Of Like Mind: The (Mostly) Similar Mentalizing of Robots and Humans

Jaime Banks

College of Media & Communication, Texas Tech University

Mentalizing is the process of inferencing others' mental states and it contributes to an inferential system known as Theory of Mind (ToM)—a system that is critical to human interactions as it facilitates sense-making and the prediction of future behaviors. As technological agents like social robots increasingly exhibit hallmarks of intellectual and social agency—and are increasingly integrated into contemporary social life—it is not yet fully understood whether humans hold ToM for such agents. To build on extant research in this domain, five canonical tests that signal implicit mentalizing (white lie detection, intention inferencing, facial affect interpretation, vocal affect interpretation, and false-belief detection) were conducted for an agent (anthropomorphic or machinic robots, or a human) through video-presented (Study 1) and physically copresent interactions (Study 2). Findings suggest that mentalizing tendencies for robots and humans are more alike than different; however, the use of nonliteral language, copresent interactivity, and reliance on agent-class heuristics may reduce tendencies to mentalize robots.

Keywords: mentalizing, social robots, heuristics, social presence, human–machine communication

Supplemental materials: <https://doi.org/10.1037/tmb0000025#supplemental-materials>

R2-D2—the canister-shaped droid of *Star Wars* fame—does not speak aside from clicks and beeps, has no discernible face aside but for a single photoreceptor, and does not gesture except to rotate its top, tilt forward and back, and manipulate various accessories. Yet through these minimal cues, audiences can make inferences about the robot's state of mind in context, ranging from fear and surprise at being attacked to determination and focus in engineering an escape. Such internal-state inferences are illustrative of *mentalizing*: A metarepresentational process by which people make sense of each other by inferring others' mental states (cf., Leslie, 1987) toward social-cognitive schemes known as Theory of Mind (ToM). ToM (though often defined differently across various fields; Apperly, 2012) is usually characterized as a system of aggregated inferences about another's mental states; the comparison of those

inferences against understandings of one's own mental states gives rise to predictions about the other's beliefs, attitudes, and intentions (Premack & Woodruff, 1978).

Although a rich body of work details the dynamics of ToM as core to human–human interaction (see Goldman, 2012), it is not yet fully understood whether and how mentalizing may function in human–machine communication as technological agents increasingly mirror human behavior and capabilities. This is an important consideration as social robots and other technological agents develop in sophistication and are increasingly integrated into society at large (Cross et al., 2019), especially as mind perception may be a point of human affinity and comprehension (Anjomshoae et al., 2019) or a threat to human distinctiveness (Ferrari et al., 2016). Understanding fundamental social cognitions in human–machine communication is the

Action Editor: Danielle S. McNamara was the action editor for this article.

ORCID iD: Jaime Banks <https://orcid.org/0000-0002-7598-4337>

The author has no conflicts of interest to report. This material is based upon work supported by the Air Force Office of Scientific Research under award number FA9550-19-1-0006. The author gratefully acknowledges the assistance of the following people for their support in the production of this article: Ambrosia Luzius (lab data collection), Zack Hlatky (lab data collection and data coding), Cloris Chen (lab data collection and data coding), Madison Wedge (lab data collection and survey video production), Shianne Ferrell (lab data collection and survey video production), Nicholas David Bowman (robot mechanics and analysis consulting), and Zachary Stiffler (robot animation). A portion of this work was conducted at West Virginia University, Department of Communication Studies, and the author thanks WVU for its support. Study materials (including stimuli, data, and analysis documentation) are available at: <https://osf.io/9r73y>

Open Science Disclosures:

The data are available at <https://osf.io/9r73y>

The experiment materials are available at <https://osf.io/9r73y>

Open Access License: This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (CC-BY-NC-ND). This license permits copying and redistributing the work in any medium or format for noncommercial use provided the original authors and source are credited and a link to the license is included in attribution. No derivative works are permitted under this license.

Disclaimer: Interactive content is included in the online version of this article.

Contact Information: Correspondence concerning this article should be addressed to Dr. Jaime Banks, Texas Tech University, Box 43082, Lubbock, TX 79409, USA. Email: j.banks@ttu.edu

foundation for human-centric development and integration (cf., Henschel et al., 2020). The present studies offer conceptual replications and extensions of recent work to address (a) whether and how people may engage in mentalistic sense-making of humanoid robots' behaviors; (b) whether implicit and explicit indicators of mentalizing point to (mis)aligned preconscious and conscious sense-making; and (c) whether mentalistic processing is linked to trust outcomes. Findings comport with and depart from prior findings, indicating that people may similarly mentalize robots and humans except when elaborative processing is required; however, there are no clear links between indicators of implicit and explicit mentalizing and neither form is linked with trust outcomes.

Review of Literature

ToM in Human–Robot Interaction

The human ToM inferential system facilitates a recognition that another's knowledge, attitudes, and experiences are different from one's own (Premack & Woodruff, 1978), supporting relational processes and outcomes such as prosocial tendencies (i.e., helping and comforting; Imuta et al., 2016) and problematic behaviors like bullying (Smith, 2017) as well as more general moral decision-making (Bzdok et al., 2012). Among neurotypical individuals, it emerges in infancy and develops through early childhood. This development is debated to unfold either through innate theorizing tendencies and social learning (Gopnik, 1998; Leslie, 1987) or internally simulating others' cognitive activities (Gordon, 1986; see Carruthers & Smith, 1996 for a review of the debate).

Just as mentalizing processes are central to human–human interaction, evidence suggests they play a role in human–machine communication. Humans unconsciously and automatically engage technologies as social agents when they exhibit self-similarity (Lee et al., 2006) in both appearance and movement (Bartneck et al., 2009) and in holistic social behaviors (Epley et al., 2007). Preconscious processing of social information delivered by robotic agents can result in apparent mentalizing of robots similar to mentalizing of humans (Banks, 2020c). Extant scholarship suggests that as an agent's appearance or behavior increases in human-likeness (Eyssel & Pfundmair, 2015) or in apparent benevolence (Tanibe et al., 2017) or is portrayed as grounded in experience (Fu et al., 2020), the more likely an observer will draw on human-native, mind-attributive mental models for interpreting its behavior. Underpinning those visual and relational dynamics, mentalizing relies on preconscious processes to engender baseline “social attunement,” or the multiplex social-cognitive orientation of agents during an interaction (Perez-Osario & Wykowska, 2020, para. 1). Although social attunement promoting mentalizing may promote acceptance of robots (Wiese et al., 2017), inferences of a capacity to think and feel may lead to discomforts (Appel et al., 2020).

However, the social-cognitive mechanisms promoting mentalizing of robots may break down under some conditions: When elaboration on the agent's behavior makes salient the machine-ontological status of the robot (Banks, 2020c), when such salience activates mental models including expectations of mindlessness of machines (cf., Perez-Osario & Wykowska, 2020; Thellman et al., 2020), or when a robot's social cues are uninterpretable (Banks, 2020c). Moreover, the manner in which mentalizing is inferred by researchers is of particular importance, as verbal and direct metrics

(i.e., self-reports) do not always comport with nonverbal and indirect measures (i.e., behavioral indicators; Banks, 2020c; Thellman et al., 2020), likely because they are, respectively, associated with explicit/logical and implicit/intuitive processes, respectively (Takahashi et al., 2013).

The state of this science is, however, unfortunately limited as a matter of methodological shortcomings (see Thellman & Ziemke, 2019). With few exceptions, much of the work on perceived robot mind relies on explicit self-report scales (principally Gray and colleagues' [2007] two-factor minded experience/agency metric), on survey-only procedures with no physical robot presence, on resource-intensive psychophysiological measures, on narrow scope (a single-mentalizing signal), on unrealistic (crudely drawn) stimuli, and/or on designs that do not include a human control. These limitations present conspicuous gaps in knowledge about robot mentalizing processes related to its theorized implicit, multiplex, and applied nature (Byom & Mutlu, 2013), the role of interaction and copresence versus mediated observation (cf., Schreiner et al., 2017), and how human–human interactions serve as experiential grounds for human–robot interactions (Spence et al., 2014).

Questions of Ontological Categorization, Explicitness, and Trust

Both implicit and explicit mentalizing processes contribute to the ToM system. Implicit processes are subconscious, automatic, spontaneous, nonconceptual, and procedural; explicit processes are conscious, controlled, conceptual, and declarative (Low & Perner, 2012). That is, people process social cues both (a) automatically and relatively abstractly and then (b) purposively and concretely (i.e., as a dual-process phenomenon; cf., Chaiken & Trope, 1999). Social-cognitive processes such as mentalizing are understood to be at least initially implicit (Greenwald & Banaji, 1995), initiated as prototypical human features and cues are interpreted (Khalid et al., 2016). Because of this implicitness, implicit measures are favored because explicit reporting suffers from inability to introspect and from error as a result of implicit/explicit process independence (Hofmann et al., 2005). To extend the current work relying primarily on explicit measures, I engage implicit measures to ask (RQ1): Do humans exhibit implicit mentalizing for social robots as they do for other humans?

