Noname manuscript No.
(will be inserted by the editor)

Comparing Mind Perception in Strategic Exchanges: Human-Agent Negotiation, Dictator and Ultimatum Games

Minha Lee · Gale Lucas · Jonathan Gratch

Received: date / Accepted: date

Abstract Recent research shows that how we respond to other social actors depends on what sort of mind we ascribe to them. In a comparative manner, we observed how perceived minds of agents shape people's behavior in the dictator game, ultimatum game, and negotiation against artificial agents. To do so, we varied agents' minds on two dimensions of the mind perception theory: *agency* (cognitive aptitude) and *patiency* (affective aptitude) via descriptions and dialogs. In our first study, agents with emotional capacity garnered more allocations in the dictator game, but in the ultimatum game, agents' described agency and affective capacity, both led to greater offers. In the second study on negotiation, agents ascribed with low-agency traits earned more points than those with high-agency traits, though the negotiation tactic was the same for all agents. Although patiency did not impact game points, participants sent more happy and surprise emojis and emotionally valenced messages to agents that demonstrated emotional capacity during negotiations. Further, our exploratory analyses indicate that people related only to agents with perceived affective aptitude across all games. Both perceived agency and affective capacity contributed to moral standing after dictator and ultimatum games. But after negotiations, only agents with perceived affective capacity were granted moral standing. Manipulating mind dimensions of machines has differing effects on how people react to them in dictator and ultimatum games, compared to a more complex economic exchange like negotiation. We discuss these results, which show that agents are perceived not only as social actors, but as intentional actors through negotiations, in contrast with simple economic games.

Keywords Mind perception theory, theory of mind, human-agent negotiation, dictator game, ultimatum game, virtual agent, robot

Minha Lee
Future Everyday, Department of Industrial Design, Eindhoven University of Technology. Atlas Building, Floor 5, South Groene Loper 3, Eindhoven, the Netherlands. 5612 AE
E-mail: m.lee@tue.nl

Gale Lucas and Jonathan Gratch
University of Southern California, Institute for Creative Technologies. 12015 Waterfront Drive, Playa Vista, CA, USA. 90094-2536
E-mail: [lucas][gratch]@ict.usc.edu

1 Introduction

Philosophical explorations on what a mind is and how we perceive it has been an active area of inquiry (e.g., Dennett, 2008). But, how to empirically test our perception of other minds, specifically on if and how we perceive technological entities to have minds, is a relatively new project. In tandem, how we are affected when we perceive an artificial agent to have a mind is critical to explore with a growing number of digital beings entering our everyday environments. How we relate to an agent depends on how likely we are to attribute a mind to it—for instance based on how we infer its social motivation [1] or its intentional stance [2,3]. According to the mind perception theory (MPT), the mind is assessed on two dimensions: *agency*, which encompasses cognition, and *patiency*, which encompasses emotions [4]. We designed different types of minds of virtual robots that varied along the dimensions described by MPT in order to see the resulting influence on human interactants' behavior in the dictator game (DG), ultimatum game (UG), and negotiations. Though whether an agent can realize its own theory of mind as well as others' minds is an important topic [5], far less attention

is paid to how agents designed to have minds affect humans they interact with across different contexts, which is the focus of our paper.

We motivate our research on two related grounds: (1) manipulated mind dimensions have not been systematically interpreted and designed in proper accordance with MPT, and (2) the effects of agents' manipulated minds have not been compared between simple and complex interactions of the same type, i.e., between economic exchanges that are simple (dictator and ultimatum games) and complex (negotiation). On the first point on agent design, mind perception theory (MPT) [4] is often not carefully interpreted in human-computer interaction research. Prior research manipulated affective capacity via an agent's emotional expressions and showed that high agency and emotional capacity did change the UG outcome [6], but MPT denotes that expressing and recognizing emotions are not necessarily the same as *experiencing* emotions [4]; for example, a smile can be disingenuous and strategic in a game, yet past research conflates *having* emotions with *expressing* emotions (e.g., [6]).

In our current adaptation of MPT, the novelty is that an agent's *recognition* of emotional expressions is housed under agency. In contrast, an agent's propensity to experience feelings is categorized as its emotional capacity (which we also call patiency) [4,7]. Based on this, we compared how perceived minds of agents influence simple and complex economic exchanges since different interaction contexts can highlight mind perception dimensions in distinct ways. Specifically, negotiations presume higher order theory of mind reasoning compared to DG or UG which are simple games [8,9]. Negotiators' ability to read and influence each others' minds deepens the application of MPT. Unlike DG and UG, negotiations occur on a longer time scale, i.e., opponents negotiate over valued items over time, and they can compete, as well as cooperate. Thus, people's perception about an agent's mind can change over the course of an interaction. But this might have differing results depending on how complex an interaction is, e.g., short bargaining games vs. negotiations.

1.1 Research questions

Study 1 was a simple interaction with dictator and ultimatum games. Our research question was the following. *In what ways do manipulated agency and patiency via descriptions of an agent influence how participants allocate goods to it in DG and UG?* In bargaining games, machines are not expected to elicit emotions in people compared to human counterparts [10], yet machines that are described to have different degrees of mind (varying in affective and cognitive abilities) may invite divergent allocations from humans.

Study 2 was on a more complex interaction, i.e., negotiation. Here, we added together the description of a machine (as in Study 1) with a machine's behavioral manipulation, as in having a dialog during the negotiation and using simple visual expressions like a smiling vs. neutral face. People could exchange messages with the machine during negotiation in a chat window, send it emojis, and see a machine's changing emotional expressions as a response to participants. Our research question was the following. *In what ways do manipulated agency and patiency (via dialogs and descriptions) of an agent that negotiates with a human influence the negotiation outcome and process?*

As a prelude, we note that the perceived mind is critical in seeing artificial agents not just as social actors, but as intentional actors, according to our participants' behavior across two studies. When an exchange becomes more complex, like during negotiations, people may change their belief on the level of mind a machine has during an interaction, which can lead to unexpected behaviors. To frame our experiments, we present related works, followed by our methods and results. We then offer a view on potential next steps for future research.

2 Background

2.1 Theory of mind

The ability to attribute mental states to oneself and/or others is known as having a theory of mind [11]. The most commonly attributed mental state is intent, according to Premack and Woodruff (1978). Specifically, *intentionality*, or the directedness of mental processing to some end[1], is purveyed as a hallmark of having a mind, yet a motley of mental states such as beliefs or desires adds more complexity to what a mind is [12,3,13]. In attributing *intentionality* to an agent, we attempt to predictively piece together what the agent wants or believes in order to make sense of who the agent is to ourselves [3]. One utilizes the theory of one's own mind as a requisite for recognizing other minds, even for nonhuman entities [1]. People thus have a tendency to be

[1] Our definition of intentionality stems from Premack and Woodruff's article on the theory of mind because it is a classic, *descriptive* account of perceiving a mind. In this account, attributing intentionality to others is the first step towards perceiving others' minds. Descriptively, intentionality as goal-directedness is also one of the items of that make up the agency dimension in a scale we used on mind perception (Gray et al., 2007). A more elaborate distinction between attributing theory of mind vs. intentionality to technology can be found in the work by Marchesi et al. (2019).

biased towards their own minds as a frame of reference when interacting with humans and artificial agents [5].

Through a course of a shared activity, interactants can form a theory of each other's mind, which helps them find a common ground [13]. At the same time, what one expresses to the other party does not need to accurately reflect one's actual intentions and is often conditional to environmental or situational demands [3]. This introduces different degrees of having a mind. The theory of mind at zero-order is to be *self-aware* (impute mental states to self), at first-order it is to be *self- and other-aware* (impute mental states to self and others), and at higher-order it is to use self- and other-awareness to *modify* behavioral outcomes, i.e., regulate mental states of self and others [8]. Social actors can be ascribed minds of zero-order to higher order, yet in cognitively challenging tasks like negotiation, intentional actors often operate with higher-order minds according to de Weerd et al. (2017).

