What's on Your Virtual Mind? Mind Perception in Human-Agent Negotiations

Minha Lee
m.lee@tue.nl
Human-Technology Interaction
Eindhoven University of Technology
Eindhoven, North Brabant, the Netherlands

Gale Lucas, Johnathan Mell, Emmanuel Johnson, Jonathan Gratch
{lucas,mell,ejohnson,gratch}@ict.usc.edu
Institute for Creative Technologies
University of Southern California, Playa Vista, CA, USA

ABSTRACT

Recent research shows that how we respond to other social actors depends on what sort of mind we ascribe to them. In this article we examine how perceptions of a virtual agent's mind shape behavior in human-agent negotiations. We varied descriptions and communicative behavior of virtual agents on two dimensions according to the mind perception theory: *agency* (cognitive aptitude) and *patiency* (affective aptitude). Participants then engaged in negotiations with the different agents. People scored more points and engaged in shorter negotiations with agents described to be cognitively intelligent, and got lower points and had longer negotiations with agents that were described to be cognitively unintelligent. Accordingly, agents described as having low-agency ended up earning more points than those with high-agency. Within the negotiations themselves, participants sent more happy and surprise emojis and emotionally valenced messages to agents described to be emotional. This high degree of described patiency also affected perceptions of the agent's moral standing and relatability. In short, manipulating the perceived mind of agents affects how people negotiate with them. We discuss these results, which show that agents are perceived not only as social actors, but as intentional actors through negotiations.

CCS CONCEPTS

• **Human-centered computing → HCI theory, concepts and models**; **Empirical studies in HCI**.

KEYWORDS

Virtual agent, mind perception theory, theory of mind, human-agent negotiation, IAGO negotiation platform

ACM Reference Format:
Minha Lee and Gale Lucas, Johnathan Mell, Emmanuel Johnson, Jonathan Gratch. 2019. What's on Your Virtual Mind? Mind Perception in Human-Agent Negotiations. In *ACM International Conference on Intelligent Virtual Agents (IVA '19), July 2–5, 2019, PARIS, France.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3308532.3329465

1 INTRODUCTION

While philosophical explorations on what a mind is and how we perceive it has been an active area of inquiry, how to empirically test our perception of other minds, specifically that of technological entities, is a relatively new project that is becoming more and more relevant with a growing number of digital beings entering our everyday environments (see, e.g., Dennett [16]). The perception of another's mind is especially relevant to human-agent interactions since 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 [18]. According to the mind perception theory (MPT), the mind is assessed on two dimensions: *agency*, which encompasses cognition, and *patiency*, which encompasses emotions [23]. We designed different types of minds of virtual robot negotiators that varied along the dimensions described by MPT in order to see the resulting influence on human interactants' behavior in negotiations. Though whether an agent can realize its own theory of mind as well as others' minds is an important topic [31], far less attention is paid to how agents designed to have minds affect humans they interact with, which is the focus of our paper.

We motivate our research on two grounds: (1) negotiations presume higher order theory of mind reasoning and therefore are fitting for empirically testing minds of various complexities [14] and (2) negotiators' ability to read and influence each others' minds depend more on their mind perception, which can be observed by people's behavior towards agents that are systematically designed to have minds of various orders. Designing intelligent virtual negotiators can push the boundaries of AI via algorithmically-imbued agency [22]. To do so, these systems should be designed with the realization that what drives people's behavior is their perception of a machine's agency; human-agent negotiation research benefits from looking into how an agent's perceived agency impacts human negotiators. Negotiations are a robust context for exploring how artificial minds of machines lead to divergent behaviors in humans.

Previous research showed that the ascribed mind of an agent affects the outcome of a simple game. For example, in the ultimatum game, people gave more money to an agent based on perceived high-agency and high-patiency [11]. However, (unlike [11]) we attentively distinguished between *recognition* of and *control* over emotions (agency) and the ability to *feel* emotion states (patiency) as per MPT [23] in our design; emotional expressions can reflect authentic feelings (patiency) or strategic motives (agency) in negotiations. We thus systematically manipulated artificial minds with agents' descriptions and dialog states to see the resulting influence in a more complex game than the ultimatum game: negotiation.

