

Effects of Interoceptive Inference in Mixed-Motive Human-Robot Teams

Proceedings of the Human Factors and
Ergonomics Society Annual Meeting
1–3

© The Author(s) 2026



Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/10711813261475208
journals.sagepub.com/home/pro



Trevor Patten¹  and Mengyao Li¹

Abstract

Humans are biased against cooperating with robots and machines under mix-motives. Current interventions rely heavily on exteroceptive cues from robots to provide social cues to human teammates. While exteroceptive cues are known to be effective, no prior evidence describes how AI or robot teammates can support interoceptive inference in humans. This study aimed to understand the role of interoceptive inference with robot teammates. Findings suggested that symbolic representations of internal states may be insufficient to reduce exploitation bias toward artificial agents. More broadly, the findings differentiate framing from embodied interoceptive social cues.

Keywords

theory of mind, HRI, decision-making, interoception

Introduction

Robots range in application, from social companions to coworkers. Smooth interactions with humans are challenging because they are reciprocal, with humans and robots evaluating each other's actions and adjusting to task demands. Humans evaluate each other by reading social cues from their partner to anticipate and account for the actions of their teammate (Ondobaka et al., 2017). This ability to anticipate what another is thinking is Theory of Mind (ToM), but it may be under-discussed in HRI. Social signals, like speech or gestures, improve teaming with robots in tasking where individual goals are just as valuable as team goals (de Melo & Terada, 2019). These signals allow humans to refine their ToM for their robot teammate; however, social cues represent one method of improving the ToM. Inferred visceral activity refines the ToM between humans (Ondobaka et al., 2017), but robots do not support this inference. This study aimed to evaluate methods of supporting human interoceptive inference from robot teammates.

Background

Mixed-motives are tasks where individual gain is at odds with collective gain. Individuals who cooperate take on the risk that their teammates may not be cooperative, but they stand to benefit the team should other teammates also cooperate. Individuals who are competitive choose not to risk being taken advantage of for the chance of taking resources that could have gone to other teammates.

Humans have a competition bias toward machines under mixed-motives (Ishowo-Oloko et al., 2019). This bias is minimized through the belief that the machine is a person, or with social cues. Social cues, like facial expressions, improve cooperative behavior under mixed-motive (de Melo & Terada, 2019). The interpretation of social information can be discussed through the lens of the ToM.

In humans, the ToM is the ability to understand how others think and feel. The ToM is more than just reading social cues. The ToM is built on exteroceptive, proprioceptive, and interoceptive information (Ondobaka et al., 2017). Humans utilize all three modes to construct the ToM. Social cues like utterances and facial expressions are exteroceptive, but interoceptive inference is not acknowledged within HRI. Interoception is unique because it relies on inferring the internal activity of another. While social cues from exteroceptive and proprioceptive cues can be adapted to HRI through utterances and gestures, robots do not possess viscera for humans to infer about, so interoceptive inference is not supported.

Developing an accurate ToM contributes to cooperation. In human interactions, those with lower ToM ability face social issues and struggle to navigate complex social scenarios (Hochhauser et al., 2015). Better ToM ability is also

¹Georgia Institute of Technology, Atlanta, USA

Corresponding Author:

Trevor Patten, Georgia Institute of Technology, Atlanta, GA 30332, USA.
Email: tpatten7@gatech.edu

Table 1. Iterated Prisoner's Dilemma Payoff Matrix.

		Robot	
		Cooperate	Defect
Human	Cooperate	6,6	0,10
	Defect	10,0	4,4

linked with an increased propensity to compete with others (Ridinger & McBride, 2025). One likely reason for this is that low ToM people cannot evaluate if others can be trusted; thus, cooperation is risky (Ridinger & McBride, 2025). Building accurate ToM in humans with robot teammates may be supported by supporting interoceptive inference in HRI.

It is unclear if supporting interoceptive inference for humans improve cooperation, as both belief and social cues alter cooperation behavior. This study evaluated the belief in interoceptive ability and the presence of social cues. We hypothesized that those who believed in the interoceptive ability of a robot teammate would also cooperate more than those who did not, and social cues would increase cooperation with robot teammates.