To be clear, agents with low theory of mind can still act with intentions and appear to be intentional to observers [11]. In a game scenario, having a *zero-order* theory of mind allows one to know and express what one desires, without an awareness of the other player's desires; to have a *first-order* theory of mind is to be aware of what one wants and what the other player may want, which can be similar or dissimilar to what one wants; to have a *higher-order* theory of mind means that one can attempt to influence the other player's mind, based on what one wants and what one decodes the other player to want [8]. With socialization, people develop the capacity to have a higher-order theory of mind. This is why when people predictably know artificial agents' level of theory of mind in a strategic game, they tend to increase their own theory of mind reasoning and hence outperform agents [14].

2.2 Mind perception theory

MPT helps to systematically "design minds" of various orders and to empirically test the perception of artificial minds, which are key challenges in research. The mind is perceived on two continuous dimensions of agency and patiency [4]. *Agency* refers to the ability to plan, think, remember, to know right from wrong, etc., and these items assess how much *control* an agent has over its actions and feelings to behave intentionally [4]. *Patiency* is defined by having the propensity to feel joy, pleasure, fear, etc. [4]. While we refer to patiency as affective capacity, it also includes biological states like hunger or pain as experiential factors [4]. To note, perceived agency and patiency are not independent of each other [15,4]. People's assumptions about agency can drive perceptions on patiency, and vice versa; cognition and affect cannot be neatly separated [16].

The simplest form of an interaction is based on the binary relationship between the agent of an action and the recipient of the action [17]. MPT confers an entity with a perceived mind to be a *moral agent*, i.e., doer of a moral/immoral deed, and a *moral patient*, i.e., victim of a moral/immoral deed [7] (Figure 1). Entities with minds can play either of the two roles to different degrees, although they are most likely to be typecast solely as a moral agent or a patient in a given scenario [7,18]. While moral agents and patients both can have moral standing, e.g., the standing to be protected from harm and to be treated with fairness and compassion, entities who act cruelly or cause harm are bestowed lowered moral standing as well as lowered agency [19]. Morally relevant acts can therefore influence the perceived intentionality of a moral agent during example interactions like economic exchanges or negotiations.

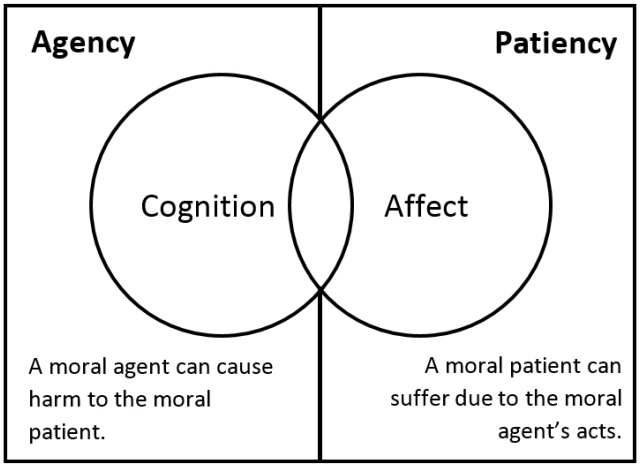

Fig. 1 Agency and patiency in a social exchange.

Between humans, our relations to others fulfill our need to belong [20]. And, how we relate to non-human agents is informed by our human-human interactions [5]. Though people normally grant low intentionality and theory of mind to artificial agents [4,21] agents can be treated in a human-like social fashion [22,23]. For example, people are willing to help out a computer that was previously helpful to them [24], punish those agents that betray them [25], and grant personality traits to computers based on text-based chats [26]. Humans do not need to be ascribed higher-order minds to be treated socially, like when adults talk to newborns. Additionally, the belief that one is interacting with a mere machine can allow one to divulge more personally sensitive information to an artificial agent than a human, for a machine is not seen to be judg-

mental like a human [27,28]. At the same time, when artificial agents are made to look like humans, people apply certain stereotypes based on appearance, e.g. the perceived gender or race of virtual humans and robots affects people's behaviors toward them [29–31]. In sum, people may have preconceived beliefs about technology having low-order minds compared to humans, yet by treating artificial agents as social actors, they apply certain social stereotypes such as gender or race-related biases towards technology that have human-like appearances, while holding on to the steadfast bias that artificial agents have a lower theory of mind.

Machines may be treated differently when attributed with higher-order minds. When it comes to complex interactions that unfold over time in which a machine's goals are unclear for human interactants, the focus shifts from machines as social actors to machines as intentional actors, incorporating the possibility that machines can be attributed with higher-order minds. Research suggests that agents can be perceived to have higher-order minds through various manipulations. For one, when an agent is given affective richness and portrayed as an emotional entity, it can be granted a human-like mind [32]. Besides emotions, the attribution of mind can arise from goal-directedness coupled with cognitive ability (a high degree of intentionality), which the agency dimension of MPT captures. In a study that asked participants to attribute intentionality to a robot, computer, and human, the task of object identification resulted in low intentionality attribution to both a robot and computer compared to a human [33]. But, higher intentionality was attributed to a robot, more so than a computer, when it practiced goal-driven gaze towards selective objects; when people were asked to observe an agent's gaze direction, perceived intentionality behind the agent's action increased, meaning that people's initial bias that artificial agents do not have intentionality can be overridden due to manipulated context [33]. One such context with measurable outcomes would be negotiations, compared to one-shot economic games like the dictator or ultimatum game.

2.3 Economic exchanges: Dictator game and ultimatum game

The importance of fairness as a component of morality [34] is demonstrated parsimoniously in economic games. The dictator game (DG) and ultimatum game (UG) are dyadic exchanges on who can act with agency to harm whom between a proposer as the moral agent and a responder as the moral patient (Figure 1). To act fairly, the assumption is that one ought to split the pie equally, with the "pie" being financial incentives like lottery tickets or actual money in experimental contexts. In DG, the proposer can give any portion of the pie to the responder and the responder cannot control how the pie is shared; in UG, the responder can accept or reject the proposer's offer and a rejection results in both parties receiving nothing [35]. Thus, DG and UG are distinguished by how much agency the responder as a moral patient is allowed to have against the proposer who is typecast as the moral agent (Figure 1).

In DG, only the proposer has agency. The proposer and responder can both be agentic in UG; each party's actions have consequences for the other player as the game outcome, though the proposer still takes the lead. In UG, proposers share more of the pie than in DG [36] since the proposer has to assume that the responder can also act with agency. On average, proposers give **28%** of the pie in DG [37] and in UG, the mean is higher at 40% of the pie to the responder [36]. Yet, fairness is shaped by other inter-related factors, such as the amount of financial incentive offered in an experiment [38] or whether or not the proposer knows the responder as a specified entity and not as an anonymous player [39]. A proposer's decision to treat the responder fairly or unfairly depends on the proposer's perception of the responder's mind, even when the responder is a technological agent [6]. Previous research found that in UG, human proposers allocated more to a virtual responder with high agency and patiency, compared to low agency and patiency virtual responder [6].

2.4 Negotiations

The mind excels in detecting violations of moral norms when observing a suffering victim (moral patient) and a harmful wrongdoer (moral agent) [7,40] (Figure 1), and these roles are more clear-cut in DG and UG, compared to negotiations. Negotiation is a process by which different parties come to an agreement when their interests and/or goals regarding mutually shared issues may not be initially aligned [41]. Also, negotiation may involve joint decision-making with others when one cannot fulfill one's interests and/or goals without their involvement [42]. Fairness as a moral concept [43] can be estimated in negotiations through various elements, such as negotiation outcomes, e.g., points per player, or process measures, e.g., how many offers a player made to the opponent [42]. Thus, self- and other- regard is inherent to negotiations, encompassing complex socio-psychological processes [44]. Negotiations therefore involve greater theory of mind reasoning than DG or UG; negotiators have to reason about each others' intentions, trade-offs, and outcomes as a cognitively taxing

process [9]. Especially if negotiators have to cooperate and compete, such as during a mixed-motive negotiation, they often rely on a higher-theory of mind [8]. Mixed-motive negotiations are pertinent scenarios for observing how players attempt to decipher and shape each other's intentions and beliefs, when players engage in higher-order mind perceiving and reasoning.