IVA '19, July 2–5, 2019, PARIS, France

ACM ISBN 978-1-4503-6672-4/19/07…$15.00
https://doi.org/10.1145/3308532.3329465

Negotiations allow people to *interactively* perceive an agent's mind; changes in perceived mind *over time* affect human-agent negotiations, which is the primary exploration underlying our study. Our results demonstrate that people earned higher points against an agent that appeared to have high-agency, and participants did worse against and negotiated longer with an agent purported to have low-agency. Agents that had high-patiency did not directly impact game points, but did affect people's behavior. People sent them more emojis and messages laced with emotional language. People's behavior thus suggests that the perceived mind is critical in seeing agents not just as social actors, but as intentional actors. To frame our study, we present relevant works on theory of mind, mind perception, and negotiation in relation to human-agent interactions, followed by our method and results. We then discuss implications of our findings.

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 [41]. The most commonly attributed mental state is intent [41]. This *intentionality*, or the directedness of mental processing to some end, 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 [15, 16, 30]. In attributing *intent* 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 [15]. One utilizes the theory of one's own mind as a requisite for recognizing other minds, even for non-human entities [18]. People thus have a tendency to be biased towards their own minds as a frame of reference when interacting with humans and agents [31].

Through a course of a shared activity, interactants can form a theory of each other's mind, which helps them find a common ground [30]. 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 [15]. 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 (regulate mental states of self and others) [14]. Social actors can be ascribed minds of zero-order to higher order, yet intentional actors often require higher-order minds, especially in cognitively challenging tasks like negotiation [14].

The 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 [23]. *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 [23]. *Patiency* is defined by having the propensity to feel joy, pleasure, fear, etc. [23]. While we refer to patiency as affective capacity, it also includes biological states like hunger or pain as experiential factors [23]. To note, perceived agency and patiency are not independent of each other [23, 40]. People's assumptions about agency can drive perceptions on patiency, and vice versa; cognition and affect cannot be neatly separated [8]. More broadly, 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 [19]. Competence items evoke agency and warmth items are reminiscent of patiency, though the aims of two scales differ [28]. To generalize, items on agency and competence have more to do with cognitive reasoning while patiency and warmth relate to affective qualities.

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., recipient of a moral/immoral deed [26]. 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 [26]. 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 [29]. Morally relevant acts can therefore influence the perceived intentionality of a moral agent, which makes these acts relevant to negotiations.

Between humans, our relations to others fulfill our "need to belong" [1]. And, how we relate to non-human agents is informed by our human-human interactions [31]. Though people normally grant low intentionality and theory of mind to agents [23, 49] agents can be treated in a human-like social fashion [6, 39]. For example, people are willing to help out a computer that was previously helpful to them [20], punish those agents that betray them [35], and grant personality traits to computers based on text-based chats [37]. 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 agent than a human, for a machine is not seen to be judgmental like a human [33, 36]. At the same time, when 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 [3, 17, 44]. In sum, people may have preconceived beliefs about agents having low-order minds compared to them, yet by treating agents as social actors, they apply certain social stereotypes such as gender or race-related biases towards agents that have human-like appearances.

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 [25]. 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 [32]. 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 agents do not have an intentional stance can be overridden based on manipulated context [32]. One context that is ripe for manipulating the perceived mind of an agent is negotiation.

2.2 Human-human vs. human-agent 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 [7]. Also, negotiation may involve joint decision-making with others when one cannot fulfill one's interests and/or goals without their involvement [47]. The concept of fairness as a component of morality [21] can be estimated in negotiations through measurable components, such as negotiation outcomes (e.g., points per player) or process measures (e.g., how many offers a player made to the opponent) [47]. Thus, self- and other- regard is inherent to negotiations, encompassing complex socio-psychological processes [46]. Negotiations therefore involve theory of mind reasoning; negotiators have to reason about each others' intentions, trade-offs, and outcomes as a cognitively taxing process [22]. Especially if negotiators have to cooperate and compete, such as during a mixed-motive negotiation, they often rely on a higher-theory of mind [14]. 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 [9] or human negotiators [4, 38]. An agent's expressed anger, regret, or joy (both facial and textual expressions) influence how human opponents play against it [9], extending the view that emotions in human-human negotiations reveal strategic intentions and influence outcomes [4, 38]. To add, priming people's belief about the negotiation (emphasizing cooperation vs. exploitation at the start) impacts human-agent negotiations [13], echoing how framing of a game in itself for human-human negotiations results in divergent outcomes [42]. Increasingly, agents are capable of using complex human-like strategies in negotiation, and the perceived gap between humans and agents may continue to shrink [2].

