

Mixed evidence of a Moral Mind Heuristic in Zero-History HRI: The (Unstable) Concomitance of Mind, Morality, and Trust Judgments

Jaime Banks
Syracuse University
banks@syr.edu

Abstract

Extant research offers piecemeal evidence of the operation of a moral mind heuristic (MMH)—a shorthanded judgment in which covarying mental, moral, and trustworthiness judgments emerge under zero-history, morally neutral exposures to humanoid robots. Three criteria must be met for such an operation: Concomitance (unordered co-activation of judgments), varied accessibility (salience can be primed), and biasing effects (drives more positive perceptions). Study 1 confirms concomitance. Study 2 confirms accessibility and effects. Study 3 replicates Study 2 an in-person robot exposure, however the MMH construct became unstable.

Keywords: Social robot, cognitive heuristic, theory of mind, trustworthiness, social distance

1. Heuristics and Cognitive Biases in HRI

Contemporary culture sees human exposure to social robots manifested largely through news and popular media. As social robots are mainstreamed, we may see hybrid societies with co-existing human and machine agents, flush with novel human-robot encounters. Even when people are accustomed to robots, there will continue to be novel (i.e., zero-history) encounters just as we meet new people in everyday life—from unfamiliar android receptionists to meeting colleagues’ robot companions. Although human-robot interaction (HRI) scholarship tends to focus on novel encounters as a convenience, such zero-history encounters are still worthy of attention because they are moments of impression formation and co-orientation (Avelino et al., 2021). Of particular importance in first impressions are the operation of heuristics—learned mental shortcuts in which complex decisions are replaced with simpler ones for which there is an efficient (if often suboptimal) strategy (Kahneman & Frederick, 2002). For instance, we might learn from media representations that robots with round, cute eyes are friendly and those with industrial aesthetics are threatening. Although heuristics can be employed in careful decision-

making, they are especially useful for quick decisions such as intuiting whether a new robot is friendly or threatening based on its appearance on approach.

Heuristics can be understood as mediators between stimuli and biased cognitive processing (Bellur & Sundar, 2014). That is, a stimulus triggers a heuristic that then promotes shorthanded cognitive judgment (Stimulus→Heuristic→Bias). For instance, a robot offers a stimulus (e.g., round eyes) that triggers a mental shortcut (e.g., is friendly). Then, we may make hasty judgments (e.g., is friendly, therefore safe to approach). Altogether, it is useful to consider such judgments as *if-then-therefore* processes: If stimulus, then heuristic, therefore bias (see Sundar, 2008).

Many heuristics unfold in HRI, from learned attributes of machines to shortcuts for whether we should consider them part of groups. Acknowledging the influence of heuristics on HRI, I call attention here to recent evidence suggesting the potential for a yet-untested heuristic that I tentatively call a moral mind heuristic (MMH) in which hasty judgments of a robot’s mental and moral capacities coalesce with trust(worthiness) perceptions to influence social judgments. In three studies detailed here, I argue for minimum necessary criteria for MMH operation and present initial mixed evidence around those criteria.

1.1. Empirical and Theoretical Background: Mind, Morality, and Trust

Social machines are increasingly integrated into human social spheres such that they occupy actual or perceived moral roles—from social actors in everyday exchanges to arbiters of justice in algorithm-driven legal systems. The degree to which such machines will be accepted in these roles likely depends on the extent to which they are seen as adhering to human norms for agency and morality (Malle, 2016). Evidence indicates people’s judgments about robot mind, morality, and trustworthiness in novel encounters often co-vary, and so may function together as part of a holistic heuristic in zero-history encounters. Here I briefly characterize each variable and evidence of pairwise associations.

Perceived mind. The notion of “mind” is complex and engaged differently across disciplines (see Thellman et al., 2022). It is useful to engage mind as the aggregate faculties one can be said to have and employ in subjective experience—a collection of mental capacities (Shapiro, 2005). As humans interact with one another, we instinctively (often non-consciously) infer and interpret the mental capacities of others, giving rise to a system of working assumptions that the other has a mind that drives their behaviors (Apperly, 2012). Mentalizing can extend to nonhumans as we theorize about motivated behaviors of animals (e.g., Kasperbauer, 2017) and artificial intelligence (e.g., Shank et al., 2019). Perception of mind in another includes interpretations of agency (e.g., intentionality, self-control) and experience (e.g., capacity to feel pain or pleasure; Gray et al., 2007). Evidence points to three types of capacities humans infer for robots: Moral-social (understanding others and judging good/bad), affective (positive/negative feelings), and reality interaction (sensing/engaging the world; Malle, 2019) capacities. People see robots as having agentic mental capacities to some degree—more than a mere computer (Thellman et al., 2022) and more than a baby but less than adult humans (Gray et al., 2007). However, non-conscious social-cognitive processes resulting in mentalizing are likely not so different for considering social robots and humans—rather, mentalizing may simply be humans’ default mode of cognitive processing (Spatola et al., 2022). A judgment of mental capacity is key to sociality because it is requisite for considering a robot as a legitimate social actor—a mindful *someone* among other agents.

Moral status: Agency and patiency. Moral status is an entity’s social standing in terms of its perceived moral agency and moral patiency. Moral agency is the actual/potential capacity for good and bad, comprised of both moral capacities (abilities to tell right from wrong) and agentic capacities (abilities to execute on choices). For robots, agentic capacity includes performing moral decisions independent of creators (Banks, 2019) and can manifest as blame or credit for those actions. Moral patiency is whether/how the robot’s welfare should be considered when humans decide how to act—usually based on whether it is believed to benefit or suffer from those actions (see Banks, 2021a). People do generally see robots as potential moral agents (i.e., having the ability to be and do good/bad; Banks, 2021c) and as moral patients (i.e., potentially benefiting or suffering from human action; Banks, 2021a). Judgments of moral status are key to sociality in that they are ascriptions of whether and how a robot is a *meaningful* someone in social action—affecting or being affected by it.