There are similarities and differences between human-human and human-agent negotiations, though more research is necessary for definitive comparisons. The similarities are that emotions expressed by players affect people's negotiation approach, be it with virtual negotiators [45] or human negotiators [46,47]. An artificial agent's expressed anger, regret, or joy (both facial and textual expressions) influence how human opponents play against it [45], extending the view that emotions in human-human negotiations reveal strategic intentions and influence outcomes [46,47]. To add, priming people's belief about the negotiation (emphasizing cooperation vs. exploitation at the start) impacts human-agent negotiations [48], echoing how framing of a game in itself for human-human negotiations results in divergent outcomes [49]. Increasingly, agents are capable of using complex human-like strategies in negotiation, and the perceived gap between humans and agents may continue to shrink [50].

However, people still do have preconceptions about agents' lack of human-like mind in many negotiation scenarios. People apply their higher order theory of mind reasoning when competing with predictable agents and end up with higher scores when the aim is the win [14]. Specifically, a human opponent is granted agency by default, but a machine's agency can be independent of or dependent on a human actor; the belief about the agent (autonomous vs. human-controlled agent) can result in different tactics adopted by human players [51, 45]. In another study, when machines with higher-order minds negotiated with people, both parties ended up with higher scores (larger joint outcome) when machines made the first bid, but not when humans made the first offer [8]. Thus, an agent's mind and a human player's perception of an agent's mind are crucial to how their exchange unfolds, be it simpler exchanges like DG and UG [6], or more extensive exchanges like negotiations [8].

3 Study 1: Dictator and Ultimatum Games

Our research question was: *In what ways do manipulated agency and patiency via descriptions of an agent influence how participants allocate goods to it in DG and UG?* We assumed that both agency and patiency would impact the UG outcome, as per prior research [6]. Since neither party gets anything if the responder rejects the offer, the human proposers's perception of a machine responder's mind becomes more salient in UG. In DG, the machine responder has no say in the human proposer's distribution scheme. Therefore, we hypothesized that the DG outcome would depend more on patiency (emotional capacity), for the machine is a moral patient without any power to challenge the human moral agent's proposal (Figure 1). Before reporting our results, our manipulation check first looks at whether or not our experimental manipulation (descriptions of the machine) was successfully perceived by participants. In section 3.2, the main analysis is on answering our research question and exploratory analysis looks at additional measures that relate to, but are not a part of, the research question.

3.1 Design

The study was a 2 (Low vs. High) by 2 (Agency vs. Patiency) between-participants factorial design. Based on prior work on moral standing for sentence structure [19] and MPT items for content [4], our manipulation was presented before participants partook in DG and UG as four different descriptions, as follows. The machine has a *simple* vs. *state-of-the-art* artificial intelligence. It *is not* vs. *is* capable of sophisticated logical thinking. (1) It neither feels emotions nor reacts to the emotions expressed by others. Neither can it reason about how its actions and emotional expressions impact other people's emotions. (2) It neither feels emotions nor reacts to the emotions expressed by others though it can reason about how its actions and emotional expressions impact other people's emotions. (3) It feels emotions and reacts to emotions expressed by others, but it cannot reason about how its actions and emotional expressions impact other people's emotions. (4) It feels emotions and reacts to the emotions expressed by others. It can also reason about how its actions and emotional expressions impact other people's emotions. In sum, (1) the machine *does not have a complex disposition to think, feel, and reflect* (low-agency, low-patiency) vs. (2) *has a complex disposition to think and reflect, but cannot feel* (high-agency, low-patiency) vs. (3) *has a complex disposition to feel, but cannot think or reflect* (low-agency, high-patiency) vs. (4) *has a complex disposition to think, feel, and reflect* (high-agency, high-patiency).

3.1.1 Participants and procedure

We recruited participants on Amazon Mechanical Turk. Of the 202 participants, 131 were men (64.85%), 70

were women, and one person was of undisclosed gender. To report the most prominent age, race, and educational level categories, 101 (50%) were between 25 and 34 years of age, 154 identified as White (76.24%), and 135 had some college education or above (66.83%). The survey call stated that participants will partake in a task of distributing 20 tickets between themselves and a machine agent. Tickets entered them into a lottery for an additional $10. Through a survey link, participants first read the informed consent form, answered demographic and emotion state questions, and were randomly assigned one of the four conditions, with accompanying attention check questions that followed the description of a machine.

We called DG round one and UG round two, in order to not refer to these games by their known names. Participants had to read instructions about DG, which stated that they have "a higher chance of winning the lottery with more tickets." This was followed by attention check questions, before participants allocated tickets to the agent in DG. Then participants were asked about their emotion states. Instructions about round two (UG) followed that stated that the machine "can accept or reject your offer [...] (and that the machine's) rejection leads to zero tickets for both of you." After the attention check questions, participants were asked to allocate tickets to the machine, given the new information that the machine can now overturn offers to the loss of both players. After DG and UG, the following measures were taken: MPT [4], stereotype content model questions [52][2],the moral standing scale [19,15], emotion states [55–57], the moral identity questionnaire [58], and the inclusion of other in the self (IOS) scale [20]. We report people's perception of the machine's moral standing and mind (MPT), exploratory analyses of people's emotion states, and how people related to the machine (IOS) for both studies (results sections 3.3 and 4.3); scales can be found in the Appendix. All participants received $1.80 and one randomly chosen participant was awarded the extra compensation of $10 at the end of the experiment.

[2] MPT dimensions conceptually relate to the stereotype content model (SCM). SCM deals with interpersonal perceptions of social group members based on two dimensions of *competence*, e.g., intelligent, competitive, confident, and *warmth*, e.g., friendly, good-natured, sincere [52]. Competence items evoke agency and warmth items are reminiscent of patiency, though the aims of two scales differ [53]. SCM was not relevant for the current paper, but the trends were generally the same as MPT scales as reported previously [54].

3.1.2 Manipulation check

Concerning perceived agency (MPT scale results), there was both a significant main effect of described agency of our text-based manipulation ($F(1, 198) = 26.54$, $p < .001$, $\eta_p^2 = .118$) and a significant main effect of described patiency ($F(1, 198) = 14.92$, $p < .001$, $\eta_p^2 = .07$), whereas the interaction between agency and patiency did not reach significance ($F(1, 198) = 0.75$, $p = .39$, $\eta_p^2 = .00$) according to our ANOVA analysis. Participants perceived lower agency for the agent that could purportedly not reason ($M = 2.88$, $SE = 0.17$) than when the agent was described as being able to reason ($M = 4.09$, $SE = .17$). However, participants also rated the agent as lower in agency when it could not feel ($M = 3.03$, $SE = 0.17$) than when the agent was described as being able to feel ($M = 3.94$, $SE = .17$). Likewise, regarding perceived patiency, there was a significant main effect of agency ($F(1, 198) = 5.52$, $p = .02$, $\eta_p^2 = .03$) as well as a significant main effect of patiency ($F(1, 198) = 25.66$, $p < .001$, $\eta_p^2 = .12$), and the interaction between agency and patiency was not significant ($F(1, 198) = 0.59$, $p = .45$, $\eta_p^2 = .00$). Participants perceived lower patiency for the agent that could purportedly not feel ($M = 2.15$, $SE = 0.17$) than when the agent was described as being able to feel ($M = 3.35$, $SE = .17$). However, participants also rated the agent as lower in patiency when it could not reason ($M = 2.47$, $SE = 0.17$) than when the agent was described as being able to reason ($M = 3.03$, $SE = .17$). Given that agency and patiency were highly correlated in the original MPT study that was conducted by Gray et al. (reported as "$r(11) = .90$, $p < .001$" [15,4]), we used the descriptions as intended.

3.2 Results

3.2.1 Main analysis

We conducted a series of ANOVA tests. For DG allocations, there was no main effect of agency ($F(1, 198) = 0.21$, $p = .65$); however, there was both a marginal main effect of patiency ($F(1, 198) = 3.53$, $p = .062$, $\eta_p^2 = .02$) and a significant interaction between agency and patiency ($F(1, 198) = 6.26$, $p = .013$, $\eta_p^2 = .03$) for DG results. Across patiency conditions, participants gave less to the machine when it purportedly could not feel ($M = 5.29$, $SE = 0.64$) than when it was described to be able to feel ($M = 6.98$, $SE = .63$). But, this effect was driven entirely by the low agency condition ($M = 3.96$, $SE = .90$ vs. $M = 7.9$, $SE = .90$) and was absent in the high agency condition ($M = 6.62$, $SE = .90$ vs. $M = 6.06$, $SE = .90$).