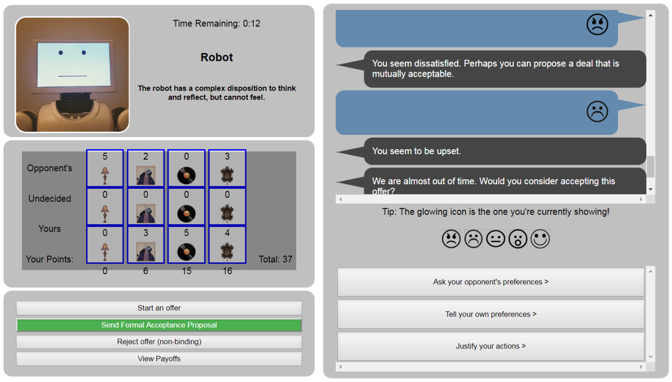
However, people still do have preconceptions about agents' lack of human-like mind in many negotiation scenarios. Specifically, a human opponent is granted agency by default, but a machine's agency can be independent of or dependent on a human actor; the belief about the agent (autonomous vs. human-controlled agent) can result in different tactics adopted by human players [9, 12]. 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 [14]. In simple games, people are likely to allocate more goods to agents that have a high degree of human-like mind, based on perceived agency and patiency [11]. Thus, an agent's mind and a human player's perception of an agent's mind are crucial to how their negotiation unfolds [11]. We focused on the latter, the perception of an agent's mind through negotiations as an interactive context.

2.3 Research question

The research question of our study is as follows. *In what ways do manipulated agency and patiency via dialog states and descriptions of a virtual 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 patiency would increase 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 [11]. In addition, we were interested in whether or not MPT dimensions relate to competence and warmth (corresponding SCM dimensions), as well as participants' judgment of the agent's moral standing and relatability. We looked at adjacent concepts such as SCM and moral standing to more holistically understand how minds of artificial agents are viewed.

3 METHODS

3.1 Design



Time Remaining: 0:12

Robot

The robot has a complex disposition to think and reflect, but cannot feel.

Opponent's	5	2	0	3	
Undecided	0	0	0	0	
Yours	0	3	5	4	
Your Points:	0	6	15	16	Total: 37

You seem dissatisfied. Perhaps you can propose a deal that is mutually acceptable.

You seem to be upset.

We are almost out of time. Would you consider accepting this offer?

Tip: The glowing icon is the one you're currently showing!

Start an offer
Send Formal Acceptance Proposal
Reject offer (non-binding)
View Payoffs
Ask your opponent's preferences >
Tell your own preferences >
Justify your actions >

Figure 1: Negotiation interface

Our agent was a virtual robot that was simple in appearance (Figure 1), without any gender, race, or other highly anthropomorphic traits that may trigger people's biases [3, 17, 44], which helped to drive the perception of its mind based on its behavior rather than its looks. We used a configurable negotiation platform (IAGO) that

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

allows 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) [34].

We employed a 2X2 between-participants factorial design of High vs. Low Agency and High vs. Low Patiency 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 1) during the experiment. We also modified dialog states of the agent, i.e., how it “talked” (Table 1 lists excerpts). The sentence structure of our descriptions was modeled after previous research on moral standing [29, 40]. We used the items of the MPT scale [23] to construct the content of the descriptions and the dialogs. 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 1). 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 virtual 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 [11], 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 [11]).

We piloted our manipulation (descriptions and dialogs) before the main experiment. After reading agency dialogs, pilot study participants assigned them with both perceived agency F(1, 308) = 4.95, p = .027, and patiency F(1, 308) = 7.39, p = .007, and there was no interaction F(1, 308) = .348, p = .556. Patiency dialogs resulted in significance for perceived patiency F(1, 308) = 12.20, p = .001 and non-significance for perceived agency F(1, 308) = .783, p = .377, with no interaction F(1, 308) = .104, p = .748. Agency descriptions were assigned perceived agency F(1, 249) = 42.09, p = .00 and perceived patiency F(1, 249) = 29.98, p = .00 and no interaction was found F(1, 249) = 1.31, p = .254. The patiency descriptions were significant for both perceived agency F(1, 249) = 17.49, p = .00 and perceived patiency F(1, 249) = 59.86, p = .00 with a significant interaction F(1, 249) = 5.78, p = .017. We concluded that perceived agency and patiency of descriptions and dialogs were significant for the corresponding dimensions, even if they were not perfectly orthogonal. In fact, Gray et al.’s original MPT data showed a high correlation between two dimensions at r(11) = .90, p < .001 [23, 40]. We therefore proceeded with using dialogs and descriptions in the main experiment based on our pilot study.

