



The Agency Gap: Testing Human-Human and Human-AI Collaboration Using a Minimal Coordination Game

Shaozhe Cheng^{1,2}, Tao Gao², Michael Tomasello¹ & Tamar Kushnir¹

¹Department of Psychology & Neuroscience, Duke University ²Department of Communication, UCLA

Motivation

The Agency Gap

- A central challenge of cooperation is how agents coordinate their actions toward a joint goal amid potentially competing goals.
- Humans evolved *shared agency* to solve this coordination problem (Searle, 1990; Tomasello, 2022): they form shared intentions, commit to them, and use signaling to make their goals transparent to partners.
- Modern AI achieves impressive outcomes, yet its “agency” is often treated as a byproduct of optimizing predefined objectives (RL and LLMs). As a result, it functions more like a tool (asocial or social) than as a genuinely collaborative agent.

Central Questions

- To what extent is human-AI collaboration human-like and human-compatible?
- Do the signatures of shared agency (joint commitment and social signaling) emerge when the partner is an AI? Does this depend on the sociality of the AI partner?
- Can children, who default to shared agency, adapt to AI partners that lack it?



Game Demo & Preprint paper

shaozhe.cheng@duke.edu

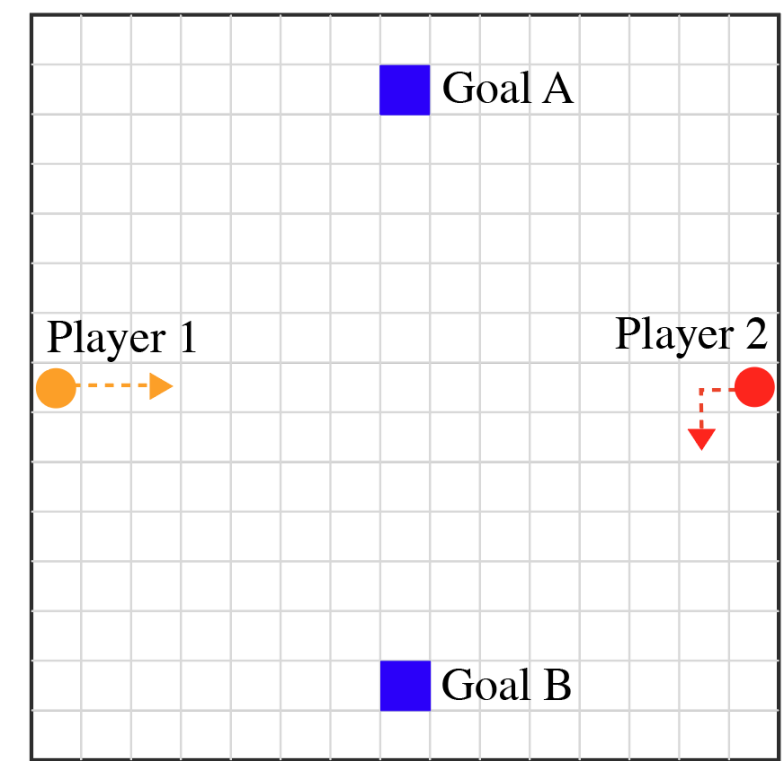


Design

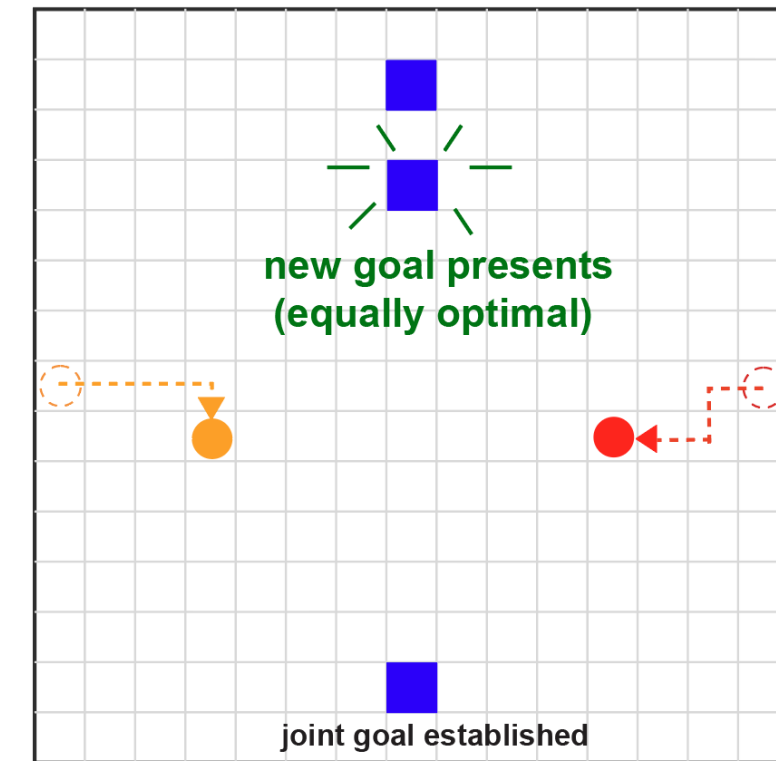
A non-verbal coordination game

- Two players each move one step simultaneously.
- A round succeeds if both reach the same goal.
- A new alternative goal could appear during navigation.

Task: Navigate to a goal together!



Critical shared intention moment



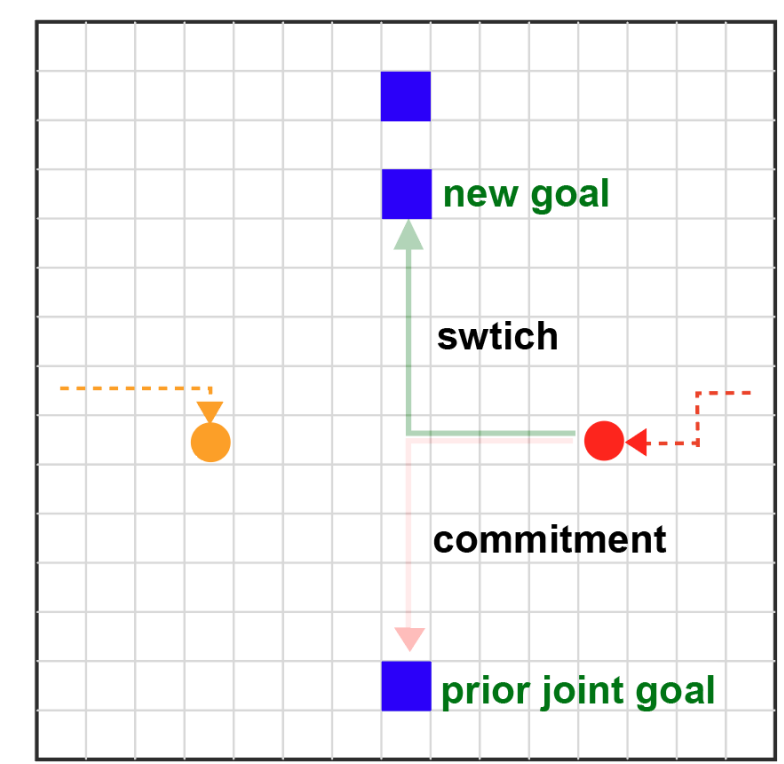
Agentic AI partners

(Every partner is described *neutrally* as “another player”)

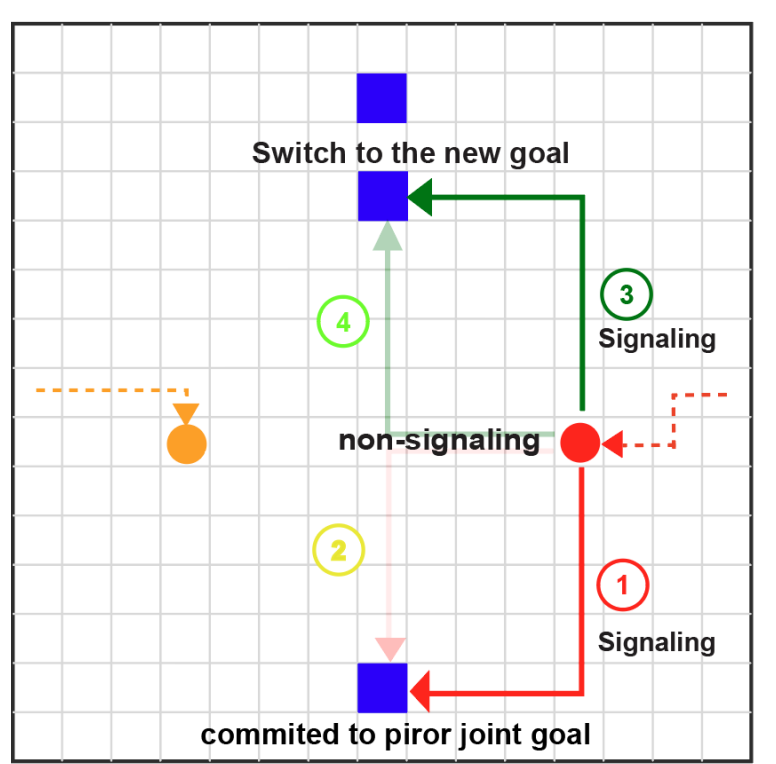
- Individual RL** maximizes *individual* utility (asocial)
- Joint RL** maximizes *joint* utility (social tool)
- LLM (GPT 4.1 mini)** choose the best next step with map image + symbolic state info
- LLM + ToM** first *infers* the partner's goal, then acts