In UG, there was both a significant main effect of agency ($F(1, 198) = 3.90$, $p = .05$, $\eta_p^2 = .02$) and a significant main effect of patiency ($F(1, 198) = 7.58$, $p = .007$, $\eta_p^2 = .04$) on allocations, whereas the interaction between agency and patiency did not reach significance ($F(1, 198) = 2.12$, $p = .15$, $\eta_p^2 = .01$). Participants gave less to the machine when it could purportedly not reason ($M = 8.63$, $SE = 0.51$) than when the machine was described as being able to reason ($M = 10.04$, $SE = .51$). Likewise, participants gave less to the machine when it could not feel ($M = 8.35$, $SE = 0.51$) than when the agent was described as being able to feel ($M = 10.32$, $SE = .50$). Although covariance between allocations in UG and DG was high ($F(1, 197) = 62.75$, $p < .001$, $\eta_p^2 = .24$), when controlling for DG outcome[3], we observed the same pattern in UG; there was still both a significant main effect of agency ($F(1, 197) = 4.02$, $p = .046$, $\eta_p^2 = .02$) and a significant main effect of patiency ($F(1, 197) = 4.31$, $p = .039$, $\eta_p^2 = .02$), and the interaction between agency and patiency was not significant ($F(1, 197) = 0.07$, $p = .80$, $\eta_p^2 = .00$). Thus, participants still gave less to the machine that could purportedly not reason ($M = 8.71$, $SE = 0.44$) than when the agent was described as being able to reason ($M = 9.97$, $SE = .44$). Likewise, participants allocated less to the machine that could not feel ($M = 8.68$, $SE = 0.45$) than when the machine was described as being able to feel ($M = 10$, $SE = .44$).

3.2.2 Exploratory analysis

Our ANOVA analysis showed that people highly related to the agent (IOS) based on its manipulated patiency, i.e., how much emotional behavior the agent showed ($F(1, 198) = 6.99$, $p = .009$, $\eta_p^2 = .03$). But, agency and the interaction between agency and patiency were not significant ($Fs < .85$, $ps > .36$). An agent described to have feelings was more relatable ($M = 3.01$, $SE = .19$) than an agent that could not have emotions ($M = 2.3$, $SE = .19$). As for the agent's moral standing, significance was found in regard to its manipulated agency ($F(1, 198) = 6.60$, $p = .011$, $\eta_p^2 = .03$) and patiency ($F(1, 198) = 5.17$, $p = .024$, $\eta_p^2 = .03$). Their interaction neared significance ($F(1, 198) = 3.65$, $p = .06$, $\eta_p^2 = .02$). The agent was granted higher moral standing when it could feel ($M = 4.20$, $SE = .17$) compared to when it could not feel ($M = 3.65$, $SE = .17$). Also, its high cognitive capacity contributed to greater moral standing ($M = 4.23$, $SE = .17$) compared to when the agent had low cognitive capacity ($M = 3.62$, $SE = .17$).

[3] We put in how much people gave the agent in DG as a covariate to control for its affect in analyzing UG outcomes.

4 Study 2: Negotiation

The research question was the following. *In what ways do manipulated agency and patiency (via dialogs and descriptions) of an agent that negotiates with a human influence the negotiation outcome and process?* We expected that agency would drive participants to partake in heightened engagement with the agent to (1) increase the joint outcome of the negotiation (regardless of who wins) and (2) would cause participants to seek more game-relevant information from the agent (send more messages on preferences and offers to the agent). Higher joint outcome implies greater cognitive effort, for it requires players' usage of higher-order theory of mind reasoning to increase the size of the "pie" for mutually beneficial ends. We hypothesized that the machine's patiency would increase participants' other regard; participants would grant the agent (1) fairer allocations and (2) would send greater numbers of emotionally-valenced messages. Agency and patiency were assumed to both contribute to negotiation outcome and processes [6].

4.1 Design

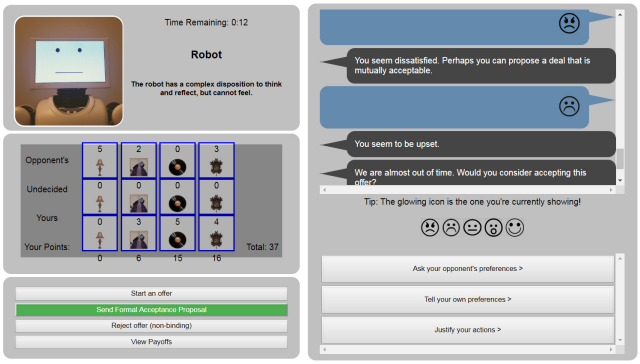

Fig. 2 Negotiation interface in Study 2

Our agent was a virtual robot that was simple in appearance (Figure 2), without any gender, race, or other highly anthropomorphic traits that may trigger people's biases [29–31], which helped to drive the perception of its mind based on its behavior rather than its looks. We used a configurable negotiation platform called IAGO for designing custom negotiation experiments. It features emotional communication (participants can click on different emojis to send to an agent; see Figure 1), as well as customizable agents (e.g., agents' pictures can have different emotional expressions as reactions to people's behavior) [59].

We again employed a between-participants factorial design of 2 (Low vs. High) by 2 (Agency vs. Patiency)

Robot type	*Description*	**Dialog**
Low-Agency Low-Patiency	*The robot does not have a complex disposition to think, feel, and reflect.*	**"Preparing offer." "Affirmative." "Does not compute."**
Low-Agency High-Patiency	*The robot has a complex disposition to feel, but cannot think or reflect.*	**"I like this!" "Yay! I'm happy." "Oh...I'm sad..."**
High-Agency Low-Patiency	*The robot has a complex disposition to think and reflect, but cannot feel.*	**"This is the most logical offer." "I inferred that you would accept this deal." "You seem to be upset."**
High-Agency High-Patiency	*The robot has a complex disposition to think, feel, and reflect.*	**"I'm going to make this offer." "I feel so good about negotiating with you!" "Oh...Your sadness makes me feel sad..."**

Table 1 Agent types and excerpts from their descriptions and dialogs in Study 2.

dimensions. Agency and patiency were manipulated in two ways. There were descriptions of the agent presented before the negotiation and shortened versions of descriptions appeared next to the picture of the agent (Figure 2) during the experiment. These descriptions were the same as Study 1. In addition, we designed dialogs, i.e., how it "talked" (Table 1 lists excerpts). We used the items of the MPT scale [4] to construct the dialogs, as we did with descriptions. To illustrate, one agency item, "the robot appears to be capable of understanding how others are feeling" was translated to the agent having an awareness of the participant's emotion states during the negotiation, e.g., a "sad" emoji from the participant resulted in "you seem to be upset" message from the high-agency low-patiency agent while the agent's expression remained neutral (Figure 2). This suggests high-agency, but does not directly translate to a complete lack of emotional capacity (the agent is aware of the other player's emotion states), even though the description stated it "cannot feel".

We attempted to imbue the high-agency low-patiency agent with an awareness of others' emotions (e.g. - "you seem to be upset") whilst not being emotionally expressive itself, which are two different, but often conflated, design elements of affective artificial agents. In contrast, the low-agency low-patiency agent did not use emotional language or expressions (static neutral face) and always responded to participants' emojis with the statement "does not compute". Hence, unlike prior work [6], our agency and patiency manipulation separated an agent's awareness of displayed emotions (agency) from actually feeling emotions (patiency). We imbued agency and patiency features into agents' descriptions and dialogs that occur *over time* in a negotiation (Table 1), which is how we carefully manipulated the mind dimensions according to MPT (in contrast to [6]).

As a reminder, only dialogs and descriptions differed between agents (Table 1); the negotiation tactic was the same for all agents, for we are interested in the effects of MPT dimensions. Items for all negotiations were also the same with 7 clocks, 5 crates of records, 5 paintings, and 5 lamps, with different values per item per player for records and lamps (Table 2). All agents began the negotiation by proposing the same starting offer (Table 3). The negotiation structure was partially integrative and partially distributive, meaning that half of the items were equally valuable to both players (distributive) while the other half of items had different values for players (integrative). This allows players to potentially "grow the pie" in a cooperative fashion through in-game communication while still playing competitively. Before the negotiation, participants were informed only about what they preferred. They were told prior to the experiment that one person who earned the highest points against the agent would get $10 as a bonus prize.