The value of implicit evaluations acknowledged, explicit assessments of robot mental capacity are also important because conscious ascription of machine mind may be performed when an agent becomes important enough (through relatedness-need satisfaction, relationship tenure, or interdependence) that it is insufficient to think of the robot merely as a machine (Wiese et al., 2019). Whereas implicit indicators signal mental processes, explicit indicators signal experience; the former illuminates attitude-based reactions that may not be accessible for conscious consideration and the latter points to sense-making grounded in cultural knowledge (Nosek, 2007). Both are legitimate indicators of how people may interpret mental states in robots, but represent different functional paths that may or may not align: Merely because one can infer the states of others does not mean they necessarily will (Keysar et al., 2003). These distinct mechanisms and potential dissociations are particularly important for mentalizing's roles in social processes. Explicit mind ascription may signal in-group status and social engagement (Deska et al., 2018), such that understanding its antecedents and boundary

conditions may advance interventions to foster technology acceptance. That same understanding could help define the experiential boundaries between humans and machines to support an explainable artificial intelligence (XAI) objective of transparent, mechanistic explanations for robot behavior over anthropomorphic (i.e., mentalistic) explanations (see Doran et al., 2017). In light of importance of both implicit and explicit mentalizing indicators of their potential divergence, I ask (RQ2) the following: What relationship exists among implicit and explicit indicators of mentalizing for social robots?

Regarding the potential for implicit and explicit social cognitions to be differentially associated with practical relational outcomes, trust reactions are of particular importance because mentalizing is linked to trust through co-operation and reciprocity (McCabe et al., 2001) and through social affiliation (Prochazkova et al., 2018). Furthermore, trust may form through the same folk-psychological social learning as may ToM (Clément et al., 2004) and technological agents with more human-like cues tend to activate more mentalistic responses aligning with neural correlates of trust (Lee, 2018). Regarding the anthropomorphic features of some social robots, however, it may also be that machine delivery of human-like cues could diminish trust through the uncanny crossing of agent-class boundaries (Gray & Wegner, 2012). Thus, I ask (RQ3): (How) are (a) implicit and (b) explicit mentalizing associated with trust in a social robot?

Research Approach

To address the posed questions, two studies were conducted; each conceptually replicated and extended the author's prior work (Banks, 2020c) by addressing shortcomings in that work and in other investigations of perceived robot mind. That original study drew on both implicit and explicit mentalizing indicators (similarly addressing RQ1 and RQ2) for three small social robots and a human, finding that implicit indicators were generally consistent across the agents as long as the social cues delivered by robots were interpretable and similar to those given by humans; there was no association between implicit and explicit indicators. In that work, some stimuli were interpreted by participants in unexpected ways, limiting some claims. That original study also relied only on photorealistic comic-strip stimuli (rather than animate depictions or presences of robots, which could heighten person-perception and social presence; see Schreiner et al., 2017), because the comic robots may have been engaged as media characters rather than actual agents.

The present studies address these limitations by replicating that study but (a) correcting problematic ambiguities in the stimuli that afford opportunity for misinterpretation; (b) presenting stimulus scenarios via videos (as observed situations, Study 1) and in person (as live interactions, Study 2); (c) leveraging a human-sized android with a anthropomorphic or machinic features (human-like or machine-like face, respectively) to more carefully consider the role of visual anthropomorphic cues in mentalizing; and (d) evaluating a relational outcome of any mentalizing (i.e., trust). See the project materials (Banks, 2020a) for a detailed accounting of the source tests, the Banks' (2020c) adaptations, and the present studies' adjustments. Because these stimulus and format adjustments have the potential to rather dramatically shift outcomes, no specific predictions are offered based on the prior study; instead, the more general exploratory questions are engaged. The details and

findings of these studies are presented separately below, and then, discussed in aggregate. All project materials (instrumentation, stimuli, session scripts, data, and analysis outputs) are available online at <https://osf.io/9r73y/>; project materials are identified with prefixes S0 (relevant to both studies), S1 (Study 1), or S2 (Study 2).

Study 1: Mentalizing in Observed Human-Robot Interactions

An online experiment was conducted in which participants viewed and evaluated an agent (anthropomorphic robot, machinic robot, or human control), presented through a series of videos. The study was acknowledged as exempt by the West Virginia University Office of Research Integrity and Compliance, protocol #1802969640.

Method

Recruitment

Participants were recruited through social media posts and university research pool announcements, inviting them to complete a 20-min online survey on "perceptions of robot and human behavior" and offering entry into a drawing for a \$150 gift card as compensation. Simultaneously, participants were recruited through Mechanical Turk (MTurk); those participants were each paid \$1 for responding. Although a convenience sample, mentalizing is understood to be a universal form of social-cognitive activity among neurotypical humans (Carruthers & Smith, 1996) so the convenience sample is adequate in that findings should generalize to most populations.

Procedure

After acknowledging informed consent information and passing a check to ensure visibility of video content, participants were randomly assigned to one of the three agent conditions, identical in content except for agent type. The survey consisted of eight sections, in order: Introduction to the agent and solicitation of first impression, five randomly ordered ToM tests, evaluations of the agent, and capture of final impressions and demographics (S1-01a,b).

Stimulus Agents

All agents were named "Ray" (an easily pronounced and remembered name). The stimulus robot was a Robothespian 4 with Socibot head and projected face (Engineered Arts). To assess potentials for mentalizing to be associated with the degree of human-likeness (Eyssel & Pfundmair, 2015), participants interacted with a robot exhibiting one of the two faces: anthropomorphic female (Pris face) or machine (Robot 1 face), randomly assigned. In both cases, it had a female American-English accent (Heather voice) so that any effects of the visual facial cues could be discerned apart from any voice differences. The stimulus human was a young adult female with visual and aural features similar to the robot: Fair-skinned and thin, speaking in a measured pace and even tone, performing gestures similar to the robot's. In all conditions, Ray was framed as a helper who worked in the research lab, and the videos were framed as lab training videos (to mitigate critiques emerging in pilot tests that the videos did not seem natural).

ToM Tests and Dependent Measures

The introduction and each of the five ToM tests comprised a stimulus video followed by questions. Each test video reflected a mentalizing mechanism manifested in the source studies, remaining as true as possible to the source scenario while adapting it for human–robot interaction plausibility. See S1-02a–S1-03y for stimulus videos.

Introduction to Agent. Participants first encountered a video of the agent introducing itself, including its name, agent type, and key embodied features (height and abilities). Participants were asked to describe their first impressions.

Implicit ToM Tests. Following the introduction, in random order, participants viewed and responded to the five ToM scenario videos. Implicit empirical tests for ToM generally address one of the three markers indicating activation of the ToM inferential system (Byom & Mutlu, 2013). The first is the ability to recognize intentionality based on shared knowledge of the world, indicated by inferring another's motivations or intentions (Happé, 1994; Sarfati et al., 1997). Intentionality inferencing was tested through the White Lie and Next Action scenarios. Second is the ability to

interpret nonverbal social cues toward inferences of another's cognitive or affective states, indicated by identifying a target's emotional state through facial expressions (Baron-Cohen, 2001) or vocal expressions (McDonald, et al., 2006). Social-cues inferencing was tested via the Mind in the Eyes and Sarcasm scenarios. Finally, mentalizing is signaled by the recognition of divergent experience, tested through false-belief tasks in which individuals override their own factual knowledge about a situation, engage in perspective-taking, and acknowledge that a target agent (without complete facts) holds a false belief about the situation (Baron-Cohen et al., 1985). Divergent-experience inferencing was tested through the False Belief scenario. See Table 1 for summaries of the five tests, respective implicit indicator questions, and coded variables; see S1-01a,b for the complete survey in which the stimuli were embedded.

Other Measures. Evaluations of trust and explicit mind ascription were presented in random order. A binary measure (no/yes) captured explicit ascription of mind (“Do you think Ray has a mind?”) and to explain their answer (open). To evaluate trust-related intention, participants were asked whether they would accept an invitation for them to meet the agent and collaborate on a project.

Table 1
Summary of ToM Scenario Videos, Questions, and Coded Variables

Interaction summary	Evaluation questions	Variables coded
White lie (intention inference; strange stories, Happé, 1994)		
Agent was promised a gift for working in the lab and hoped for a program for dancing (robot) or dance lessons (human). Lab worker enters and gives the gift. The agent acts surprised, and the administrator asked if the agent liked the gift; agent replied yes, it was the only thing desired.	Consider Ray's very last statement about liking the gift—is it true? Why did Ray answer that way?	1. Lie detection* 2. Untruth detection 3. Presence of first-order attributions [§] 4. Presence of second-order attributions [§]
Mind in the eyes (facial affect; Baron-Cohen, 2001)		
Four videos of the agent exhibit a different facial expression in each (with no verbal cues). They were normatively happy, sad, angry, and afraid.	Type in any word that you think BEST describes what Ray's internal state was when the video was recorded. What led you to believe those were Ray's internal states?	5. Number of emotion words, out of four [†] 6. Presence of affective rationales
Sarcasm scenario (vocal affect; TASIT, McDonald et al., 2006)		
Agent responds to questions about what it is like to work in the lab. When finally asked about whether Ray enjoys the work, agent says that it is a “party” and a “real vacation.”	What did Ray mean by the last statement about working in the lab? Why did Ray answer that way? What did the human think Ray meant?	7. Sarcasm detection
Next actions (intention inference; Sarfati et al., 1997)		
Agent and lab worker are working on a task. Worker says she needs to break for lunch. Agent laughs at the need to break and finishes the task. Agent gets an internal alert that power was very low and it needed to find a power source (robot) or stomach growls and hurts (human).	Now, what do you think Ray would do next? Why do you think Ray would do that?	8. Prediction of optimal/sensical next behavior 9. Presence of intention inferences
False belief (divergent experience; Baron-Cohen et al., 1985)		
Agent asks lab worker if she has the \$5 owed, and worker challenges the robot to a double-or-nothing game. She puts the money in a blue box; agent must close eyes and guess the location. While closed, the money is moved to a red box.	Where would Ray look first for the money? Why would Ray look in that spot first?	10. Location of Ray's first look* 11. Presence of agent knowledge rationales

Note. ToM = Theory of Mind

* Indicates manifest, multiple choice variables; all others are latent, human-coded variables. [†] Indicates a continuous variable calculated from the four subscenario codes; all others are 0/1 absence/presence codes. [§] First-order attributions are inferences about another person's thoughts; second-order ToM attributions are inferences about another's thoughts about a third person's thoughts (Perner & Wimmer, 1985).