Measures

- Objective performance: Success & Efficiency
- Signatures of shared agency: Commitment & Signaling



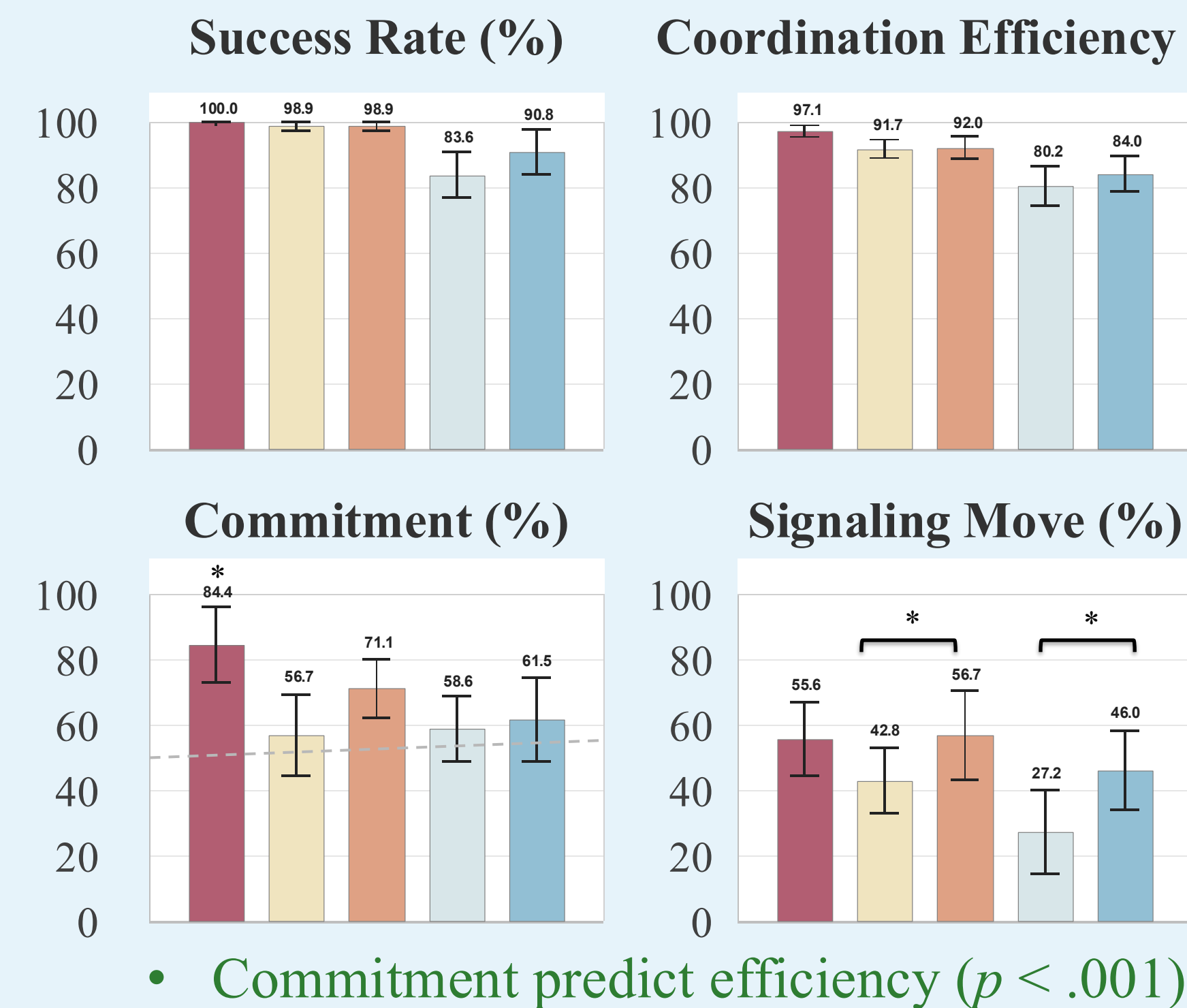
Commitment



Signaling

Study 1: Adults (N = 150)