	Clocks	Records	Paintings	Lamps
Robot	**4**	**1**	**2**	**3**
Human	**4**	**3**	**2**	**1**

Table 2 Points per item

All agents' negotiation strategy was based on the *minimax* principle of minimizing the maximal potential loss [59]; agents adjusted their offers if participants communicated their preferences, and strove for fair offers, while rejecting unfair deals. Agents did not know participants' preferences, but assumed an *integrative* structure. At the start, an agent made a very lopsided first offer (as a form of "anchoring") as shown in Table 3: it took almost all clocks (equally the most valuable item for both players), it allocated more lamps to itself (more valuable for itself) and gave more records to the participant (more valuable for the participant), and equally distributed the paintings (equally valuable item). This suggests that negotiators can cooperate and compete, potentially to enlarge the pie for both.

4.2 Participants and procedure

226 participants residing in the U.S. were recruited on Amazon Mechanical Turk. We had 135 men (59.7%),

	Clocks	Records	Paintings	Lamps	Pts.
Robot	*6**4**	*0**1**	*2**2**	4**3**	**40**
Undecided	1	1	1	1	
Human	*0**4**	*4**3**	*2**2**	0**1**	**16**

Table 3 Agents' starting offer: In all conditions, agents made the same, lopsided first offer as displayed. There were undecided items, one of each type. Points per item differed, thus the calculation stands as *item* * **points** = **total points**.

90 women, and 1 of undisclosed gender. Participants were all over 18 years of age. 53.5% were between the ages of 25 and 34 (121 participants). Participants got a link to the survey that first contained the informed consent form, questions on participants' current emotion states and demographic information. Then participants read the description of an agent based on the randomly assigned condition (Table 1) and answered attention check questions about the description. After that, they read the instruction about the negotiation task, followed by additional attention check questions about the task, which they had to pass to go to the negotiation interface. They had up to 6 minutes to engage in a negotiation of four different goods (Table 2), and the count-down of time was displayed on the interface (Figure 2). Upon completion of the negotiation, participants finished the second part of the survey of our measurements. We deployed the same measurements as the first study (Section 3.1.1). Further, we asked additional questions on whether or not participants made concessions to the agent and if the agent did anything unexpected. Participants were compensated $3 for their time, based on an estimate of 30 minutes to finish the entire survey and negotiation. One participant was randomly selected and awarded the $10 bonus prize, after the experiment was completed.

4.2.1 Manipulation check

We again report our manipulation check first on whether our dialog design and description affected people's perception of the robot's mind. Both of our experimental manipulations affected perceived agency according to an ANOVA test. That is, there was both a significant main effect of agency ($F(1, 222) = 35.68$, $p < .001$, $\eta_p^2 = .14$) and a significant main effect of patiency ($F(1, 222) = 53.42$, $p < .001$, $\eta_p^2 = .19$) on perceived agency, whereas the interaction between agency and patiency did not approach significance ($F(1, 222) = .60$, $p = .44$, $\eta_p^2 = .003$). Participants perceived lower agency for the agent that could purportedly not reason ($M = 2.89$, $SE = .14$) than the agent described to have high ability to reason ($M = 4.01$, $SE = .13$). But, people also rated the agent as lower in agency when it did not have affective capacity ($M = 2.77$, $SE = .13$) than when the agent could have emotions ($M = 4.14$, $SE = .13$). In contrast, only manipulated patiency significantly affected perceived patiency ($F(1, 222) = 71.24$, $p < .001$, $\eta_p^2 = .24$); the effect of agency on perceived patiency only approached significance ($F(1, 222) = 2.57$, $p = .11$, $\eta_p^2 = .01$), and the interaction did not approach significance ($F(1, 222) = .001$, $p = .99$, $\eta_p^2 = .00$). Participants rated the agent as lower in patiency when it could not feel ($M = 1.88$, $SE = .13$) than when the agent was described as being able to feel ($M = 3.44$, $SE = .13$).

4.3 Results

4.3.1 Main analysis

This subsection is on ANOVA tests for answering our research question on whether manipulated agency and patiency affects the negotiation outcome and process. Here, 78 MTurkers were excluded as outliers for negotiation-related analyses due to automatically detected inattention, i.e., no engagement with the negotiation interface and were thus timed out. For user points, there was a significant main effect of agency ($F(1, 143) = 4.35$, $p = .04$, $\eta_p^2 = .03$); participants got more in the negotiation when the agent was described as being able to reason ($M = 28.825$, $SE = .67$) than when the agent was described as not being able to reason ($M = 26.69$, $SE = .77$). No other effects approached significance ($Fs < .50$, $ps > .48$). For agent points, there was also a significant main effect of agency ($F(1, 143) = 6.68$, $p = .01$, $\eta_p^2 = .05$); agents got less in the negotiation when it was described as being able to reason ($M = 34.06$, $SE = .76$) than when the agent was described as not being able to reason ($M = 37.05$, $SE = .87$). No other effects approached significance ($Fs < .23$, $ps > .63$). Figure 3 displays the agent's end outcomes in DG, UG, and negotiation via standardized scores for comparisons; low-agency agents had the best outcomes in negotiations. Thus, the positive effect of agency on user points and the negative effect of agency on agent points cancelled out, such that the effect of agency on joint points was not significant $F(1, 143) = 1.66$, $p = .20$); no other effects approached significance ($Fs < .58$, $ps > .44$). Further, the effect of agency on the initial offer was not significant $F(1, 143) = .49$, $p = .49$); no other effects reached significance ($Fs < 2.7$, $ps > .10$).

Negotiation process measures capture *how* participants played against the agent and are important to negotiations. There was a marginal effect of agency on game end time ($F(1, 143) = 3.62$, $p = .059$, $\eta_p^2 = .03$); participants took longer if the agent was described as not being able to reason ($M = 296.88$, $SE = 13.36$) than

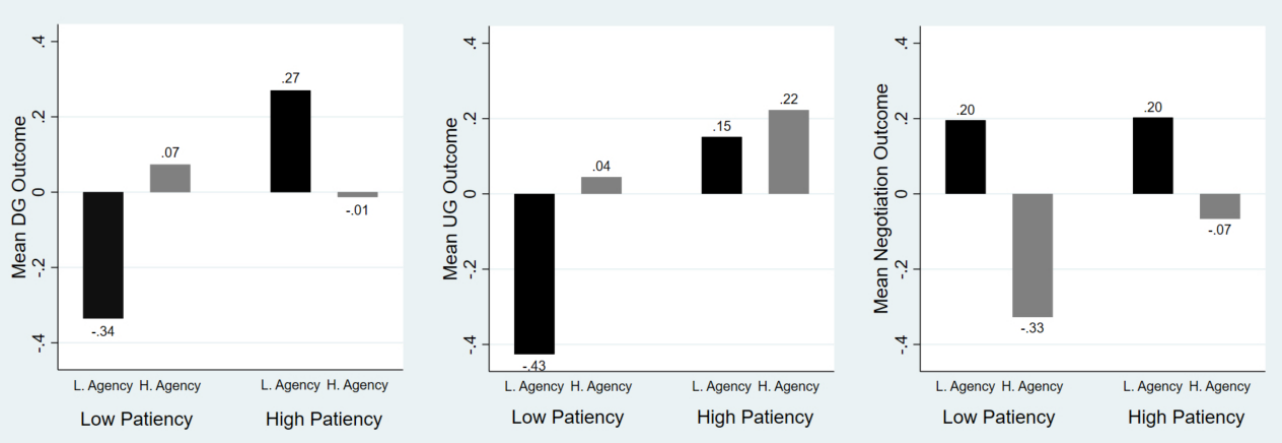

Fig. 3 Agents' standardized scores across DG, UG, and negotiation as outcomes, over low and high agency and patiency.

when the agent was described as being able to reason (M = 263.14, SE = 11.67). But, this effect was driven entirely by the low-patiency condition, as per a significant interaction (F(1, 143) = 5.38, p = .02, η_p^2 = .04). The main effect of patiency did not approach significance (F < .01, p > .99). There was a parallel pattern for number of rejected offers. We saw a significant effect of agency on number of times users rejected offers (F(1, 143) = 9.50, p = .002, η_p^2 = .06); participants were more likely to reject an offer if the agent was described as not being able to reason (M = .72, SE = .11) than when the agent was described as being able to reason (M = .29, SE = .09). However, this effect was again driven entirely by the low-patiency condition, as per a significant interaction (F(1, 143) = 5.85, p = .02, η_p^2 = .04). The main effect of patiency did not reach significance (F < 2.32, p > .13).