Approach

This study used a 2 (interoceptive belief, no belief) by 2 (social cue, no cue) between-subjects design. A total of 74 participants were recruited from a university population (46 female, $M_{age} = 19$, $SD = 1.95$). The first manipulation altered participants' beliefs about the robot teammate's decision-making. Half of the participants were presented with a video introduction describing how the robot teammate used processes similar to how human biology informs human decision-making. A manipulation check ensured that beliefs were influenced by the video. Half of the participants viewed the robot teammate's reported valence and arousal during the task, while the other half did not. This functioned as a low-resolution cue of internal activity akin to sweat or a wavering voice in humans (Ondobaka et al., 2017).

Participants played 15 rounds of an iterated prisoner's dilemma with a robot teammate. For each round, participants indicated if they wanted to cooperate or defect. The points awarded after each round were determined by the human and the robot's decisions indicated in Table 1. The robot's behavior was predetermined, as was its valence and arousal. Rounds 1 to 5 used a fixed pattern as a mask for the tit-for-tat strategy used in the remaining 10 rounds (Hsieh & Cross, 2022). Participants in the social cue cells could view the robot's response on a valence/arousal scale (Bradley & Lang, 1994).

Outcome

A *t*-test confirmed that belief in the robot teammate's interoceptive ability was successful ($t(72) = 5.1$, $p < .001$). The final model was run with 60 participants, 15 per cell, due to data loss from network outages. A linear mixed model regressed the participant's choice to cooperate or defect over the belief and cue independent variables, with participants as the random effect.

The model did not return significant results of belief in the robot teammates decision-making ability ($\beta = -.25$, $\chi^2 = .04$, $p = .84$, CI [0.91, 0.74]), cues of interoceptive activity ($\beta = -.13$, $\chi^2 = .10$, $p = .76$, CI [-0.95, 0.67]), nor their interaction $\beta = .31$, $\chi^2 = .27$, $p = .60$, CI [-0.86, 1.47]).

Conclusion

This study presented an application of a theoretical model of the ToM as it relates to building better robot teammates. Particular emphasis was placed on the supporting human interoceptive inference because it is an unused part of the ToM that generally does not apply to robot teams. This study did not find conclusive results to suggest supporting interoceptive inference through belief, nor did interoceptive cues improve cooperation relative to control. Null results could be caused by the interoceptive cue being too weak to generate an effect relative to conventional social cues. Despite the success of the manipulation check, belief may have suffered due to the physical presence of the robot interfering with the belief in the interoceptive quality of the robot.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

ORCID iD

Trevor Patten  <https://orcid.org/0009-0008-3581-9434>

References

- Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25(1), 49–59. [https://doi.org/10.1016/0005-7916\(94\)90063-9](https://doi.org/10.1016/0005-7916(94)90063-9)
- de Melo, C. M., & Terada, K. (2019). Cooperation with autonomous machines through culture and emotion. *PLOS ONE*, 14(11), Article e0224758. <https://doi.org/10.1371/journal.pone.0224758>
- Hochhauser, M., Weiss, P. L., & Gal, E. (2015). Negotiation strategies of adolescents with high-functioning autism spec-

- trum disorder during social conflicts. *Research in Autism Spectrum Disorders*, 10, 7–14. <https://doi.org/10.1016/j.rasd.2014.10.022>
- Hsieh, T.-Y., & Cross, E. (2022). People's dispositional cooperative tendencies towards robots are unaffected by robots' negative emotional displays in prisoner's dilemma games. *Cognition and Emotion*, 36(5), 995–1019. <https://doi.org/10.1080/02699931.2022.2054781>
- Ishowo-Oloko, F., Bonnefon, J.-F., Soroye, Z., Crandall, J., Rahwan, I., & Rahwan, T. (2019). Behavioural evidence for a transparency–efficiency tradeoff in human–machine cooperation. *Nature Machine Intelligence*, 1(11), 517–521. <https://doi.org/10.1038/s42256-019-0113-5>
- Ondobaka, S., Kilner, J., & Friston, K. (2017). The role of interoceptive inference in theory of mind. *Brain and Cognition*, 112, 64–68. <https://doi.org/10.1016/j.bandc.2015.08.002>
- Ridinger, G., & McBride, M. (2025). Theory-of-mind ability and cooperation. *Journal of Economic Behavior & Organization*, 230, Article 106895. <https://doi.org/10.1016/j.jebo.2025.106895>