Then, they were asked to think about the things they might consider in deciding whether to accept, and (in relation to that decision) were asked to then respond to the 20-item multidimensional scale for trust in robots (Ullman & Malle, 2018). That scale comprises trust dimensions of capability ($\alpha = .906$), reliability ($\alpha = .883$), ethicality ($\alpha = .850$), and sincerity ($\alpha = .869$). Because all dimensions were highly correlated (range $r = .626-.839$, $p < .001$ for each pair) indicating a good deal of shared variance, exploratory factor analysis was performed (principal axis factoring, and oblimin rotation). In the unrotated solution, all items loaded onto a single factor with correlations above .60 and no cross-loadings above .60, so all items were combined into an omnibus trust metric ($\alpha = .957$; S1-00).

Open-ended items captured demographics via self-reports for age (in years), gender (standardized to male/female/nonbinary), and race/ethnicity (standardized into six categories).

Coding and Analytical Approach

Adapting the canonical in-person ToM tests to online survey formats also required adaptation of the analytical approach. The original tests were designed to be conducted in face-to-face settings and, for some tests, analysis is conducted in real-time through interviewers' adaptive probes. Here, the asynchronously gathered open-ended responses were instead systematically coded for key mentalizing markers (see Table 1, Variables Coded). The entire data set was coded by two independent raters trained for ~6 hr total, coding by detailed instructions (S1-04). Because (a) intercoder reliability statistics excessively punish slight divergences in low-frequency codes (see Krippendorff, 2011) and (b) some latent mentalizing indicators rely on subtle and highly subjective indicators, a negotiated-agreement approach was taken to determine the acceptability of data codes (see Campbell et al., 2013): Independent raters' evaluations were compared by the researcher and discrepancies flagged, and the raters discussed and resolved all discrepancies.

Results

Participants ($N = 182$) were $M = 31.86$ years ($SD = 11.324$, range 18–71, two not reporting). They identified as 51.1% female, 46.7% male, and 1.1% nonbinary (1.1% not reporting). A majority identified as White/Caucasian (74.2%), followed by Asian (7.1%),

with other racial/ethnic identifications comprising 14.3% of the sample, 4.4% not reporting. Just over half originated from MTurk (54.4%, $n = 99$) and the remainder from social media and pool recruitment. Random assignment resulted in condition counts of machinic robot $n = 54$, anthropomorphic robot $n = 72$, and human $n = 56$. See S1-05 for the complete data set. Zero-order correlations (S1-06) indicated near-perfect correlation ($r = .938$, $p < .01$) between the manifest and coded white-lie detection variables, so only the manifest (yes/no as to whether the agent's statement was the truth) was retained for analysis.

Implicit Mentalizing of Agents (RQ1)

To evaluate the extent to which participants engaged in implicit mentalizing of robotic agents in comparison to humans, mentalizing indicators were compared across agent conditions. Because the indicators signal discrete mentalizing mechanisms, each was analyzed separately via chi-square tests; the exception was the ratio-level emotion-word sum, analyzed via ANOVA. There was no difference across conditions for white-lie detection, white-lie second-order mentalizing, optimal behaviors, false-belief divergent experience, or false-belief mentalistic explanations; the model was significant for affect explanations of facial expressions, but the groups did not significantly differ (Table 2; see S1-07). Neither was there a difference in total emotion words, $F(2, 179) = 1.70$, $p = .19$. There was a difference, however, for white-lie first-order mentalizing and next-action intentionality, with humans most likely to elicit mentalizing for both. There was also a difference in sarcasm detection, for which the human was more likely than the machinic robot to elicit the indicator, though nondifferent from the anthropomorphic robot.

That there were significant differences among robots and a human for explanations of white lies (face-saving deception), predicted behavior (minded next actions), and sarcasm (nonliteral language) suggests that differences in mentalizing may emerge when people are prompted to elaborately make sense of agent behaviors. More indirect indicators (lie detection and box selection in a false-belief test) were frequent but did not significantly differ as a function of agent type. This suggests that mentalizing for robots and humans may be more alike than different when it comes to nonlaborative reactions to behavior. Acknowledging that statistical nondifference does not necessarily mean equivalence, these findings are tentatively

Table 2

Chi-Square Results for Implicit Mentalizing Indicators Across Agent-Type Conditions (Mediated Interaction)

	Machinic robot n (%)	Anthro. robot n (%)	Human n (%)	$\chi^2(2)$
WL: Lie detection	47 (87.0) _a	62 (86.1) _a	46 (82.1) _a	.606, $p = .739$, $V = .058$
WL: First-order	18 (34.0)_a	34 (48.6)_a	40 (71.4)_b	15.668, $p < .001$, $V = .296$
WL: Second-order	14 (26.4) _a	27 (38.6) _a	23 (41.1) _a	2.944, $p = .230$, $V = .128$
ME: Affect explain*	2 (3.7) _a	3 (4.2) _a	8 (14.3) _a	6.150, $p = .046$, $V = .184$
S: Sarcasm detection	32 (59.3)_a	49 (68.1)_{a,b}	46 (82.1)_b	6.995, $p = .030$, $V = .196$
NA: Optimal behavior	18 (33.3) _a	32 (45.1) _a	31 (55.4) _a	5.398, $p = .067$, $V = .173$
NA: Intention inferred	9 (16.7)_a	15 (21.4)_a	35 (62.5)_b	32.906, $p < .001$, $V = .428$
FB: Divergent experience	41 (75.9) _a	54 (75.0) _a	36 (64.3) _a	2.386, $p = .303$, $V = .115$
FB: Knowledge explain	31 (57.4) _a	46 (64.8) _a	34 (60.7) _a	.717, $p = .699$, $V = .063$

Note. WL = white lie; ME = Mind in the Eyes test; S = Vocal Sarcasm; NA = Next Action; FB = False Belief.

Values indicate number and percentage of individuals within conditions that presented indicators of mentalizing. Bolded frequencies with different subscripts are statistically different by condition. *Multivariate test was statistically significant, but pairwise comparisons yielded no clear group differences.

interpreted to suggest that (RQ1) people may similarly infer mental states in robots as in humans for superficial behavior observations, but mentalizing for robots is diminished when more deeply elaborating to explain agent behaviors.

Implicit Versus Explicit Mentalizing (RQ2)

To investigate relationships among mentalizing indicators, logistic regression was conducted to evaluate how the 10 implicit indicators, agent type, and interactions thereof were associated with explicit binary mind ascription (0/1). Implicit mentalizing indicators and agent types (dummy coded with human as referent group, 0) were entered in the first block, and mentalizing/agent-type interaction terms in the second.

The Block 1 model was statistically significant, correctly classifying 77.5% of cases (Tables 3 and S1-08). Those who explicitly ascribed mind to the agent were far less likely to have identified the agent's white lie as an untruth (compared to those who did not ascribe mind), but more than twice as likely to

have used mentalistic explanations for white lies. There was also a main effect of agent type: Those viewing either robots were far less likely to ascribe mind compared to those viewing a human agent. Finally, considering agent/indicator interaction terms, Block 2 was significant, correctly classifying an additional 3.4% of cases over Block 1. There were significant agent/indicator interactions for the next-action test: For the machinic robot only, those explaining behavior as intentional were 25 times more likely to ascribe mind, whereas those predicting an optimal next action far less likely to do so.