As a reminder, only dialogs and descriptions differed between agents (Table 1). Also in all negotiations, there were 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 [34]; the agent adjusted its offers if the participant communicated his/her preferences, and strove for fair offers, while rejecting unfair deals. The agent did not know participants’ preferences, but assumed an *integrative* structure. The 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).

	Clocks	Records	Paintings	Lamps	Points
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: Starting offer as item * points = total points

3.2 Participants

226 participants residing in the U.S. were recruited on Amazon Mechanical Turk. We had 135 men (59.7%), 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). As for other participants, 17 were between 18-24 years of age, 47 were between 35-44 years of age, 26 were between 45-54 years of age, 13 were between 55-64 years of age, and 2 were between 65 and 74 years of age. 87.2% identified as White/Caucasian (197 participants), and 10 as Black/African Americans, 6 as Native Americans/American Indians, 14 as Asian Americans, and 1 identified as Black/African and Asian American. 60.6% had some college education or above.

3.3 Procedure and measurements

Participants got a link to the survey on Amazon Mechanical Turk, which 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 shows four conditions) 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 actual 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 1). Upon completion of the negotiation, participants finished the second part of the survey of our measurements.

We deployed following measurements: MPT (agency and patiency) [23], SCM (competence and warmth) [19], the moral standing scale [29, 40], emotion states [10, 27, 45], the moral identity questionnaire [5], the inclusion of other in self (IOS) scale [1] as a measure of relatability. We asked additional questions on whether or not participants made concessions to the agent and if the agent did anything unexpected. We only report relevant measures in our results. 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 RESULTS

4.1 Manipulation check

Both of our experimental manipulations affected perceived agency; that is, there was both a significant main effect of agency (F(1, 222) = 35.68, p < .001) and a significant main effect of patiency (F(1, 222) = 53.42, p < .001) on perceived agency, whereas the interaction between agency and patiency did not approach significance (F(1, 222) = .60, p = .44). Serving as a manipulation check, participants perceived lower agency for the agent that could purportedly not reason (M = 2.89, SE = .14) than when the agent was described as being able to reason (M = 4.01, SE = .13). However, participants also rated the agent as lower in agency when it could not feel (M = 2.77, SE = .13) than when the agent was described as being able to feel (M = 4.14, SE = .13). In contrast, only manipulated patiency significantly affected perceived patiency (F(1, 222) = 71.24, p < .001); the effect of agency on perceived patiency only approached significance (F(1, 222) = 2.57, p = .11), and the interaction did not approach significance (F(1, 222) = .001, p = .99). Participants rated the agent as lower in patiency when it could not feel (M = 1.88, SE = .13) than when the agent is described as being able to feel (M = 3.44, SE = .13). The manipulations thus showed a same trend as in our pilot study. As aforementioned, agency and patiency dimensions were highly correlated in the original MPT study [23, 40].

4.2 Main analysis

We next looked into negotiation outcomes. For user points, there was a significant main effect of agency (F(1, 143) = 4.35, p = .04); 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); 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). 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). However, the effect of agency on initial offer was not significant F(1, 143) = .49, p = .49); no other effects reached significance (Fs < 2.7, ps > .10).

Process measures capture *how* participants played against the agent and thus are also important elements of negotiations. There was a marginally significant effect of agency on game end time (F(1, 143) = 3.62, p = .059); participants took longer if the agent was described as not being able to reason (M = 296.88, SE = 13.36) than 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). 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); 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). 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 UI) 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) 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 this message "I gave a little here; you give a little next time" (F(1, 143) = 5.18, p = .02). 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). No other effects reached significance (Fs < .85, ps > .36). No other effects for any other message options reached significance (Fs < 2.17, ps > .14). There were other process related measures that were relevant, but not significant: the effect of agency on number of times users made offers, accepted offers, declared their preferred items, posed queries, and sent messages to the agent.

