Can't Do or Won't Do? Social Attributions in Human–Agent Cooperation

(Extended Abstract)

Philipp Kulms
Bielefeld University (Germany)
pkulms@techfak.uni-bielefeld.de

Nikita Mattar
Bielefeld University (Germany)
nmattar@techfak.uni-bielefeld.de

Stefan Kopp
Bielefeld University (Germany)
skopp@techfak.uni-bielefeld.de

ABSTRACT

We highlight how human–agent cooperation is linked to the attribution of critical qualities like trustworthiness and competence to the cooperation partner. To investigate these aspects in a systematic fashion, we devise a novel paradigm of an interactive cooperation game that goes beyond commonly adopted economic scenarios. Our results indicate how a less skillful agent that shows little appreciation for its human partner's suggestions regarding the next action is ascribed negative trustworthiness, but also positive competence. This suggests that perceived competence and cooperativeness interact with each other in important ways that can be studied jointly in an interactive cooperation game scenario.

1. INTRODUCTION

Agents are social actors and are hence perceived in ways similar to how we perceive humans. In human–human interaction and cooperation, humans interpret the behavior of others in terms of two underlying universal dimensions of social cognition, namely, warmth and competence [1]. We hypothesize that these mechanisms transfer to human–agent cooperation. That is, we assume that artificial agents engaged in cooperative interaction with human partners elicit the perception and attribution of qualities similar to warmth and competence, and this assessment is important for future cooperative behavior.

Researchers in HCI, robotics and artificial intelligence have been investigating the social dimensions of human–agent interaction. Considering the effects of even minimal social cues [6], it is clear that addressing only task-related aspects is insufficient to understand human–agent cooperation. Agents must also be willing and trusted to act toward the joint cooperative goal. This adds an important social dimension on which agents need to assess and affect their partners. To this date, there is no framework that describes how complex goal-directed behavior of two agents and social outcome variables evolve and interact over time. Standard cooperative games as used in behavioral game theory are limited when it comes to studying qualities like warmth and competence and how they evolve, since problem solving is translated to the decision between cooperation and non-cooperation based on reasoning about the other agent's corresponding intention. The possibility to include and manipulate qualities like task competence or the willingness to cooperate is very limited in such models. In this paper, we present work towards a more comprehensive account of the relevant *perceived* qualities of intelligent agents in human–agent cooperation. Based on related work in human–human and human–agent interaction, we hypothesize that the attribution of competence, trustworthiness and cooperativeness are important in human–agent cooperation. For example, seemingly non-cooperative behavior can be due to a lack of cooperativeness and/or a lack of competence. Likewise, competent behavior is necessary for attaining goals but in itself is not sufficient in cooperative settings, in which decisions about whom to entrust a certain, potentially crucial task are needed. We hence take trustworthiness to be an important and relevant sub-concept of warmth, and we aim to study how perceived competence and trustworthiness are interlinked with each other, and how they are differentially related to cooperativeness. Our results demonstrate that both the quality of task-related action as well as (non-)compliance with recommendations of the partner affect how humans perceive *both* competence and trustworthiness of an autonomous agent, and this differentially correlates with the attribution of cooperativeness.

2. METHOD

We devise an interaction framework in which we manipulate and analyze key characteristics of cooperative behavior systematically. We focus on how an autonomous agent in a cooperative interaction is perceived in terms of its competence and trustworthiness, and how this is related to attributed cooperativeness. The interaction paradigm is a turn-based scenario in which two partners solve a puzzle cooperatively. The setting involves two players working together to place two kinds of Tetris-like blocks in a manner that minimizes required space. The game allows us to easily manipulate the agent's task-related behavior and thus its perceived competence. Moreover, we can induce both individual (as individual payoff for a block) and cooperative goals (as joint payoff added to the individual payoff for completed rows). Two different player roles are introduced: a human block recommender and an agent decider (see [3] for more details). The recommender suggests the next block for the agent to choose. The agent then decides whether it accepts the offer or picks the other block. Importantly, in each round, both blocks are drawn by the players from an urn without replacement.

Appears in: *Proceedings of the 15th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2016), J. Thangarajah, K. Tuyls, C. Jonker, S. Marsella (eds.), May 9–13, 2016, Singapore.*

Both players receive points individually for every block they place. There are two kinds of blocks: a T-shaped block and a U-shaped block that is harder to place. The U-block yields twice as many points than the T-block, leading to an individual benefit in placing more U-blocks. In addition, both players obtain bonus points when a certain amount of rows are completed horizontally. We refer to this as cooperative or joint goal, since it can only be achieved if the players act cooperatively by recommending and choosing blocks that maximize the number of completed rows and by placing the blocks well. The interactive cooperation game paradigm was designed with the goal in mind to enable the study of *perceived* competence, trustworthiness, and cooperation. These factors are operationalized as follows:

Cooperation: The two players attain the cooperative goal by alternately placing blocks in a coordinated efficient manner. Additionally, the agent accepts or rejects the human's suggestion for its next block. Cooperative actions are thus based on the context in which they occur, there is no action that is universally cooperative. Participants are asked to assess agent cooperativeness.