Addressing RQ2, few implicit ToM indicators were associated with explicit mind ascription. There was a negative main effect of agent-type on mind ascription with lower likelihood for either robot compared to a human, as well as main effects of white-lie indicators (positive for first-order mentalistic explanation and negative for more general lie detection). It may be that mentalistic explanations and explicit mind ascription both draw on elaborated thinking (and so are aligned). Notable interaction effects for the machinic robots (positive for intentionality and negative for

Table 3

Logistic Regression of Explicit Mind Ascription Upon Agent Type, Implicit Mentalizing Indicators, and Interactions Thereof

	<i>B</i>	<i>SE</i>	Wald	<i>p</i>	Exp(<i>B</i>)	CI
<i>Block 1: $\chi^2(12) = 13.30, p < .001$, Nagelkerke $R^2 = .430$</i>						
Implicit mentalizing indicators						
WL: Lie detection	-2.04	.65	9.74	.002	.13	.04/.47
WL: First-order	.80	.39	4.15	.042	2.23	1.03/4.83
WL: Second-order	.31	.41	.56	.453	1.36	.61/3.05
ME: Emotion words	-.33	.20	2.83	.092	1.39	.95/2.05
ME: Affect explain	.18	.76	.06	.812	1.20	.27/5.32
S: Sarcasm detection	-.03	.46	.00	.955	.98	.40/2.38
NA: Optimal behavior	.77	.41	3.52	.061	2.16	.97/4.83
NA: Intention inferred	.33	.46	.52	.469	1.39	.57/3.40
FB: Divergent experience	-.13	.57	.05	.819	.88	.29/2.66
FB: Knowledge explain	-.13	.52	.07	.798	.88	.32/2.42
Agent type (human as referent)						
Anthro. robot	-1.65	.49	11.24	.001	.19	.07/.50
Machinic robot	-2.18	.55	15.79	<.001	.11	.04/.33
<i>Block 2: $\chi^2(20) = 34.77, p = .021$, Nagelkerke $R^2 = .590$</i>						
Interaction terms: Agent type $\times$ ToM Indicators						
Anthro $\times$ Lie detect	-3.340	1.911	3.054	.081	.035	.00/1.50
Anthro $\times$ Lie first	-1.136	1.471	.596	.440	.321	.02/5.74
Anthro $\times$ Lie second	-1.615	1.518	1.132	.287	.199	.01/3.90
Anthro $\times$ Emotion	.413	.907	.208	.649	1.512	.26/8.95
Anthro $\times$ Affect	3.811	2.621	2.115	.146	45.202	.27/7690.30
Anthro $\times$ Sarcasm	-.070	1.531	.002	.964	.933	.05/18.75
Anthro $\times$ Optimal	-2.863	2.329	1.511	.219	.057	.00/5.48
Anthro $\times$ Intentional	1.358	1.354	1.006	.316	3.889	.2755.27
Anthro $\times$ Divergent	2.131	2.244	.902	.342	8.423	.10/685.01
Anthro $\times$ Knowledge	2.338	2.105	1.233	.267	10.359	.17/641.77
Machin $\times$ Lie detect	-.971	1.745	.310	.578	.379	.01/11.58
Machin $\times$ Lie first	-1.836	1.528	1.444	.230	.159	.01/3.19
Machin $\times$ Lie second	-.786	1.577	.249	.618	.456	.02/10.02
Machin $\times$ Emotion	.389	.971	.161	.689	1.476	.22/9.91
Machin $\times$ Affect	4.801	2.667	3.240	.072	121.679	.65/22684.09
Machin $\times$ Sarcasm	-.030	1.501	.000	.984	.971	.05/18.40
Machin $\times$ Optimal	-6.260	2.465	6.449	.011	.002	.00/.24
Machin $\times$ Intentional	3.209	1.570	4.180	.041	24.761	1.14/536.95
Machin $\times$ Divergent	1.794	2.388	.565	.452	6.014	.06/648.04
Machin $\times$ Knowledge	2.136	2.166	.972	.324	8.463	.12/590.36

Note. WL = white lie; ME = Mind in the Eyes test; S = Vocal Sarcasm; NA = Next Action; FB = False Belief.

Results significant at $p < .05$ are presented in bold. Main effects are reported from Block 1 and Interaction terms from Block 2.

optimal next-action) indicate that robotic-agent characteristics moderate some implicit–explicit mentalizing relationships. These patterns suggest that explicit mind-ascription may be promoted/reduced when robots appear to violate/uphold common mental models for robots: The machinic robot exhibited expected mechanical properties cuing the heuristic that mindless machines make systematic and optimal decisions (see Sundar, 2020) but also appeared to *mean* to make that decision, promoting mind ascription.

Mentalizing and Trust (RQ3)

In considering potential relationships between implicit/explicit mentalizing and trust in the agent, separate analyses were conducted for the two trust measures because they are distinct operationalizations of trust as (a) indirect evaluation of the agent's trustworthiness traits (the omnibus metric, continuous) and (b) as a hypothetical behavioral intention indicator (accepting the collaboration invitation or not, categorical); there is a moderate correlation between the two, $r = .429$, $p < .001$.

Regressing the omnibus trust metric upon mentalizing indicators and agent type, Model 1 (main effects) was significant, $F(13, 164) = 3.37$, $p < .001$, adj. $R^2 = .148$ (Table 4). Explicit mind ascription and more frequent use of emotion words to explain nonverbals were positively associated with trust; those viewing the anthropomorphic robot had higher trust ratings than those considering the human (which was not the case for the machinic robot). Although Model 2 (interactions) was significant, $F(35, 142) = 1.55$, $p < .040$, adj. $R^2 = .097$, no individual terms significantly predicted omnibus trust. See S1-09 for complete regression results.

Logistic regression of invitation acceptance upon mentalizing indicators did not result in a significant model for main effects, $\chi^2(13) = 19.33$, $p = .113$, or for agent $\times$ mentalizing interaction effects, $\chi^2(22) = 21.13$, $p = .513$. See S1-09 for complete regression results.

Addressing RQ3, a tendency to explain agent expressions in emotional terms and explicitly ascribing mind positively

contributes to trustworthiness evaluations (though not to trust-related intentions).

Study 2: Mentalizing in Participatory Human–Robot Interactions

Study 1 relied on survey-presented video stimuli of others interacting with a robot, as is common in human–robot interaction studies. However, mediated robot stimuli may reduce perceived social presence and interactivity (Schreiner et al., 2017), potentially limiting the social cuing and contextualizing information required for mentalizing (Achim et al., 2013). Thus, the Study 1 protocol was replicated with live human–robot interactions in Study 2. The study was approved by the West Virginia University Office of Research Integrity and Compliance, protocol #1802982775.

Method

Recruitment and Procedure

Participants were recruited through a university research participation pool and offered course credit and entry into a drawing for a \$150 gift card as compensation. They completed a presession online survey (S2-01) and scheduled a lab session. In that session, participants acknowledged informed consent information and were guided into a 10-foot $\times$ 12-foot room; the stimulus agent was positioned opposite a bistro table and two stools, approximately 8 feet apart. The researcher introduced participants to the agent, asked them to be seated on a stool, and guided them through interactive versions of the five randomly ordered ToM scenarios (Table 1). The Study 1 survey scenarios were adapted for live interaction: (a) participants directly observed the agent rather than viewing a video; (b) participants directly interacted with the agent instead of watching other humans interact; and (c) the researcher asked questions aloud after each scenario rather than presenting them via screen. During each questioning procedure, the agent was put “in a bubble” (the robot's listening/visual capabilities were described as being turned off; the human was blindfolded with

Table 4

Linear Regression of Agent Trust Evaluations Upon Implicit/explicit Mentalizing Indicators and Agent Type

	B(SE)	β	t	p	CI
Mentalizing indicators					
WL: Lie detection	-.30(.28)	-.09	-1.09	.277	-.85/.25
WL: First-order	-.26(.19)	-.11	-1.42	.157	-.63/.10
WL: Second-order	-.02(.19)	-.01	-.09	.926	-.39/.36
ME: Emotion words	.21(.08)	.19	2.55	.012	.05/.38
ME: Affect explain	-.56(.33)	-.12	-1.69	.093	-1.21/.10
S: Sarcasm detection	-.22(.20)	-.08	-1.06	.292	-.62/.19
NA: Optimal behavior	.16(.19)	.07	.83	.410	-.22/.54
NA: Intention inferred	.15(.22)	.06	.68	.498	-.28/.58
FB: Divergent experience	-.48(.26)	-.18	-1.88	.062	-.99/.03
FB: Knowledge explain	.09(.24)	.04	.39	.701	-.37/.55
Explicit mind ascription	.80(.21)	.33	3.91	<.001	.40/1.21
Agent type (human as referent)					
Anthro. robot	.52(.24)	.21	2.22	.028	.06/.99
Machinic robot	.33(.26)	.13	1.27	.206	-.18/.85

Note. WL = white lie; ME = Mind in the Eyes test; S = Vocal Sarcasm; NA = Next Action; FB = False Belief. Results significant at $p < .05$ are presented in bold.

noise-canceling headphones) so that the participant would feel free to respond without concern for the agent's feelings or judgment (S2-02 for complete session script). The procedures were video recorded for later response coding. Finally, participants were guided to a separate room to complete a survey capturing mind-perception and trust measures (S2-03).

Stimulus Agents

The stimulus robot, "Ray," was again a Robothespian 4 with the Socibot animated face—either the human female (Pris) or a machine (Robot 1) face, both with the female American-English accent (Heather voice). The human was the same young adult female as in Study 1. Notably, due to an inability to move the large robot between sessions to allow for complete random assignment to conditions, participants were randomly assigned to one of the two robot conditions for the first 74 cases—at which point the robot was disassembled and relocated—and the remainder of participants were assigned to the human condition. This approach was favored over randomly assigning separate lab spaces due to confounds of varied lighting and acoustics that could have influenced social cueing. Nonrandom agent assignment is acknowledged as a limitation.