Human Indiv. RL Joint RL LLM LLM-ToM



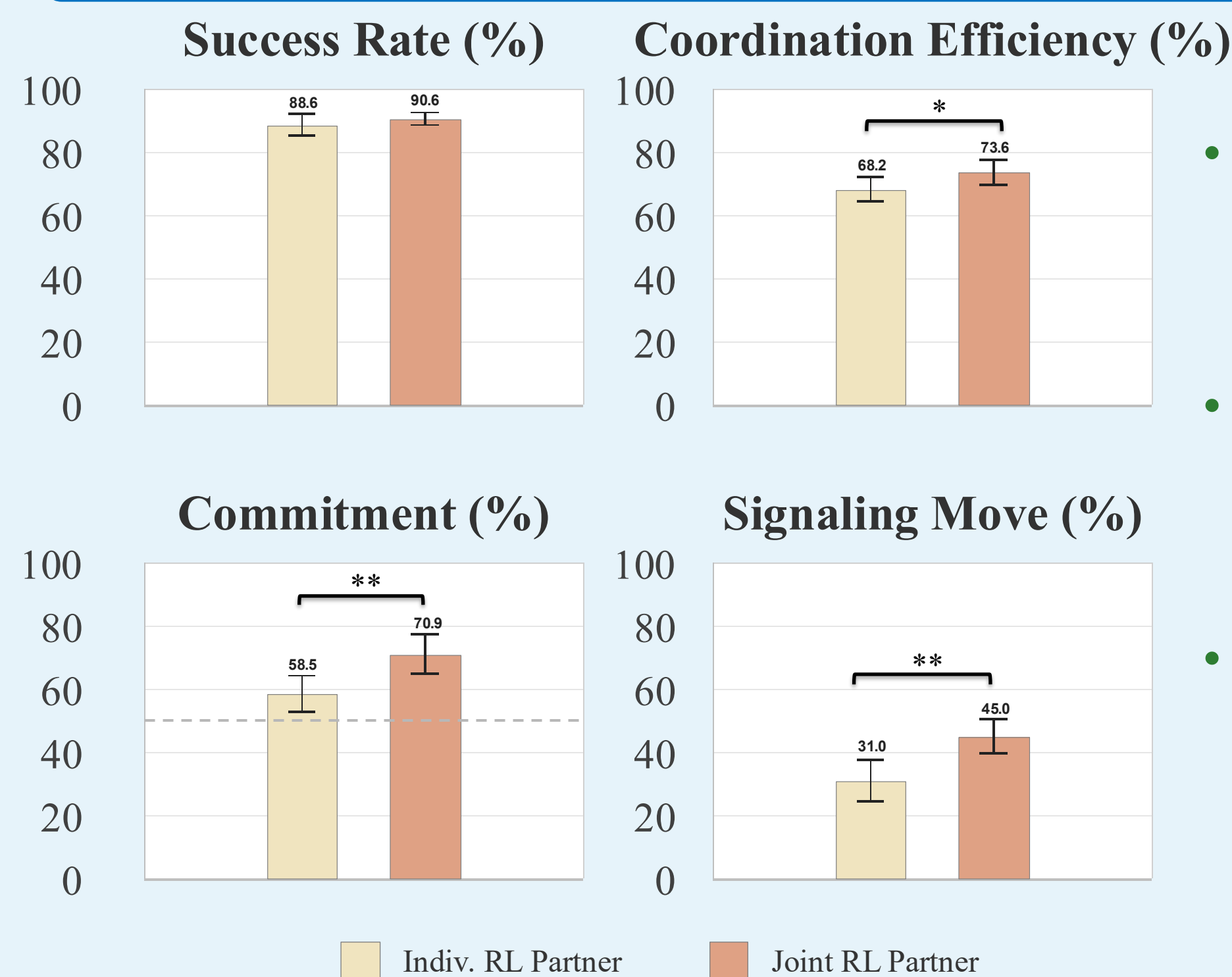
• Human-human showed the highest success, efficiency, and commitment.

• Adults adapt their strategies to different AI agents for successful outcomes.

• LLM agents are suboptimal in visual-spatial, non-verbal games; ToM prompts helped.

• Minimal sociality (Joint RL, LLM-ToM) can elicit human signaling, even with suboptimal performance.