4.3 Exploratory analysis

We examined the impact of agency and patiency dimensions on competence, warmth, IOS (relatability), and moral standing. Both of our experimental manipulations affected perceived competence, meaning that there was both a significant main effect of agency (F(1, 222) = 1.30, p = .002) and a significant main effect of patiency (F(1, 222) = 19.20, p < .0001) on perceived competence, whereas the interaction between agency and patiency did not approach significance (F(1, 222) = .08, p = .77). Participants perceived the agent that purportedly could not reason as lower in competence (M = 3.13, SE = .09) than when the agent was described as being able to reason (M = 3.52, SE = .08). However, participants also rated the agent as lower in competence when it could not feel (M = 3.06, SE = .09) than when the agent is described as being able to feel (M = 3.59, SE = .09). Likewise, both of our experimental manipulations affected perceived warmth; there was a significant main effect of agency (F(1, 222) = 6.71, p = .01) and a significant main effect of patiency (F(1, 222) = 4.06, p < .001) on perceived warmth, whereas the interaction between agency and patiency did not approach significance (F(1, 222) = .03, p = .86). Participants rated the agent as lower in warmth when it could not feel (M = 2.33, SE = .10) than when the agent was described as being able to feel (M = 3.20, SE = .10). However, participants also perceived the agent that could purportedly not reason as lower in warmth (M = 2.59, SE = .10) than when the agent was described as being able to reason (M = 2.95, SE = .09).

Only manipulated patiency significantly affected psychological distance (IOS) from the agent (F(1, 222) = 29.1, p = .002); the effect of agency on IOS and the interaction did not reach significance (Fs < 1.16, ps > .28). Participants reported that they identified with the agent more when the agent 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); 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 DISCUSSION

Our focus was on how people played against virtual agents that were designed to have divergent minds. The results on negotiation outcomes and processes, two paradigmatic measures in negotiation research [47], overall did not align with our hypotheses. Unlike previous findings that showed both agency and patiency to matter in game outcomes, e.g., the ultimatum game [11], we only noted a significant effect of agency, in a different direction than anticipated. Agency did not contribute to greater joint outcomes. Rather, participants scored higher and played shorter games when the agent was portrayed to have high-agency, and they scored lower via longer negotiations with low-agency agents. Joint outcomes hence were insignificant. Participants also did not seek more information from the agent with high-agency. While game points did not hinge on patiency, people did utilize more emotive messages and emojis with an agent with high-patiency.

Emotions matter in how people negotiate [4, 48]. Participants sent emojis and emotionally relevant messages to a high-patiency agent which was described to be affective. Yet, patiency in itself did not affect negotiation results. Rather, it only affected a few process measures. Potentially there was more "noise" for people to interpret when they 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. Additional analyses demonstrated that patiency played a large part in attributing moral standing and relatability (IOS) to an agent. But, how much participants identified with an agent and to what degree they thought the agent had moral standing did not align with how well they played against it.

When an agent that was described to be less cognitively intelligent (low-agency) interacted with participants in a cognitively taxing task (negotiation over goods), 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. Human-like qualities such as an agent's emotions, moral standing or relatability are in essence, distracting points when it comes to game mechanics, yet these distractors could have (wrongly) gained greater traction as part and parcel of the game, especially since harm salience regarding a moral patient increases with time pressure [24]. Agents demonstrating low-agency traits can seem inconsistent with highly agentic tasks like negotiation they partake in, which can mean that people do poorly when they cannot conclude to what degree an agent has a perceived mind.

Negotiations serve as a context for adjusting preconceptions on technological entities' minds. We buttress this on three premises. First, people have preconceived beliefs about virtual agents' minds; agents are seen to have low-order theory of minds [23, 49] (at least presently) even if people interact with agents socially [6, 39]. Second, the perceived mind of an agent can be adjusted, be it through patiency (affective richness [25]) or agency (behavioral intentionality [32]). Third, negotiations require cognitively effortful participation that involves theory of mind reasoning [14, 22], especially when it comes to mixed-motive negotiations [14, 43]. Through negotiations, an agent's behavioral intentionality can be called into question, providing people opportunities to reformulate an agent's degree of conferred mind via agency.

In our experiment, participants' perception of how agents negotiated was the main point, not how agents actually negotiated. All agents appeared to calculatively negotiate, but not with any sophisticated AI; their offer strategies were not affected by emotional communications from players. Agents, however, adjusted their offers if participants communicated about preferences [34]. Participants were incentivized to do well, i.e., extra monetary compensation for the best player, and were purposefully not provided information on the agent's preferences; communication was necessary to cooperate and compete as an interaction paradigm.