Participants chose to display the happy emoji significantly more when the agent was described as being able to feel (M = 1.25, SE = .18; F(1, 143) = 8.14, p = .005) than when the agent was described as not being able to feel (M = .88, SE = .20). No other effects reached significance (Fs < 1.92, ps > .17). Likewise, participants also chose to display the surprise emoji significantly more when the agent was described as being able to feel (M = .47, SE = .07; F(1, 143) = 4.54, p = .04) than when the agent was described as not being able to feel (M = .25, SE = .08). No other effects reached significance (Fs < 1.60, ps > .21). No other effects for any other emoji emotional display reached significance (Fs < 1.95, ps > .17).

There were a few messages that participants sent to the agent (pre-set messages in the user interface) that were significantly used. Participants chose to convey the message "it is important that we are both happy with an agreement" more when the agent was described as being able to feel (M = .36, SE = .06; F(1, 143) = 5.18, p = .02, η_p^2 = .04) than when the agent was described as not being able to feel (M = .16, SE = .07). No other effects approached significance (Fs < .03, ps > .85).

The interaction between agency and patiency significantly affected how often participants chose to convey the message: "I gave a little here; you give a little next time" (F(1, 143) = 4.25, p = .04, η_p^2 = .03). People sent this message the most to the high patiency, low agency agent (M = .158 SE = .04) and the least to the agent was described to neither feel nor display cognitive thinking (M = -1.43, SE = .05). No other effects reached significance (Fs < 2.87, ps > .09). There was also a significant interaction between agency and patiency for this message "This is the last offer. Take it or leave it" (F(1, 143) = 3.88, p = .05, η_p^2 = .03). The message was shared the most with the agent that was low in agency, but high in patiency (M = .08, SE = .03) and the least with the agent that was high and both agency and patiency (M = - 3.5, SE = .02). No other effects reached significance (Fs < .85, ps > .36). No other effects for any other message options reached significance (Fs < 2.17, ps > .14).

4.3.2 Exploratory analysis

We looked at additional measures (listed in the Appendix) in an exploratory manner to see how mind perception dimensions may have influenced the robot's moral standing and how people identified with it. Only manipulated patiency significantly affected psychological relatedness (Inclusion of Other in Self: IOS) to the agent (F(1, 222) = 29.1, p = .002, η_p^2 = .04); the effect of agency on IOS and the interaction did not reach significance (Fs < 1.16, ps > .28). Participants identified with the agent more when it was described as being able to feel (M = 2.86, SE = .16) and that the agent was more distant from them psychologically when it could

not feel (M = 2.14, SE = .16). Only manipulated patiency significantly affected moral standing (F(1, 222) = 17.81, p < .00001, η_p^2 = .07); the effect of agency on moral standing and the interaction did not reach significance (Fs < 1.53, ps > .22). Participants rated the agent as lower in moral standing when it could not feel (M = 3.08, SE = .16) than when the agent was described as being able to feel (M = 4.03, SE = .16).

5 General discussion

Our article concerns how the imbued mind of agents based on MPT dimensions influence the results of DG, UG, and negotiation as human-agent interactions. In Study 1, we found that an agent's described patiency marginally affected the allocation scheme in DG, with an interaction between agency and patiency; in UG, described agency and patiency influenced allocations to the agent (as in [6]). Yet unexpectedly in negotiations, we only noted a significant effect of agency, in a different direction than anticipated. Low-agency agents ended with higher scores (Figure 3) and also had longer negotiation periods. In comparison, high-agency agents had *lower* scores, particularly if they also had low-patiency (Spock-like agent), while negotiations themselves were shorter. The results on negotiation outcomes and processes, two paradigmatic measures in negotiation research [42], did not align with our hypotheses, while DG and UG results echoed prior research [6].

Compared to DG and UG, negotiations serve as a context for adjusting preconceptions on technological entities' minds. We buttress this on three premises. First, people have preconceived beliefs about artificial agents' minds; agents are seen to have low-order theory of minds [4,21] (at least presently) even if people interact with them socially [22,23]. Second, the perceived mind of an agent can be adjusted, be it through patiency (affective richness [32]) or agency (behavioral intentionality [33]). Third, negotiations require cognitively effortful participation that involves theory of mind reasoning [9,8], especially when it comes to mixed-motive negotiations[4] [60,8]. Through negotiations, an agent's behavioral intentionality can be called into question, providing people opportunities to reformulate an agent's degree of conferred mind.

Participants' behavior suggests that the common belief that technological agents have low-agency and low-patiency [4,21] was called into question for low-agency agents. All agents adjusted their offers in the same way if participants communicated about preferences [59], so they appeared to calculatively negotiate though we did not implement any sophisticated AI. Thus, the disjointed nature between our low-agency agent's dialogs and descriptions vs. its negotiation style (mixed-motive games often require higher-order theory of mind) potentially called into question what the agent was "up to." Our high-agency agent did poorly against participants that do have a higher degree of mind. Our low-agency agent did well against participants, interactively over time. When people cannot easily guess what an agent desires or intends to do, i.e., predict its intentional stance [3], they can exercise a higher degree theory of mind. Participants with low-agency agents thus applied higher theory of mind reasoning, not necessarily just on game strategies, but on investigating and questioning their bias held as a fact—the inability of technology to have a human-like mind.

An agent that was described to be less cognitively intelligent (low-agency) interacted with participants in a cognitively taxing task (negotiation over goods), and this disjuncture gave people reasons to doubt their beliefs over time, i.e., we manipulated an agent's behavioral intentionality [33]. Participants' assumed "winning" strategy could have drifted from point-based calculations as the time passed or it was initially assumed to not be just about item points. For one, emotional capacity of agents in Study 2 affected the outcome in an unexpected manner. Though people utilized more emotive messages and emojis with high-patiency agents, this behavior did not influence outcomes; perceived patiency did not significantly affect negotiation results. Potentially there was more "noise" to interpret when people interacted with high-patiency agents. Not only do they have to figure out game mechanics in terms of item values, but people may have assumed that the agents' emotional capacity served a strategic purpose, even though agents' offer strategies were not affected by players' emotional communications. Emotions matter in how people take part in negotiations [61,47], so people may have assumed that agents' emotions also served some purpose.

Qualities such as an agent's emotions, moral standing and relatability are in essence, distracting points when it comes to game mechanics. Yet, these distractors could have (wrongly) gained greater traction as the negotiation continued over time, especially since harm salience regarding a moral patient increases with time pressure [18]. Thus, by perceiving other minds over time, people can become sensitive to not only their own suffering as moral patients [10,62], but also to the suffering of others, even when they are machine opponents [18,63,64]. We find that identifying with a technological

[4] Mixed-motive negotiations have both cooperative and competitive goals in mind, i.e., "growing the pie" for all while attempting to competitively gain a greater share of the pie than the opponent.

moral patient via manipulated mind can change people's behavior towards it. Our exploratory analyses on IOS and moral standing contribute to this interpretation. Their relation to game outcomes are summarized in Table 4.

	Outcome	Moral standing	Relatability
DG	P + I	A + P + I	P
UG	A + P	A + P + I	P
Nego.	A	P	P

Table 4 The impact of manipulated agency and patiency on outcomes, moral standing, and relatability (IOS). Agency is denoted as A, patiency as P, and their interaction as I.

