6:30 p.m. EDT

github.com/msclar/symmtom