Our approach alluded to differing degrees of agency and patiency *over time* through descriptions (pre- and in-game- manipulation), dialogs (in-game manipulation), facial expressions (in-game manipulation) and the experimental context (negotiation) in itself was suggestive of agency. The common belief that technological agents have low-agency and patiency [23, 49] can be solidified or not called into question in an interaction, unless people have reasons to adjust their beliefs, e.g., manipulated behavioral intentionality based on an agent's action [32]. Thus, the disjointed nature between the 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". The high-agency agent could have been more straightforward to "read" for participants. It negotiated, talked, and was described as if it could have a higher-order mind. But, this manipulation in itself would not necessarily overturn people's belief that an agent has low-agency and low-patiency compared to them.

An agent that calculates offers or talks in an overly logical fashion ("Spock-like") is not granted high-agency by default. An agent that smiles or frowns is not granted high-patiency by default. These manipulations alone do not greatly challenge people's notion that they are interacting with a mere machine. Our high-agency agent did poorly against participants that do have a higher degree of mind. Our low-agency agent did well against participants, over time. When we cannot easily guess what an agent desires or intends to do, i.e., predict its intentional stance [15], we can exercise our higher degree theory of mind, investigating and questioning the bias we hold as a fact—the inability of technology to have a human-like mind.

We note that human-agent negotiations can greatly aid research on mind perception. There are potential areas for broadening future research, such as processes that challenge people's steadfast beliefs about forms of technology. Specifically, we recommend a thorough look at how the mind is judged on *continuous* dimensions of agency and patiency as there are tiered degrees of having a mind. If so, to treat agency and patiency as *discrete* dimensions in designing virtual agents, e.g., agency as random vs. intentional actions, and patiency as facial expressions vs. no expressions in a simple game [11], or to test these dimensions without interactive conversations, e.g., computational approaches to modelling agents' minds without conversational assessments [14], leaves much out when the combination of mind perception and human-agent negotiation as conjoined research areas can richly inform each other.

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. MPT focuses more on the latter case. Minds of various beings were judged through a survey; it is about people's pre-existing beliefs at a single point in time [23]. The novelty of our study is that people seem to be revising their opinion of the agent's perceived mind; the human attribution of seeing a mind in a machine may be misguided, but people can question their own beliefs through an interaction. 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 negotiations, 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 virtual agents that are truly intentional actors like humans. But, the degree to which agents are perceived to have agency and patiency can be observed via human-agent interactions. Through negotiations, we caught 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. The results show that 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 (emojis, messages). People also granted higher moral standing and related more to the agent when it was described to have patiency. We conjecture that a virtual agent that sends unclear or mismatched signals that people have to interpret during a complex interaction like negotiation can lead them to reconsider manipulated mind perception dimensions. What we can conclude is that in attempting to comprehend a virtual negotiator's "mind", people react to its rational and emotional capacities in divergent ways, leading to noticeable differences in how they behave.

7 ACKNOWLEDGMENTS

This work is supported by the Air Force Office of Scientific Research, under Grant FA9550-18-1-0182, and the US Army. The content does not necessarily reflect the position or the policy of any Government, and no official endorsement should be inferred.

REFERENCES

[1] Arthur Aron, Elaine N Aron, and Danny Smollan. 1992. Inclusion of other in the self scale and the structure of interpersonal closeness. *Journal of personality and social psychology* 63, 4 (1992), 596.

[2] Tim Baarslag, Michael Kaisers, Enrico Gerding, Catholijn M. Jonker, and Jonathan Gratch. 2017. When will negotiation agents be able to represent us? The challenges and opportunities for autonomous negotiators.. In *International Joint Conference on Artificial Intelligence.* 4684–4690.

[3] Jeremy N Bailenson, Jim Blascovich, Andrew C Beall, and Jack M Loomis. 2003. Interpersonal distance in immersive virtual environments. *Personality and Social Psychology Bulletin* 29, 7 (2003), 819–833.

[4] Bruce Barry, Ingrid Smithey Fulmer, Gerben A Van Kleef, et al. 2004. I laughed, I cried, I settled: The role of emotion in negotiation. *The handbook of negotiation and culture* (2004), 71–94.

[5] Jessica E Black and William M Reynolds. 2016. Development, reliability, and validity of the Moral Identity Questionnaire. *Personality and Individual Differences* 97 (2016), 120–129.

[6] Jim Blascovich, Jack Loomis, Andrew C Beall, Kimberly R Swinth, Crystal L Hoyt, and Jeremy N Bailenson. 2002. Immersive virtual environment technology as a methodological tool for social psychology. *Psychological Inquiry* 13, 2 (2002), 103–124.