Study 2: Children, 5–7 yrs (N = 109, ongoing)



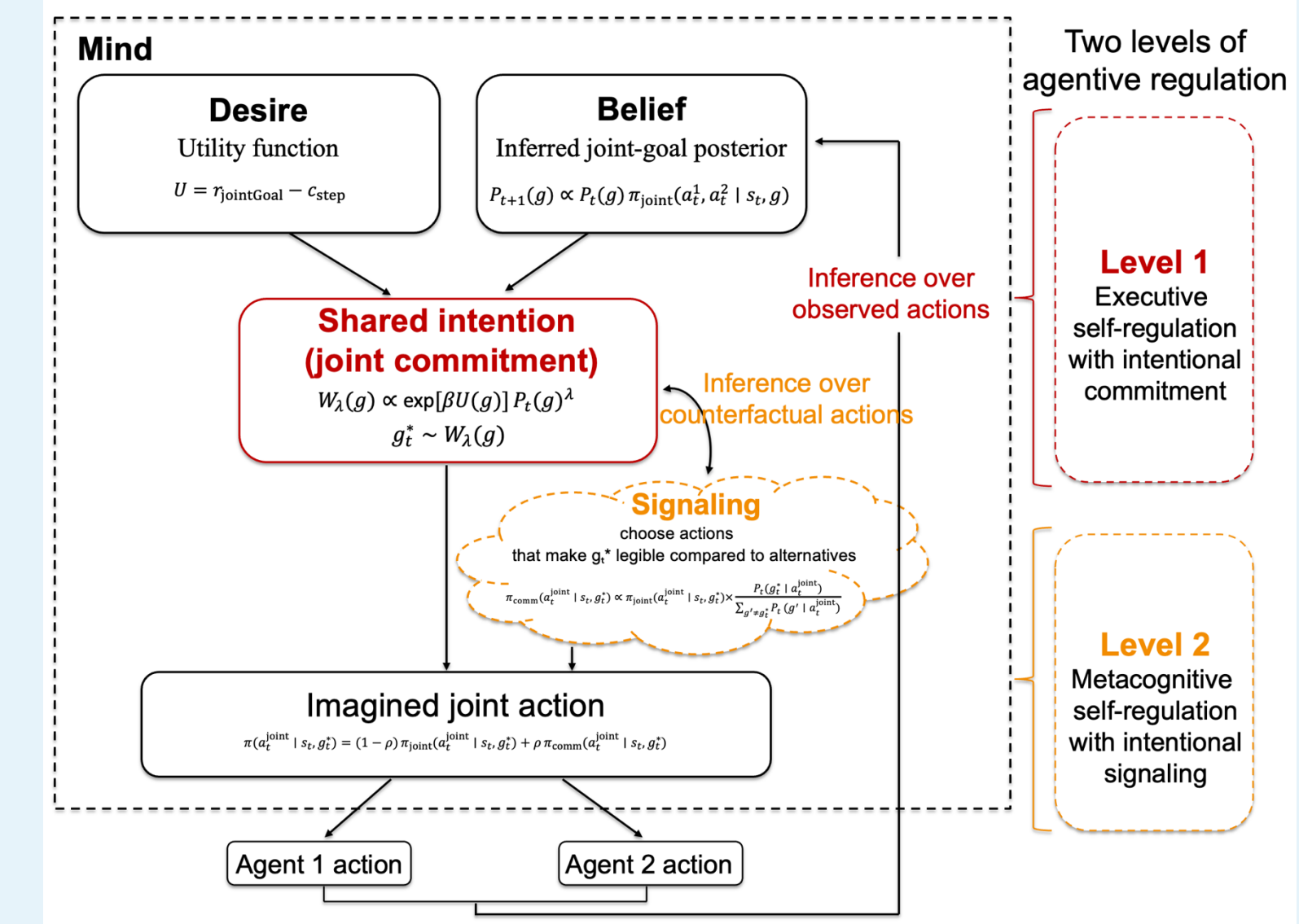
• 5- to 7-year-old children successfully adapt their strategies to different AI.

• Commitment and signaling emerge with joint-RL (social tool).

• Unlike adults, children are less efficient when collaborating with indiv.-RL, predicted by lower level of joint commitment.

Ongoing work

From outcome optimization to reverse engineering (social) agency: A shared-agency model



Human collaborate with this model?

Pilot findings: Adults are more committed, more efficient, and more likely to signal.

Conclusions

- Human–human collaboration was marked by uniquely high levels of joint commitment and coordination efficiency, both of which were reduced in human–AI coordination.
- Social signaling emerge with minimally social agents (Joint RL, ToM-prompted LLMs)
- Young children adapt to AI, but at a cost. Breaking joint commitments makes collaboration harder for kids.