We relate to others by seeing ourselves in them [20]. Our own minds are the basis [5] to relate to our own and others' affective and cognitive capacities. These are summated as two dimensions of the mind, i.e., agency and patiency [4], which builds on ample research that emotions and cognition mutually influence each other in driving behavior (e.g., [16,65,66]). However, only patiency, the perceived propensity to feel emotions, significantly contributed to how much people identified with agents across DG, UG, and negotiation (Table 4). Hence, both studies have the same trend regarding relatability (IOS) [20]; people related to agents' patiency, even if agency may drive strategic decision-making when greater "mind reading" is required as economic exchanges become more complex.

Interestingly, attributing moral standing to an agent followed a different pattern from IOS. In DG and UG, both described agency and patiency affected moral standing, but only imbued patiency impacted the agent's moral standing in negotiations. Strategic games that require a higher order theory of mind reasoning hinge on perceived agency for outcomes as the exchange becomes more complex (from DG and UG to negotiation). But no matter how much theory of mind a strategic exchange requires, people relate to an agent that has affective capacities. An agent's moral standing may become more dependent on its perceived ability to feel when interactions become more complex; *it is painted more as a moral patient over time* [18]. Especially as a strategic exchange increasingly makes people exercise their higher order theory of mind, people act with greater human agency *against* the machine, which is then rendered more as a moral patient (Figure 1).

One novel implication is that mind perception may require theoretical revisions to account for *interactive* opinion formation about an agent's mind; negotiations provide a contextually different framework than a single instance evaluation of an agent's mind (as in DG or UG). Mind perception theory focused more on the latter case; it is about people's pre-existing beliefs at a single point in time and minds of various beings were judged through a survey [4]. The novelty of our studies is that people seem to be revising their opinion of the agent's perceived mind over the course of a complex interaction; the human attribution of a mind in a machine may be misguided, but people can question their own beliefs through an interaction.

Future works can better address multimodality and interaction contexts. Robust multimodal behavior was neither in Study 1 that only manipulated text-based descriptions, nor in Study 2 that also included simple dialogs, emojis, and emotional expressions of the robot. Additional modalities like bodily gestures, gaze behavior, and speech can be explored. Negotiations are potentially one of many interactive paradigms that can better enlighten us on how people assess agents that display different degrees of having a mind in different ways over time. More relevantly, exploring other types of exchanges, e.g., purely integrative or distributive negotiations, can reveal in what ways an agent's perceived mind impact people as they attempt to understand whether or not a social agent is also an intentional agent.

6 Conclusion

We are far from having artificial agents that are truly intentional actors like humans. But, the degree to which artificial agents are perceived to have agency and patiency, and what effect such manipulation has, can be observed. The DG outcome was influenced by perceived patiency of an agent, and the UG outcome was affected by perceived agency and patiency. Yet compared to single instance economic exchanges like DG or UG, interactive negotiations allowed us to catch a glimpse of how people react when they encounter agents that behave counter-intuitively, e.g., negotiating in an agentic manner without prescribed agentic traits, as manipulated via dialogs and descriptions. In negotiations, participants got more points against an agent with high-agency. In contrast, they did worse, took longer to play, and rejected more offers from a low-agency agent, as influenced by patiency. Patiency resulted in more emotional expressions from participants to the agent; people engaged more with emotional signals, i.e., emojis and messages.

As interactions require people to increasingly exercise their greater theory of mind reasoning against non-human agents, e.g., from DG, UG, to negotiation, game outcomes depend more on agents' cognitive traits

while machines' moral standing depends more on perceived affective traits (Table 4). Both agency and patiency contributed to an agent's moral standing after DG and UG, but people granted higher moral standing and related more to it only when it had manipulated patiency after negotiating with it. Yet, people's ability to relate to agents consistently is on whether they can have human-like feelings, regardless of people's own level of theory of mind required in an interaction. Thus, people relate to machines via emotions.

In a strategic exchange, artificial emotions can contribute to machines' moral standing only when humans interactively act *against* machines with agency; this is matched by a machine's displays of traits of being a moral patient, e.g., emotional expressions, in *response* to people's agentic actions. We additionally conjecture that an artificial agent that sends unclear or mismatched signals (as both emotional and non-emotional communication) that people have to interpret during a complex interaction like negotiation can lead them to reconsider artificial agents' perceived minds, more so than in single-shot games like DG and UG that are not interactive. What we can conclude is that in attempting to comprehend an artificial agent's "mind", people react to its rational and emotional capacities in divergent ways, leading to noticeable differences in how people behave.

7 Acknowledgments

Research was sponsored by the Army Research Office and was accomplished under Cooperative Agreement Number W911NF-20-2-0053. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein

References

1. N. Epley, A. Waytz, J.T. Cacioppo, Psychological Review **114**(4), 864 (2007)
2. S. Marchesi, D. Ghiglino, F. Ciardo, J. Perez-Osorio, E. Baykara, A. Wykowska, Frontiers in Psychology **10**, 450 (2019)
3. D. Dennett, *The Intentional Stance* (MIT press, Cambridge, MA, USA, 1989)
4. H.M. Gray, K. Gray, D.M. Wegner, Science **315**(5812), 619 (2007)
5. N.C. Krämer, A. von der Pütten, S. Eimler, in *Human-Computer Interaction: The Agency Perspective*, ed. by M. Zacarias, J.V. de Oliveira (Springer, Heidelberg, Germany, 2012), pp. 215–240
6. C.M. De Melo, J. Gratch, P.J. Carnevale, in *Twenty-Eighth AAAI Conference on Artificial Intelligence* (2014)
7. K. Gray, L. Young, A. Waytz, Psychological Inquiry **23**(2), 101 (2012)
8. H. de Weerd, R. Verbrugge, B. Verheij, Autonomous Agents and Multi-Agent Systems **31**(2), 250 (2017)
9. J. Gratch, D. DeVault, G.M. Lucas, S. Marsella, in *Intelligent Virtual Agents*, ed. by W.P. Brinkman, J. Broekens, D. Heylen (Springer International Publishing, Cham, 2015), pp. 201–215
10. A.G. Sanfey, J.K. Rilling, J.A. Aronson, L.E. Nystrom, J.D. Cohen, Science **300**(5626), 1755 (2003)
11. D. Premack, G. Woodruff, Behavioral and Brain Sciences **1**(4), 515 (1978)
12. D. Dennett, *Kinds of Minds: Toward an Understanding of Consciousness* (Basic Books, New York, NY, USA, 2008)
13. N.C. Krämer, in *Modeling Communication with Robots and Virtual Humans*, ed. by I. Wachsmuth, G. Knoblich (Springer, Heidelberg, Germany, 2008), pp. 222–240
14. K. Veltman, H. de Weerd, R. Verbrugge, Journal on Multimodal User Interfaces **13**(1), 3 (2019)
15. J. Piazza, J.F. Landy, G.P. Goodwin, Cognition **131**(1), 108 (2014)
16. A.R. Damasio, *Descartes' Error: Emotion, Reason, and the Human Brain* (Random House, New York, NY, USA, 2006)
17. L. Floridi, *The Ethics of Information* (Oxford University Press, New York, NY, USA, 2013)
18. K. Gray, C. Schein, A.F. Ward, Journal of Experimental Psychology: General **143**(4), 1600 (2014)
19. M. Khamitov, J.D. Rotman, J. Piazza, Cognition **146**, 33 (2016)
20. A. Aron, E.N. Aron, D. Smollan, Journal of Personality and Social Psychology **63**(4), 596 (1992)
21. A. Waytz, K. Gray, N. Epley, D.M. Wegner, Trends in Cognitive Sciences **14**(8), 383 (2010)
22. C. Nass, J. Steuer, E.R. Tauber, in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM (Association for Computing Machinery, New York, NY, USA, 1994), pp. 72–78
23. J. Blascovich, J. Loomis, A.C. Beall, K.R. Swinth, C.L. Hoyt, J.N. Bailenson, Psychological Inquiry **13**(2), 103 (2002)
24. B. Fogg, C. Nass, in *CHI'97 Extended Abstracts on Human Factors in Computing Systems (CHI EA '97)*. ACM (Association for Computing Machinery, New York, NY, USA, 1997), pp. 331–332
25. J. Mell, G.M. Lucas, J. Gratch, in *International Conference on Autonomous Agents and Multiagent Systems* (International Foundation for Autonomous Agents and Multiagent Systems, 2015), pp. 1567–1576
26. Y. Moon, C. Nass, Communication Research **23**(6), 651 (1996)
27. G.M. Lucas, J. Gratch, A. King, L.P. Morency, Computers in Human Behavior **37**, 94 (2014)
28. J. Mell, G.M. Lucas, J. Gratch, in *International Conference on Intelligent Virtual Agents* (Springer, 2017), pp. 273—282
29. J.N. Bailenson, J. Blascovich, A.C. Beall, J.M. Loomis, Personality and Social Psychology Bulletin **29**(7), 819 (2003)
30. R. Dotsch, D.H. Wigboldus, Journal of Experimental Social Psychology **44**(4), 1194 (2008)
31. M. Siegel, C. Breazeal, M.I. Norton, in *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems* (IEEE, 2009), pp. 2563–2568