ToM Tests and Dependent Measures

The implicit ToM tests and corresponding indicators were identical to those in Study 1, except for adaptations made for verbal rather than multiple-choice responses (a single lie detection response limited to no/yes [0/1] rather than both measured and coded, given their high correlation in Study 1). Also identical were measures for explicit mind ascription (no/yes, 0/1); trust ($\alpha = .950$), willingness to collaborate (no/yes, 0/1), and demographics. Trust items were again collapsed into an omnibus metric because most items loaded onto a single factor and retention allowed for faithful comparison against Study 1 (S2-00). Of note, nonresponse to some postinteraction questions (potentially as a function of fatigue at the end of the hour-long session) resulted in low cell counts for some dependent variables and listwise deletion in some analyses, acknowledged as a limitation. Mentalizing indicators were again coded by two independent coders using the same codebook, training, and disagreement-resolution procedure as in Study 1 (S1-04); coding was conducted using session recordings. A review of Study 1 open-ended responses suggested that participants' preexisting attitudes toward robots may influence willingness to explicitly ascribe mind, so 16 positively and negatively valenced items from the Implicit Association Test (Greenwald & Banaji, 1995) were adapted for use in the presession survey as a scale to measure preexposure attitudes toward target agent types (7-point Likert scale; negative items recoded such that higher values reflect more positive attitudes; $\alpha = .836$).

Results

Participants ($N = 102$) were $M = 20.34$ years ($SD = 1.749$, range 18–26). They identified as 52% female and 48% male. A majority identified as White/Caucasian (65.7%), followed by Arab/Middle Eastern (11.8%), with other racial/ethnic identifications comprising 22.5%. Random assignment into robot conditions and

purposive assignment to the human condition resulted in counts of machinic robot $n = 39$ (38.2%), anthropomorphic robot $n = 35$ (34.3%), and human $n = 28$ (27.5%). See S2-04 for the complete data set. Zero-order correlations indicate no collinearity among mentalizing indicators (see S2-05).

Implicit Mentalizing of Agents (RQ1)

To again evaluate the extent to which participants engaged in implicit mentalizing of stimulus agents, the five ToM tests' indicators were compared across conditions. Preexisting attitude toward agents was not significantly correlated with any mentalizing indicator so was not included in analysis. Indicators were again evaluated individually via chi-square, except for emotion-word count analyzed via ANOVA. There were no significant differences across conditions for white-lie first-order or second-order mentalizing, affect rationale for facial expressions, next-action as optimal or intentional, or false-belief box selection or mentalistic explanation (Table 5; see S2-06). Lie-detection differed and was least-likely elicited for the human. Sarcasm detection also differed: It was more likely inferred for humans and least likely for the machinic robot (with considerations of an anthropomorphic robot not differing from other conditions). In addition, emotion-word count differed, $F(2, 98) = 3.20$, $p = .045$, $\eta_p^2 = .06$, with highest frequency for humans and lowest for machinic robots (with considerations for anthropomorphic robots not differing from those for other agents)—though all mean values indicate frequent use of emotion terms (M range 3.18–3.71, of a possible range 0–4).

Study 1 found that people observing an interaction superficially mentalized robots similarly to humans, but more strongly mentalized humans when elaboration was prompted. In comparison, the Study 2 in-person replication indicates that, in live interactions, people do mentalize humans and robots similarly; however, physically copresent robots (compared to humans) garner more frequent lie detection, elicit emotional explanations for behaviors less frequently, and prompt less frequent identification of sarcasm. To again address RQ1, there is often no difference in mentalizing indicators among robots and humans, but divergence manifests when in-person nonverbal cues elicit inferences and in which language must be taken nonliterally (white lies and sarcasm) or in which language is not present at all (facial expressions).

Implicit Versus Explicit Mentalizing (RQ2)

Logistic regression was again conducted to evaluate associations of implicit mentalizing indicators, agent type, and their interactions with explicit mind ascription. The Block 1 model was significant and correctly classified 77.6% of cases; only agent class predicted explicit mind ascription (Table 6 and S2-07). For Block 2, the model failed to converge. This was likely due to specification issues associated with missing data because logistic regression uses maximum likelihood estimations, which require listwise deletion for case with missing values on any one measure; this resulted in reduced degrees of freedom necessary for the large number of interaction terms and that led to issues of (quasi)complete separation (Allison, 2004). Analysis was repeated using only the two significantly predicting interaction terms from Study 1 (machinic $\times$ optimal and machinic $\times$ intentional) as the most likely to predict explicit

Table 5*Chi-Square Results for Implicit Mentalizing Indicators Across Agent-Type Conditions (Live Interaction)*

	Machinic robot <i>n</i> (%)	Anthro. robot <i>n</i> (%)	Human <i>n</i> (%)	$\chi^2(2)$
WL: Lie detection	24 (77.4)_a	18 (64.3)_a	6 (21.4)_b	20.035, <i>p</i> < .001, <i>V</i> = .480
WL: First-order	24 (66.7) _a	25 (80.6) _a	22 (78.6) _a	2.033, <i>p</i> = .362, <i>V</i> = .146
WL: Second-order	18 (50.0) _a	16 (51.6) _a	8 (28.6) _a	3.954, <i>p</i> = .138, <i>V</i> = .204
ME: Affect explain	14 (37.8) _a	15 (42.9) _a	5 (17.9) _a	4.718, <i>p</i> = .095, <i>V</i> = .217
S: Sarcasm detection	20 (57.1)_{a,b}	14 (42.4)_a	21 (75.0)_b	6.570, <i>p</i> = .037, <i>V</i> = .262
NA: Optimal behavior	21 (58.3) _a	17 (53.1) _a	19 (67.9) _a	1.370, <i>p</i> = .504, <i>V</i> = .119
NA: Intentional inferred	20 (58.8) _a	15 (50.0) _a	13 (46.4) _a	1.030, <i>p</i> = .598, <i>V</i> = .106
FB: Divergent experience	33 (84.6) _a	33 (94.3) _a	21 (75.0) _a	4.636, <i>p</i> = .098, <i>V</i> = .213
FB: Knowledge explain	30 (76.9) _a	27 (77.1%) _a	20 (71.4) _a	.345, <i>p</i> = .842, <i>V</i> = .058

Note. WL = white lie; ME = Mind in the Eyes test; S = Vocal Sarcasm; NA = Next Action; FB = False Belief.

Values indicate number and percentage of individuals within conditions that presented indicators of mentalizing. Bolded frequencies with different subscripts are statistically different by condition.

mind ascription in the live interaction. With this approach, the Block-2 model converged but was not significant.

Addressing RQ2, then, in live interactions, people explicitly mentalize robots differently than humans, without a material association with any implicit mentalizing indicators. The physical copresence of a robot (of either appearance) appears to override the contributions of implicit indicators (i.e., white-lie detection and mentalistic explanations for deception) found in Study 1.

Mentalizing and Trust (RQ3)

To consider the relationship between mentalizing and trust, data were analyzed using the same approach as in Study 1: Linear regression for omnibus trust and logistic regression for trust-related intent. Regressing the omnibus trust metric upon agent type, mentalizing indicators, and agent $\times$ indicator interaction terms, there were no main effects $F(13, 35) = .657, p = .789, \text{adj. } R^2 = -.102$ or interaction effects $F(33, 15) = 1.089, p = .446, \text{adj. } R^2 = .058$ —

though low statistical power may have contributed to these null outcomes. Logistic regression indicates there were also no main effects of mentalizing indicators or agent type on willingness to accept the invitation to collaborate [$\chi^2(13) = 15.76, p = .263$]. As with RQ2, the interaction-term model failed to converge. As there were no significant interactions in Study 1 to reevaluate in the live interaction, no further analysis was conducted (see S2-08).

General Discussion

On the grounds that mentalizing and mind ascription are central to how humans communicate with other humans and may be critical to the acceptance of technologies as social agents, the present studies conducted conceptual replications of canonical ToM tests in the perception of robotic agents compared to humans. Findings broadly suggest that implicit indicators of mentalizing are mostly nondifferent among agent types; however, the detection of nonliteral language use (i.e., deception and sarcasm) was more likely for

Table 6*Logistic Regression of Mind Ascription Upon Implicit Mentalizing and Agent Type*

	<i>B</i>	<i>SE</i>	Wald	<i>p</i>	Exp(<i>B</i>)	CI
Block 1: $\chi^2(12) = 24.50, p = .015, \text{Nagelkerke } R^2 = .387$						
Implicit mentalizing indicators						
WL: Lie detection	-.31	.77	.16	.686	.73	.16/.3.31
WL: First-order	.61	.70	.74	.389	1.83	.46/7.29
WL: Second-order	1.01	.66	2.37	.124	2.75	.76/10.00
ME: Emotion words	.53	.38	1.94	.164	1.69	.81/3.55
ME: Affect explain	.20	.66	.09	.766	1.22	.33/4.46
S: Sarcasm detect	.39	.63	.38	.537	1.48	.43/5.06
NA: Optimal behavior	-.25	.75	.11	.740	.78	.18/3.38
NA: Intention inferred	1.21	.82	2.19	.139	3.36	.68/16.68
FB: Divergent experience	-1.20	1.26	.91	.340	.30	.033/.56
FB: Knowledge explain	.99	.85	1.34	.248	2.68	.50/14.28
Agent type (human as referent)						
Anthro. robot	-2.78	.96	8.35	.004	.06	.00/.41
Machinic robot	-2.78	.98	8.07	.005	.06	.00/.42
Block 2: $\chi^2(2) = .14, p = .934, \text{Nagelkerke } R^2 = .389$						
Interaction terms: Agent type $\times$ ToM indicators						
Machin $\times$ Optimal	.380	1.593	.057	.811	1.462	.064/33.16
Machin $\times$ Intentional	.143	1.678	.007	.932	1.153	.043/30.93

Note. WL = white lie; ME = Mind in the Eyes test; S = Vocal Sarcasm; NA = Next Action; FB = False Belief.