[7] Peter J Carnevale and Dean G Pruitt. 1992. Negotiation and mediation. *Annual review of psychology* 43, 1 (1992), 531–582.

[8] Antonio R Damasio. 2006. *Descartes' error.* Random House.

[9] Celso M de Melo, Peter J Carnevale, Stephen J Read, and Jonathan Gratch. 2014. Reading people's minds from emotion expressions in interdependent decision making. *Journal of personality and social psychology* 106, 1 (2014), 73.

[10] Celso M de Melo and Jonathan Gratch. 2015. People show envy, not guilt, when making decisions with machines. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII).* IEEE, 315–321.

[11] Celso M De Melo, Jonathan Gratch, and Peter J Carnevale. 2014. The importance of cognition and affect for artificially intelligent decision makers. In *Twenty-Eighth AAAI Conference on Artificial Intelligence.* 336–342.

[12] Celso M de Melo, Jonathan Gratch, and Peter J Carnevale. 2015. Humans versus computers: Impact of emotion expressions on people's decision making. *IEEE Transactions on Affective Computing* 6, 2 (2015), 127–136.

[13] Celso M de Melo, Peter Khooshabeh, Ori Amir, and Jonathan Gratch. 2018. Shaping Cooperation between Humans and Agents with Emotion Expressions and Framing. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems.* International Foundation for Autonomous Agents and Multiagent Systems, 2224–2226.

[14] Harmen de Weerd, Rineke Verbrugge, and Bart Verheij. 2017. Negotiating with other minds: the role of recursive theory of mind in negotiation with incomplete information. *Autonomous Agents and Multi-Agent Systems* 31, 2 (2017), 250–287.

[15] Daniel Dennett. 1989. *The intentional stance.* MIT press.

[16] Daniel Dennett. 2008. *Kinds of minds: Toward an understanding of consciousness.* Basic Books.

[17] Ron Dotsch and Daniël HJ Wigboldus. 2008. Virtual prejudice. *Journal of experimental social psychology* 44, 4 (2008), 1194–1198.

[18] Nicholas Epley, Adam Waytz, and John T Cacioppo. 2007. On seeing human: a three-factor theory of anthropomorphism. *Psychological review* 114, 4 (2007), 864.

[19] Susan T Fiske, Amy JC Cuddy, Peter Glick, and Jun Xu. 2002. A model of (often mixed) stereotype content: competence and warmth respectively follow from perceived status and competition. *Journal of personality and social psychology* 82, 6 (2002), 878.

[20] BJ Fogg and Clifford Nass. 1997. How users reciprocate to computers: an experiment that demonstrates behavior change. In *CHI'97 Extended Abstracts on Human Factors in Computing Systems (CHI EA '97).* ACM, New York, NY, USA, 331–332. https://doi.org/10.1145/1120212.1120419

[21] Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P Wojcik, and Peter H Ditto. 2013. Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in experimental social psychology.* Vol. 47. Elsevier, 55–130.

[22] Jonathan Gratch, David DeVault, Gale M Lucas, and Stacy Marsella. 2015. Negotiation as a challenge problem for virtual humans. In *International Conference on Intelligent Virtual Agents.* Springer, 201–215.

[23] Heather M Gray, Kurt Gray, and Daniel M Wegner. 2007. Dimensions of mind perception. *science* 315, 5812 (2007), 619–619.

[24] Kurt Gray, Chelsea Schein, and Adrian F Ward. 2014. The myth of harmless wrongs in moral cognition: Automatic dyadic completion from sin to suffering. *Journal of Experimental Psychology: General* 143, 4 (2014), 1600.

[25] Kurt Gray and Daniel M Wegner. 2012. Feeling robots and human zombies: Mind perception and the uncanny valley. *Cognition* 125, 1 (2012), 125–130.

[26] Kurt Gray, Liane Young, and Adam Waytz. 2012. Mind perception is the essence of morality. *Psychological inquiry* 23, 2 (2012), 101–124.

[27] Jonathan Haidt et al. 2003. The moral emotions. *Handbook of affective sciences* 11, 2003 (2003), 852–870.

[28] Nick Haslam. 2012. Morality, mind, and humanness. *Psychological Inquiry* 23, 2 (2012), 172–174.

[29] Mansur Khamitov, Jeff D Rotman, and Jared Piazza. 2016. Perceiving the agency of harmful agents: A test of dehumanization versus moral typecasting accounts. *Cognition* 146 (2016), 33–47.