Competence: An agent is competent if it places blocks skillfully, that is, in a way that maximizes the likelihood of completed rows. We manipulate this skill using two different heuristics and assess the extent to which participants think the agent is competent.

Trustworthiness: An agent that is perceived as socially warm is perceived as friendly, helpful, having positive intentions and, as a recurrent factor in the literature, trustworthy [1]. We assume that trustworthiness more clearly pertains to the perception of cooperative artifacts than the broad warmth concept and use this sub-term instead. Trustworthiness, then, is an agent's ability to signal integrity, benevolence, and ability (i.e., competence) [5]. Participants are asked to assess agent trustworthiness.

3. RESULTS AND CONCLUSION

We conducted an exploratory study to answer the following RQ: How does the behavior of an agent player influence its perceived trustworthiness and competence, and how are trustworthiness and competence related to perceived cooperativeness? We investigated how human recommenders judge decider agents with varying degrees of task skill ("I am able") and compliance ("I adopt your strategy") regarding the human's suggestion. The study had a 2 (Heuristic: maximize occupied space by blocks [MS] vs. random next move [R]) X 2 (Compliance: high [HC] vs. low [LC]) between-subjects design. Seventy-seven participants took part in the experiment.

A number of significant differences emerged. First, the agent was more competent given MS vs. R, and given LC vs. HC. Second, the agent was more trustworthy given HC vs. LC, and given MS vs. R. Third, during R conditions, competence was judged significantly higher for LC vs. HC. Fourth, cooperativeness correlated strongly with trustworthiness and moderately with competence (competence and trustworthiness were uncorrelated). These patterns converge to an overall importance of skill and compliance in our game, each influencing both perceived competence and trustworthiness, as well as a modulation effect of compliance on competence such that unskilled behavior was perceived as competent.

In conclusion, the interactive cooperation game paradigm allows us to frame the perceived qualities of trustworthiness, competence, and cooperativeness into a picture that is grounded in social cognition research and theorizing in human–centered HCI. Our results indicate that competence alone does not ensure an overall favorable assessment of an agent, although competence was a necessary condition of goal attainment. Perceived trustworthiness and competence – attributions that are crucial for social interactions – were affected both by an agent's compliance and competence. This shows that in cooperation, humans are very sensitive to the agent's cooperativeness in light of its competence, and vice versa. The unskilled agent, when it only chose the high-value option, was rated nearly as competent as the actual competent agents, although it played much worse in comparison. Participants apparently assumed intentionality behind this behavior [4] and felt that the agent is able to enact its goals. Competence is self-profitable [7], the agent was thus perceived as competent enough to act selfishly. Our next steps further address the issue of how intelligent agents should be designed for cooperative settings. Given conflicting human–agent goals, agents need to consider the trade-off between individual and joint payoff [2]. While the former benefits from competence, the latter builds on trustworthiness. We attempt to interrelate the behaviors and judgments occurring in our game with established measures on the cooperation spectrum and investigate how social cues facilitate trustworthiness in the game.

4. ACKNOWLEDGEMENTS

This research was supported by the German Federal Ministry of Education and Research (BMBF) within the Leading-Edge Cluster 'it's OWL', managed by the Project Management Agency Karlsruhe (PTKA), as well as by the Deutsche Forschungsgemeinschaft (DFG) within the Center of Excellence 277 'Cognitive Interaction Technology' (CITEC).

REFERENCES

[1] S. T. Fiske, A. J. C. Cuddy, and P. Glick. Universal dimensions of social cognition: warmth and competence. *Trends in Cognitive Sciences*, 11(2):77–83, 2007.

[2] G. Klein, D. Woods, J. Bradshaw, R. Hoffman, and P. Feltovich. Ten challenges for making automation a "team player" in joint human-agent activity. *IEEE Intelligent Systems*, 19(06):91–95, 2004.

[3] P. Kulms, N. Mattar, and S. Kopp. An interaction game framework for the investigation of human–agent cooperation. In W.-P. Brinkman, J. Broekens, and D. Heylen, editors, *Intelligent Virtual Agents*, volume 9238, pages 399–402. Springer, Berlin, Heidelberg, 2015.

[4] B. F. Malle and J. Knobe. The folk concept of intentionality. *Journal of Experimental Social Psychology*, 33(2):101–121, 1997.

[5] R. C. Mayer, J. H. Davis, and F. D. Schoorman. An integrative model of organizational trust. *The Academy of Management Review*, 20(3):709–734, 1995.

[6] C. Nass and Y. Moon. Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, 56(1):81–103, 2000.

[7] G. Peeters. Relational and informational patterns in social cognition. In W. Doise and S. Moscovici, editors, *Current Issues in European Social Psychology*, pages 201–237. Cambridge University Press, Cambridge, 1983.