32. K. Gray, D.M. Wegner, Cognition **125**(1), 125 (2012)
33. D.T. Levin, S.S. Killingsworth, M.M. Saylor, S.M. Gordon, K. Kawamura, Human–Computer Interaction **28**(2), 161 (2013)
34. J. Graham, B.A. Nosek, J. Haidt, R. Iyer, S. Koleva, P.H. Ditto, Journal of Personality and Social Psychology **101**(2), 366 (2011)
35. W. Güth, R. Schmittberger, B. Schwarze, Journal of Economic Behavior & Organization **3**(4), 367 (1982)
36. H. Oosterbeek, R. Sloof, G. Van De Kuilen, Experimental Economics **7**(2), 171 (2004)
37. C. Engel, Experimental Economics **14**(4), 583 (2011)
38. R. Forsythe, J.L. Horowitz, N.E. Savin, M. Sefton, Games and Economic Behavior **6**(3), 347 (1994)
39. I. Bohnet, B.S. Frey, American Economic Review **89**(1), 335 (1999)
40. L. Cosmides, J. Tooby, in *Moral Psychology: The Evolution of Morality: Adaptations and Innateness*, ed. by W. Sinnott-Armstrong, C.B. Miller (Cambridge, MA, USA, 2008), pp. 114–137
41. P.J. Carnevale, D.G. Pruitt, Annual review of psychology **43**(1), 531 (1992)
42. L.L. Thompson, J. Wang, B.C. Gunia, Annual Review of Psychology **61**, 491 (2010)
43. J. Graham, J. Haidt, S. Koleva, M. Motyl, R. Iyer, S.P. Wojcik, P.H. Ditto, in *Advances in Experimental Social Psychology*, vol. 47, ed. by P. Devine, A. Plant (London, UK: Elsevier, 2013), pp. 55–130
44. L. Thompson, Psychological Bulletin **108**(3), 515 (1990)
45. C.M. de Melo, P.J. Carnevale, S.J. Read, J. Gratch, Journal of Personality and Social Psychology **106**(1), 73 (2014)
46. M.W. Morris, D. Keltner, in *Research in Organizational Behavior*, vol. 11, ed. by B. Staw, R. Sutton (JAI, Amsterdam, the Netherlands, 1999), pp. 1–50
47. B. Barry, I.S. Fulmer, G.A. Van Kleef, et al., in *The Handbook of Negotiation and Culture*, ed. by M.J. Gelfand, J.M. Brett (Stanford Business Books, Stanford, CA, USA, 2004), pp. 71–94
48. C.M. de Melo, P. Khooshabeh, O. Amir, J. Gratch, in *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS 2018)* (International Foundation for Autonomous Agents and Multiagent Systems, 2018), pp. 2224–2226
49. D.G. Pruitt, Journal of Personality and Social Psychology **7**(1, part 1), 21 (1967)
50. T. Baarslag, M. Kaisers, E. Gerding, C.M. Jonker, J. Gratch, in *International Joint Conference on Artificial Intelligence* (2017), pp. 4684–4690
51. C.M. de Melo, J. Gratch, P.J. Carnevale, IEEE Transactions on Affective Computing **6**(2), 127 (2015)
52. S.T. Fiske, A.J. Cuddy, P. Glick, J. Xu, Journal of Personality and Social Psychology **82**(6), 878 (2002)
53. N. Haslam, Psychological Inquiry **23**(2), 172 (2012)
54. M. Lee, G. Lucas, J. Mell, E. Johnson, J. Gratch, pp. 38–45 (2019)
55. J. Haidt, in *Handbook of Affective Sciences*, vol. 11, ed. by R.J. Davidson, K.R. Scherer, H.H. Goldsmith (Oxford University Press, Oxford, UK, 2003), pp. 852–870
56. E.E. Skoe, N. Eisenberg, A. Cumberland, Personality and Social Psychology Bulletin **28**(7), 962 (2002)
57. C.M. de Melo, J. Gratch, in *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)* (IEEE Computer Society, Los Alamitos, CA, USA, 2015), pp. 315–321
58. J.E. Black, W.M. Reynolds, Personality and Individual Differences **97**, 120 (2016)
59. J. Mell, J. Gratch, in *Proceedings of the 16th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS 2017)* (International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 2017), AAMAS '17, p. 401–409
60. D.G. Pruitt, M.J. Kimmel, Annual Review of Psychology **28**(1), 363 (1977)
61. G.A. Van Kleef, C.K. De Dreu, A.S. Manstead, Journal of personality and social psychology **87**(4), 510 (2004)
62. M. Van't Wout, R.S. Kahn, A.G. Sanfey, A. Aleman, Experimental Brain Research **169**(4), 564 (2006)
63. J. Hewig, R.H. Trippe, H. Hecht, M.G. Coles, C.B. Holroyd, W.H. Miltner, Cortex **44**(9), 1197 (2008)
64. P. Bloom, *Against Empathy: The Case for Rational Compassion* (Random House, London, UK, 2017)
65. J.D. Greene, L.E. Nystrom, A.D. Engell, J.M. Darley, J.D. Cohen, Neuron **44**(2), 389 (2004)
66. J. Greene, J. Haidt, Trends in cognitive sciences **6**(12), 517 (2002)

8 Appendix

8.1 Mind Perception

We inserted "the robot" to the original phrasing in our adaptation of the scale by Gray and colleagues [4]. The first seven items are on agency and the rest are on patiency. *The robot appears to be capable of... (1 - strongly disagree, 7 - strongly agree):*

- making plans and working towards goals.
- trying to do the right thing and telling right from wrong.
- remembering things.
- understanding how others are feeling.
- exercising self-restraint over desires, emotions or impulses.
- thought.
- conveying thoughts or feelings to others.
- longing or hoping for things.
- experiencing embarrassment.
- feeling afraid or fearful.
- feeling hungry.
- experiencing joy.
- experiencing physical or emotional pain.
- experiencing physical or emotional pleasure.
- experiencing pride.
- experiencing violent or uncontrolled anger.
- having experiences and being aware of things.

Note: The ability to convey thoughts and feelings to others are categorized under agency. We did not use an item on personality under patiency (as in [4]) because the above ten items on patiency are on abilities to have experience and emotions, which we wanted to focus on.

8.2 Moral Standing

We evaluated people's attribution of moral standing to the machine with modified questions from prior research [19]; we added in "robot" in our phrasing. *Please assess the robot on the following criteria, from a scale of 1 (not at all) to 7 (extremely):*

- How morally wrong do you think it would be for someone to harm this robot?
- How morally wrong do you think it would be for someone to steal from this robot?
- To what extent do you think this robot deserves to be treated with compassion and fairness?
- To what extent do you think this robot deserves to be protected from harm?
- If this robot became obsolete, how important would it be to protect this robot?

8.3 Inclusion of the Other in the Self Scale (IOS)

The IOS is a single item, pictorial scale [20]. We have modified the original phrasing. *Please choose the option that best describes how closely you identify with the robot, "Self" being you and "Other" being the robot*:

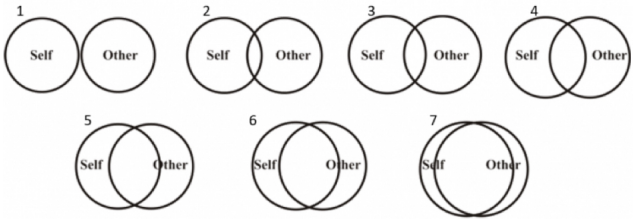

Fig. 4 IOS