Results significant at $p < .05$ are presented in bold. Main effects are reported from Block 1 and Interaction terms from Block 2.

humans across both studies. In observed interactions, intentionality inferencing in behavior explanations was more likely for humans, and in live interactions, affective interpretation of nonverbal cues was more frequent for humans than for robots (RQ1). Likelihood of explicitly ascribing mind was lower for robots (compared to a human) across both studies. In observed interactions, lie detection decreased mind-ascription likelihood and first-order mentalistic explanations for deception decreased ascription likelihood; for machinic robots only, those predicting an optimal (i.e., mindful in its rationality) next behavior were less likely to ascribe mind and those mentioning intentionality in behavior explanations were far more likely. In the live interaction, only the main effect of agent type on decreased mind-ascription likelihood was exhibited (RQ2). Finally, in observed interactions, trust ratings were linked to greater tendency to describe facial nonverbal cues using emotion words, explicit mind ascription, and viewing an anthropomorphic robot; these associations were not found in live interactions. Neither implicit nor explicit mentalizing indicators nor agent types were linked with behavioral-intention indicators of trust (RQ3).

Altogether, findings generally comport with the prior research on which this investigation was based (Banks, 2020c) in indicating that implicit markers of mentalizing for robots and humans are mostly nondifferent when social cues are similar and in that there are few links between implicit and explicit markers of mentalizing. However, where these patterns do *not* hold, two dynamics emerged as potentially important in ToM dynamics for robots: Nonliteral language may be interpreted differently for robots than for humans, and live interactions may prompt reliance on heuristics and reduction in elaborative sense-making.

NonLiteral Language Interpretation as a Mentalizing Differentiator

Across the studies, only a few implicit mentalizing indicators differed among agents. In both studies, interpretations of a white lie and detection of sarcasm favored mentalizing humans over robots, though the finding is interpreted cautiously given the nonrandom agent assignment in Study 2. The agent-specific divergence around these indicators is important because lies and sarcasm are both forms of nonliteral language. Detecting both deception and sarcasm relies on the interpretation of nonverbal cues, semantic context, and relational dynamics (cf., Attardo et al., 2003; Dunbar et al., 2014). As robots become more advanced, it may be that people expect robots to not only understand human sarcasm [made possible via emotion detection (Fung et al., 2016) and probabilistic sentiment analysis (e.g., Radpour & Ashokkumar, 2017)] but also to generate their own sarcastic expressions (e.g., Joshi et al., 2018).

The apparent importance of nonliteral language to mental-state inferencing holds implications for design practice. Robots that *effectively* employ nonliteral language by drawing on human norms for sarcasm may be more readily accepted into human social spheres and those with less convincing expressions may provoke more mechanistic—and less mentalistic—interpretations of their behaviors. Such potentials sit at the center of the debate around XAI, with some suggesting that anthropomorphic appearance and behaviors are necessary for acceptance (see Duffy, 2003) and others arguing that they leave humans susceptible to manipulation by machines (and their creators) such that transparent, mechanistic explanations

are preferable (cf., Bryson, 2019), especially when deception is of concern (cf., Danaher, 2020).

Participation May Prompt Reliance on Agent-Category Heuristics

In observed interactions, detecting deception was associated with lower likelihood of explicit mind ascription, whereas explaining deception in first-order mentalistic terms linked to a higher likelihood, in tandem with a negative main effect of viewing either robot (vs. a human). In the live interaction, however, only agent type predicted mind ascription. Again interpreted cautiously (due to nonrandom assignment to the observed/live studies), findings suggest that participating in an encounter with an interactive agent places social demands on human interlocutors (see Bowman, 2018). Those demands may have diminished opportunities to engage in the third-person cognitive elaborations and moved them to instead rely more on agent-type heuristics. In other words, it may be that mentalizing robots follows dual-process theories for social cognition (see Evans & Stanovich, 2013): Observing affords disengagement required for people to engage in slower, more purposeful sense-making, but interacting with a robot people moves people to engage in faster, more automatic mental shortcuts regarding that agent class.

An agent's ontological category—the *type* of thing an agent is thought to be—is known to provoke powerful heuristics (Guzman, 2020): Often in spite of social cues, sometimes in co-ordination with cues, and sometimes overridden by cues. That people directly interacting with robots trended toward rejecting mind ascription suggests that, in the absence of time and distance to reflect on the agents, mental models may govern conscious responses to evaluating robots' minds (Banks, 2020b). These models likely include the heuristics for "mindless machines" or "iron idiots" often presented in media representations of robots (Bruckenberg, et al., 2013). Based on the current studies' designs, however, it is difficult to disentangle whether the shift to relying on ontological-category heuristics may be a function of copresence (vs. mediation) or of active participation (vs. more passive observation). Regarding copresence, physical immediacy promotes increased social presence (Schreiner et al., 2017) making a robot's outgroup status more salient or even novelty effects that prompt reliance on media tropes (Banks, 2020b). Regarding participatory engagement, it may be that live interactions placed higher social demands on participants, moving them to reallocate cognitive resources from sense-making of agent behavior to the interactive activities at hand. Resources reallocated, they may have had to rely more on mental shortcuts for mindlessness rather than elaborating on the possibility of mindfulness (cf., Bowman, 2018; Sundar, 2020). Because heuristics are powerful in fostering both social and technological acceptance (cf., Nass et al., 1994), future research should work to disentangle the mechanisms by which people are moved to invoke heuristics over elaborative considerations—especially in relation to contexts in which robots align with or deviate from people's mental models for those agents.

Limitations and Future Research

The present study's findings are subject to limitations that call for replication and extension of the work. In addition to acknowledged

small power issues, nonrandom assignment to the human condition in Study 2, and nonrandom assignment across the two studies, a cursory qualitative reviews of behavior explanations hint at the importance of contextual information for mentalizing to occur (e.g., sense-making appeared to include consideration of environmental factors and other actors in the scenarios) such that different patterns may emerge according to domain (non)specificity. The present studies also accounted only for the humans' mentalizing of robots, when effective dyadic interactions require communicative coordination of *two* minds (Bahrami et al., 2010). Indeed, a rich body of work has long-considered the ways that robots may be programmed to model human interlocutors' mental states (see Scassellati, 2002) toward more comfortable and productive interactions via fluid and meaningful conversations (Mavridis, 2015), prediction of human intent (Görür et al., 2017), and even recognizing themselves as distinct social agents (Dautenhahn, 1994). Future research should attend to potential dynamics and effects of how robots' modeling of human mental models may actually contribute to meaningful interactions through coordinated social cognitions, as well as how these dynamics may shift according to contextual and personological factors. Finally, the convenience samples for both studies were nonrepresentative and no participants exhibited signs of cognitive impairment, so generalizability of results to other populations should be tested.

Conclusion

This study's findings suggest that automatic mentalizing of robots and humans is mostly nondifferent (cf., Thellman et al., 2017), except when nonliteral language and social demand may prompt elaborated sense-making and/or reliance on agent-category heuristics for (non)mindedness. Thus, as social technologies advance in sophistication toward the potential for intersubjective human-machine relations (Marchetti et al., 2018), it may be that humans' perceptions of those agents rely sufficiently on preconscious social-cognitive processes (cf., Reeves & Nass, 1995) such that machine agents might offer similar mind-associated relational gratifications—and encumbrances—as do humans (see de Graaf, 2016). Considering these potentials in tandem with these findings, the perceptions of mind in nonhuman agents are critical to understand and may be best explored through dual-process paradigms (Banks, 2020c, cf., Lobato et al., 2013; Evans & Stanovich, 2013) given the divergence of implicit and explicit indicators revealed in this study.