[30] Nicole C Krämer. 2008. Theory of mind as a theoretical prerequisite to model communication with virtual humans. In *Modeling communication with robots and virtual humans.* Springer, 222–240.

[31] Nicole C Krämer, Astrid von der Pütten, and Sabrina Eimler. 2012. Human-agent and human-robot interaction theory: similarities to and differences from human-human interaction. In *Human-computer interaction: The agency perspective.* Springer, 215–240.

[32] Daniel T Levin, Stephen S Killingsworth, Megan M Saylor, Stephen M Gordon, and Kazuhiko Kawamura. 2013. Tests of concepts about different kinds of minds: Predictions about the behavior of computers, robots, and people. *Human–Computer Interaction* 28, 2 (2013), 161–191.

[33] Gale M Lucas, Jonathan Gratch, Aisha King, and Louis-Philippe Morency. 2014. It's only a computer: Virtual humans increase willingness to disclose. *Computers in Human Behavior* 37 (2014), 94–100.

[34] Johnathan Mell and Jonathan Gratch. 2017. Grumpy & Pinocchio: answering human-agent negotiation questions through realistic agent design. In *Proceedings of the 16th Conference on Autonomous Agents and Multiagent Systems.* International Foundation for Autonomous Agents and Multiagent Systems, 401–409.

[35] Johnathan Mell, Gale M. Lucas, and Jonathan Gratch. 2015. An Effective Conversation Tactic for Creating Value over Repeated Negotiations. In *International Conference on Autonomous Agents and Multiagent Systems.* International Foundation for Autonomous Agents and Multiagent Systems, 1567–1576.

[36] Johnathan Mell, Gale M. Lucas, and Jonathan Gratch. 2017. Prestige Questions, Online Agents, and Gender-Driven Differences in Disclosure. In *International Conference on Intelligent Virtual Agents.* Springer, 273—282.

[37] Youngme Moon and Clifford Nass. 1996. How "real" are computer personalities? Psychological responses to personality types in human-computer interaction. *Communication Research* 23, 6 (1996), 651–674. https://doi.org/10.1177/009365096023006002

[38] Michael W Morris and Dacher Keltner. 1999. How Emotions Work: An Analysis of the Social Functions of Emotional Expression in Negotiation, Vol. 11. in B.M. Staw R.I. Sutton (Eds.), Research in organizational behavior. Amsterdam: JAI, 1–50.

[39] Clifford Nass, Jonathan Steuer, and Ellen R Tauber. 1994. Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems.* ACM, 72–78.

[40] Jared Piazza, Justin F Landy, and Geoffrey P Goodwin. 2014. Cruel nature: Harmfulness as an important, overlooked dimension in judgments of moral standing. *Cognition* 131, 1 (2014), 108–124.

[41] David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences* 1, 4 (1978), 515–526.

[42] Dean G Pruitt. 1967. Reward structure and cooperation: The decomposed Prisoner's Dilemma game. *Journal of Personality and Social Psychology* 7, 1p1 (1967), 21.

[43] Dean G Pruitt and Melvin J Kimmel. 1977. Twenty years of experimental gaming: Critique, synthesis, and suggestions for the future. *Annual review of psychology* 28, 1 (1977), 363–392.

[44] Mikey Siegel, Cynthia Breazeal, and Michael I Norton. 2009. Persuasive robotics: The influence of robot gender on human behavior. In *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems.* IEEE, 2563–2568.

[45] Eva EA Skoe, Nancy Eisenberg, and Amanda Cumberland. 2002. The role of reported emotion in real-life and hypothetical moral dilemmas. *Personality and Social Psychology Bulletin* 28, 7 (2002), 962–973.

[46] Leigh Thompson. 1990. Negotiation behavior and outcomes: Empirical evidence and theoretical issues. *Psychological bulletin* 108, 3 (1990), 515.

[47] Leigh L Thompson, Jiunwen Wang, and Brian C Gunia. 2010. Negotiation. *Annual review of psychology* 61 (2010), 491–515.

[48] Gerben A Van Kleef, Carsten KW De Dreu, and Antony SR Manstead. 2004. The interpersonal effects of emotions in negotiations: a motivated information processing approach. *Journal of personality and social psychology* 87, 4 (2004), 510.

[49] Adam Waytz, Kurt Gray, Nicholas Epley, and Daniel M Wegner. 2010. Causes and consequences of mind perception. *Trends in cognitive sciences* 14, 8 (2010), 383–388.