References

- Achim, A. M., Guitton, M., Jackson, P. L., Boutin, A., & Monetta, L. (2013). On what ground do we mentalize? Characteristics of current tasks and sources of information that contribute to mentalizing judgments. *Psychological Assessment*, 25(1), 117–126. <https://doi.org/10.1037/a0029137>
- Allison, P. D. (2004). Convergence problems in logistic regression. In M. Altman, J. Gill, & M. P. McDonald (Eds.), *Numerical issues in statistical computing for the social scientist* (pp. 238–252). John Wiley & Sons. <https://doi.org/10.1002/0471475769.ch10>
- Anjomshoe, S., Najjar, A., Calvaresi, D., & Främling, K. (2019). Explainable agents and robots: Results from a systematic literature review. *Proceedings of the 18th international conference on autonomous agents and multiagent systems* (pp. 1078–1088). IFAAMS.
- Appel, M., Izydorczyk, D., Weber, S., Mara, M., & Lischetzke, T. (2020). The uncanny mind in a machine: Humanoid robots as tools, agents, and experiencers. *Computers in Human Behavior*, 102, 274–286. <https://doi.org/10.1016/j.chb.2019.07.031>
- Apperly, I. A. (2012). What is “theory of mind”? Concepts, cognitive processes, and individual differences. *The Quarterly Journal of Experimental Psychology*, 65(5), 825–839. <https://doi.org/10.1080/17470218.2012.676055>
- Attardo, S., Eisterhold, J., Hay, J., & Poggi, I. (2003). Multimodal markers of irony and sarcasm. *Humor: International Journal of Humor Research*, 16(2), 243–260. <https://doi.org/10.1515/humr.2003.012>
- Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., & Frith, C. D. (2010). Optimally interacting minds. *Science*, 329(5995), 1081–1085. <https://doi.org/10.1126/science.1185718>
- Banks, J. (2020a, October 16). *Theory of mind in social robots*. <https://OSF.IO/9R73Y>
- Banks, J. (2020b). Optimus primed: Media cultivation of robot mental models and social judgments. *Frontiers in Robotics and AI*, 7, Article 62. <https://doi.org/10.3389/frobt.2020.00062>
- Banks, J. (2020c). Theory of mind in social robots: Replication of five established human tests. *International Journal of Social Robotics*, 12, 403–414. <https://doi.org/10.1007/s12369-019-00588-x>
- Baron-Cohen, S. (2001). Theory of mind in normal development and autism. *P.R.I.S.M.E. Psychiatrie, Recherche et Intervention en Santé Mentale de l'Enfant*, 34(1), 74–83.
- Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a “theory of mind?” *Cognition*, 21(1), 37–46. [https://doi.org/10.1016/0010-0277\(85\)90022-8](https://doi.org/10.1016/0010-0277(85)90022-8)
- Bartneck, C., Kulić, D., Croft, E., & Zoghbi, S. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics*, 1(1), 71–81. <https://doi.org/10.1007/s12369-008-0001-3>
- Bowman, N. D. (2018). The demanding nature of video game play. In N. D. Bowman (Ed.), *Video games: A medium that demands our attention* (pp. 1–24). Routledge. <https://doi.org/10.4324/9781351235266-1>
- Bruckenberg, U., Weiss, A., Mirnig, N., Strasser, E., Stadler, S., & Tscheligi, M. (2013). *The good, the bad, the weird: Audience evaluation of a “Real” robot in relation to science fiction and mass media* [Conference session]. Proceedings of the International Conference on Social Robotics, Cham.
- Bryson, J. J. (2019). Robot, all too human. *XRDS: Crossroads. The ACM Magazine for Students*, 25(3), 56–59.
- Byom, L. J., & Mutlu, B. (2013). Theory of mind: Mechanisms, methods, and new directions. *Frontiers in Human Neuroscience*, 7, Article 413. <https://doi.org/10.3389/fnhum.2013.00413>
- Bzdok, D., Schilbach, L., Vogeley, K., Schneider, K., Laird, E. R., Langner, R., & Eickhoff, S. B. (2012). Parsing the neural correlates of moral cognition: ALE meta-analysis on morality, theory of mind, and empathy. *Brain Structure & Function*, 217, 783–796. <https://doi.org/10.1007/s00429-012-0380-y>
- Campbell, J. L., Quincy, C., Osserman, J., & Pedersen, O. K. (2013). Coding in-depth semistructured interviews: Problems of unitization and intercoder reliability and agreement. *Sociological Methods & Research*, 42(3), 294–320. <https://doi.org/10.1177/0049124113500475>
- Carruthers, P., & Smith, P. K. (1996). Introduction. In P. Carruthers & P. K. Smith (Eds.), *Theories of theories of mind* (pp. 1–8). Cambridge University Press. <https://doi.org/10.1017/CBO9780511597985.002>
- Chaiken, S., & Trope, Y. (1999). *Dual-process theories in social psychology*. Guilford.
- Clément, F., Koenig, M. A., & Harris, P. L. (2004). The ontogenesis of trust. *Mind & Language*, 19(4), 360–379. <https://doi.org/10.1111/j.0268-1064.2004.00263.x>
- Cross, E. S., Hortensius, R., & Wykowska, A. (2019). From social brains to social robots: Applying neurocognitive insights to human-robot

- interaction. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 374(1771), Article 20180024. Advance online publication. <https://doi.org/10.1098/rstb.2018.0024>
- Danaher, J. (2020). Robot betrayal: A guide to the ethics of robotic deception. *Ethics and Information Technology*, 22, 117–128. <https://doi.org/10.1007/s10676-019-09520-3>
- Dautenhahn, K. (1994). Trying to imitation—A step towards release robots from social isolation. *Proceedings of PercAc*, 94, 290–301.
- de Graaf, M. M. (2016). An ethical evaluation of human-robot relationships. *International Journal of Social Robotics*, 8(4), 589–598. <https://doi.org/10.1007/s12369-016-0368-5>
- Deska, J. C., Lloyd, E. P., & Hugenberg, K. (2018). Facing humanness: Facial width-to-height ratio predicts ascriptions of humanity. *Journal of Personality and Social Psychology*, 114(1), 75–94. <https://doi.org/10.1037/pspi0000110>
- Doran, D., Schulz, S., & Besold, T. R. (2017). *What does explainable AI really mean? A new conceptualization of perspectives*. <https://arxiv.org/abs/1710.00794>
- Duffy, B. (2003). Anthropomorphism and the social robot. *Robotics and Autonomous Systems*, 42(3–4), 177–190. [https://doi.org/10.1016/S0921-8890\(02\)00374-3](https://doi.org/10.1016/S0921-8890(02)00374-3)
- Dunbar, N. E., Jensen, M. L., Tower, D. C., & Burgoon, J. K. (2014). Synchronization of nonverbal behaviors in detecting mediated and non-mediated deception. *Journal of Nonverbal Behavior*, 38(3), 355–376. <https://doi.org/10.1007/s10919-014-0179-z>
- Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, 114(4), 864–886. <https://doi.org/10.1037/0033-295X.114.4.864>
- Evans, J., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science*, 8(3), 223–241. <https://doi.org/10.1177/1745691612460685>
- Eyssel, F. A., & Pfundmair, M. (2015). *Predictors of psychological anthropomorphization, mind perception, and the fulfillment of social needs: A case study with a zoomorphic robot* [Symposium]. RO-MAN: 24th IEEE international symposium on in robot and human interactive communication, Kobe, Japan.
- Ferrari, F., Paladino, M. P., & Jetten, J. (2016). Blurring human-machine distinctions: Anthropomorphic appearance in social robots as a threat to human distinctiveness. *International Journal of Social Robotics*, 8, 287–302. <https://doi.org/10.1007/s12369-016-0338-y>
- Fu, C., Yoshikawa, Y., Iio, T., & Ishiguro, H. (2020). Sharing experiences to help a robot present its mind and sociability. [online before print]. *International Journal of Social Robotics*. Advance online publication. <https://doi.org/10.1007/s12369-020-00643-y>
- Fung, P., Bertero, D., Wan, Y., Dey, A., Chan, R. H. Y., Siddique, F. B., Yang, Y., Wu, C.-S., & Lin, R. (2016). Towards empathetic human-robot interactions. In *International conference on intelligent text processing and computational linguistics* (pp. 173–193). Springer.
- Goldman, A. I. (2012). Theory of mind. In E. Margolis, R. Samuels, & S. Stich (Eds.), *Oxford handbook of philosophy and cognitive science* (pp. 402–425). Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780195309799.003.0017>
- Gopnik, A. (1998). Explanation as orgasm. *Minds and Machines: Journal for Artificial Intelligence, Philosophy and Cognitive Science*, 8, 101–118. <https://doi.org/10.1023/A:1008290415597>
- Gordon, R. M. (1986). Folk psychology as simulation. *Mind & Language*, 8, 101–118.
- Görür, O. C., Rosman, B. S., Hoffman, G., & Albayrak, S. (2017). *Toward integrating theory of mind into adaptive decision-making of social robots to understand human intention*. Workshop on the role of intentions in human-robot interaction at the 12th ACM/IEEE international conference on human-robot interaction. ACM.
- Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science*, 315(5812), 619. <https://doi.org/10.1126/science.1134475>
- Gray, K., & Wegner, D. M. (2012). Feeling robots and human zombies: Mind perception and the uncanny valley. *Cognition*, 125(1), 125–130. <https://doi.org/10.1016/j.cognition.2012.06.007>
- Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*, 102(1), 4–27. <https://doi.org/10.1037/0033-295X.102.1.4>
- Guzman, A. L. (2020). Ontological boundaries between humans and computers and the implications for human-machine communication. *Human-Machine Communication*, 1, 37–54. <https://doi.org/10.30658/hmc.1.3>
- Happé, F. G. (1994). An advanced test of theory of mind: Understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. *Journal of Autism and Developmental Disorders*, 24(2), 129–154. <https://doi.org/10.1007/BF02172093>
- Henschel, A., Hortensius, R., & Cross, E. S. (2020). Social cognition in the age of human-robot interaction. *Trends in Neurosciences*, 43(6), 373–384. <https://doi.org/10.1016/j.tins.2020.03.013>
- Hofmann, W., Gawronski, B., Gschwendner, T., Le, H., & Schmitt, M. (2005). A meta-analysis on the correlation between the Implicit Association Test and explicit self-report measures. *Personality and Social Psychology Bulletin*, 31(10), 1369–1385. <https://doi.org/10.1177/0146167205275613>
- Imuta, K., Henry, J. D., Slaughter, V., Selcuk, B., & Ruffman, T. (2016). Theory of mind and prosocial behavior in childhood: A meta-analytic review. *Developmental Psychology*, 52(8), 1192–1205. <https://doi.org/10.1037/dev0000140>
- Joshi, A., Bhattacharyya, P., & Carman, M. J. (2018). Sarcasm generation. In *Investigations in computational sarcasm. Cognitive systems monographs* (Vol. 37). Springer. https://doi.org/10.1007/978-981-10-8396-9_5
- Keysar, B., Lin, S., & Barr, D. J. (2003). Limits on theory of mind use in adults. *Cognition*, 89(1), 25–41. [https://doi.org/10.1016/S0010-0277\(03\)00064-7](https://doi.org/10.1016/S0010-0277(03)00064-7)
- Khalid, S., Deska, J. C., & Hugenberg, K. (2016). The eyes are the windows to the mind: Direct eye gaze triggers the ascription of others' minds. *Personality and Social Psychology Bulletin*, 42(12), 1666–1677. <https://doi.org/10.1177/0146167216669124>
- Krippendorff, K. (2011). Agreement and information in the reliability of coding. *Communication Methods and Measures*, 5(2), 93–112. <https://doi.org/10.1080/19312458.2011.568376>
- Lee, K. M., Peng, W., Jin, S.-A., & Yan, C. (2006). Can robots manifest personality?: An empirical test of personality recognition, social responses, and social presence in human-robot interaction. *Journal of Communication*, 56(4), 754–772. <https://doi.org/10.1111/j.1460-2466.2006.00318.x>
- Lee, S.-Y. (2018). *Impact of human like cues on human trust in machines: Brain imaging and modeling studies for human-machine interactions*. Air Force Research Laboratory. <https://apps.dtic.mil/dtic/tr/fulltext/u2/1046152.pdf>
- Leslie, A. M. (1987). Pretense and representation: The origins of “theory of mind.” *Psychological Review*, 94(4), 412–426. <https://doi.org/10.1037/0033-295X.94.4.412>
- Lobato, E. J., Wiltshire, T. J., & Fiore, S. M. (2013). A dual-process approach to understanding human-robot interaction. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 57(1), 1263–1267. <https://doi.org/10.1177/1541931213571280>
- Low, J., & Perner, J. (2012). Implicit and explicit theory of mind: State of the art. *British Journal of Developmental Psychology*, 30(1), 1–13. <https://doi.org/10.1111/j.2044-835X.2011.02074.x>
- Marchetti, A., Manzi, F., Itakura, S., & Massaro, D. (2018). Theory of mind and humanoid robots from a lifespan perspective. *Zeitschrift für Psychologie*, 226, 98–109. <https://doi.org/10.1027/2151-2604/a000326>

- Mavridis, N. (2015). A review of verbal and non-verbal human-robot interactive communication. *Robotics and Autonomous Systems*, 63, 22–35. <https://doi.org/10.1016/j.robot.2014.09.031>
- McCabe, K., Houser, D., Ryan, L., Smith, V., & Trouard, T. (2001). A functional imaging study of cooperation in two-person reciprocal exchange. *Proceedings of the National Academy of Sciences*, 98, 2011832–2011835. <https://doi.org/10.1073/ypnas.211415698>
- McDonald, S., Bornhofen, C., Shum, D., Long, E., Saunders, C., & Neulinger, K. (2006). Reliability and validity of The Awareness of Social Inference Test (TASIT): A clinical test of social perception. *Disability and Rehabilitation*, 28(24), 1529–1542. <https://doi.org/10.1080/09638280600646185>
- Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 72–78). ACM.
- Nosek, B. A. (2007). Implicit-explicit relations. *Current Directions in Psychological Science*, 16(2), 65–69. <https://doi.org/10.1111/j.1467-8721.2007.00477.x>
- Perez-Osario, J., & Wykowska, A. (2020). Adopting the intentional stance toward natural and artificial agents. *Philosophical Psychology*, 33, 369–395. <https://doi.org/10.1080/09515089.2019.1688778>
- Perner, J., & Wimmer, H. (1985). “John thinks that Mary thinks that ...” Attribution of second-order beliefs by 5–10 year old children. *Journal of Experimental Child Psychology*, 39, 437–471. [https://doi.org/10.1016/0022-0965\(85\)90051-7](https://doi.org/10.1016/0022-0965(85)90051-7)
- Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4), 515–526. <https://doi.org/10.1017/S0140525X00076512>
- Prochazkova, E., Prochazkova, L., Giffin, M. R., Scholte, H. S., De Dreu, C. K., & Kret, M. E. (2018). Pupil mimicry promotes trust through the theory-of-mind network. *Proceedings of the National Academy of Sciences of the United States of America*, 115(31), E7265–E7274. <https://doi.org/10.1073/pnas.1803916115>
- Radpour, N. D., & Ashokkumar, V. (2017). *Non-Contextual Modeling of Sarcasm using a Neural Network Benchmark*. arXiv preprint arXiv:1711.07404.
- Reeves, B., & Nass, C. (1995). *The media equation: How people treat computers, television, and new media like real people and places*. Cambridge University Press.
- Sarfati, Y., Hardy-Baylé, M. C., Besche, C., & Widlöcher, D. (1997). Attribution of intentions to others in people with schizophrenia: A non-verbal exploration with comic strips. *Schizophrenia Research*, 25(3), 199–209. [https://doi.org/10.1016/S0920-9964\(97\)00025-X](https://doi.org/10.1016/S0920-9964(97)00025-X)
- Scassellati, B. (2002). Theory of mind for a humanoid robot. *Autonomous Robots*, 12, 13–24. <https://doi.org/10.1023/A:1013298507114>
- Schreiner, C., Mara, M., & Appel, M. (2017). *When R2-D2 hops off the screen: A service robot encountered in real life appears more real and human-like than on video or in VR* [Conference session]. MediaPsych2017: The 10th conference of the media psychology division of the German Psychological Society, Landau, Germany.
- Smith, P. K. (2017). Bullying and theory of mind: A review. *Current Psychiatry Reviews*, 13(2), 90–95. <https://doi.org/10.2174/1573400513666170502123214>
- Spence, P. R., Westerman, D. K., Edwards, C., & Edwards, A. (2014). Welcoming our robot overlords: Initial expectations about interaction with a robot. *Communication Research Reports*, 31(3), 272–280. <https://doi.org/10.1080/08824096.2014.924337>
- Sundar, S. S. (2020). Rise of machine agency: A framework for studying the psychology of Human—AI Interaction (HAI). *Journal of Computer-Mediated Communication*, 25(1), 74–88. <https://doi.org/10.1093/jcmc/zmz026>
- Takahashi, H., Saito, C., Okada, H., & Omori, T. (2013). An investigation of social factors related to online mentalizing in a human-robot competitive game. *Japanese Psychological Research*, 55(2), 144–153. <https://doi.org/10.1111/jpr.12007>
- Tanibe, T., Hashimoto, T., & Karasawa, K. (2017). We perceive a mind in a robot when we help it. *PLOS ONE*, 12(7), Article e0180952. <https://doi.org/10.1371/journal.pone.0180952>
- Thellman, S., Giagtzidou, A., Silvervarg, A., Ziemke, T. (2020). *An implicit, non-verbal measure of belief attribution to robots* [Conference session]. HRI’20: Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction, Cambridge, United Kingdom. <https://doi.org/10.1145/3371382.3378346>
- Thellman, S., Silvervarg, A., & Ziemke, T. (2017). Folk-psychological interpretation of human vs. humanoid robot behavior: Exploring the intentional stance toward robots. *Frontiers in Psychology*, 8, Article 1962.
- Thellman, S., & Ziemke, T. (2019). The intentional stance toward robots: Conceptual and methodological considerations. *Proceedings of the 41st annual conference of the cognitive science society* (pp. 1097–1103). Cognitive Science Society.
- Ullman, D., & Malle, B. F. (2018). *What does it mean to trust a robot? Steps toward a multidimensional measure of trust* [late-breaking report] [Conference session]. HRI’18 companion, Chicago, IL. <https://doi.org/10.1145/3173386.3176991>
- Wiese, E., Mandell, A., Shaw, T., & Smith, M. (2019). Implicit mind perception alters vigilance performance because of cognitive conflict processing. *Journal of Experimental Psychology: Applied*, 25(1), 25–40. <https://doi.org/10.1037/xap0000186>
- Wiese, E., Metta, G., & Wykowska, A. (2017). Robots as intentional agents: Using neuroscientific methods to make robots appear more social. *Frontiers in Psychology*, 8, Article 1663.

Received June 22, 2020

Revision received October 16, 2020

Accepted October 16, 2020 ■