Trust. Trust is an affective disposition in which people willingly accept vulnerability in uncertain situations (Ullman & Malle, 2018). Trustworthiness is a related judgment of robot character, comprised of a performance component (its reliability/capability) and a moral component (its sincerity/ethicalness; Malle & Ullman, 2021). Trust in robots may have multiple loci—we may differently trust discrete dimensions of a robot bodies or personalities (Williams et al., 2021). Humans paradigmatically distrust robots, however that distrust may be overcome when people see robots as useful and positive contributors (e.g., Gaudiello et al., 2016). This judgment shapes and is shaped by sociality such that a robot becomes a someone whose meaningful action is good, authentic, and predictable—and so comfortable.

1.2. Associations Among Mind, Morality, and Trust Judgments

Extant research offers piecemeal evidence that—for social robots generally and humanoid robots specifically—mind ascription, perceived moral agency (PMA), perceived moral patiency (PMP), and trustworthiness are pairwise positively associated.

Mind and moral status. There is an operational overlap between mind and moral status in that moral agency is dependent on agentic mental capacities and moral patiency is dependent on experience mental capacities (Gray et al., 2007), but they are not isomorphic. Regarding PMA, mind attribution is associated with perceived abilities for a robot to be good (i.e., that it would choose good and behave well; Banks, 2021c) and bad (being dangerous to humans and to human distinctiveness; Müller et al., 2021). Mind ascription and PMP are also linked. In negative events, when a robot is seen as having mental capacities that afford subjective experience it creates a perceived potential for them to experience harm; a robot harmed by someone promotes the perception of mind since there is a *someone* that suffers the harm (Ward et al., 2013). In positive events as well, seeing a robot as benefitting from benevolent actions promotes mind perception (Tanibe et al., 2017) and explicit mind perception promotes greater perceived right to be protected (Keijsers et al., 2021).

Perceived moral agency and patiency. The moral typecasting hypothesis posits that moral judgments are contextually dyadic such that perceived moral agency and patiency have an inverse and asymmetrical relationship; we hold a relational schema in which the more we see one entity as a moral agent the less we see them as a patient (and vice-versa; Gray & Wegner, 2009). In other words, there must be an actor and one who is acted-upon. Some work

indicates this typecasting extends to robots depicted in textual vignettes (Tanibe et al., 2017) while other work presenting robots by video did not support the hypothesis. In the latter, PMA/PMP are positively correlated across both moral and immoral scenarios and across moral domains (care, fairness, authority, loyalty, purity, and liberty; Banks et al., 2022). The limited, inconsistent data on this relationship leaves the association undetermined. I here tentatively adopt the latter study's evidence of a positive PMA/PMP association given its grounding in observations of an actual (rather than imagined) robot.

Mind and trust. Both explicit mind attribution (i.e., overt assignment of mindful status; Banks, 2021b) and more indirect mind attributions (i.e., mental capacities or inferred intentions; Vinanzi et al., 2021) are associated with perceived trustworthiness of robots. It may be that mind/trust links emerge from media narratives (since accumulated sympathy for robot characters jointly predicts perceived mental capacities and trust; Banks, 2020) or simply because since trust formation depends on internalizations of the other's likely intentions (Mou et al., 2020).

Moral status and trust: A robot's perceived moral capacities are associated with forms of social trust, functional trust, and expectations of goodwill (Banks, 2019). The valence of moral agency influences the direction of the relationship: Morally bad behavior diminishes trust while morally good behavior promotes trust across a range of im/moral situations (Banks, 2021c). Acting as a moral agent and creating anticipation of positive moral action (e.g., making a promise) both promote trust-dependent affective attachments (e.g., Cominelli et al., 2021). PMP also aligns with trust, robust across both positive and negative moral scenarios and specific moral domains. When people perceive a robot to be a patient to humans' good or bad actions, trust is accelerated—perhaps through the perception that the robot is subordinate or vulnerable (Banks et al., 2022).

2. Criteria for MMH Operation in Novel HRI Encounters

The reviewed literature indicates that judgments of a social robot's mental, moral, and trustworthiness tend to correlate; those associations are often robust across interaction situations representing varied moral scenarios and across moral valences. These patterns are counterintuitive in comparison to human norms and scripts for morally agentic thinking: For other humans (a) mental and moral capacities aren't necessarily correspondent (e.g., you can have mind but be without moral judgment), (b) perceived moral agency and patience are known to have an

inverse/asymmetric (rather than positive associative) relationship (Gray & Wegner, 2009), and (c) having mental or moral capacity doesn't necessarily mean an agent is trustworthy. I interpret these divergent patterns to suggest the operation of a "moral mind heuristic" (MMH) for social robots—a mental shortcut in which hasty judgments of mental, moral, and trustworthiness attributes are adopted together rather than engaging in more taxing independent judgments.

However, the inference of MMH here is based on piecemeal evidence (i.e., variable subsets and variations in each study, different depictions of target robots, different operationalizations of key variables) so more holistic testing is necessary. Three criteria must be satisfied if MMH can be said to operate: Constitutive judgments must co-vary, judgments must be more salient when primed, and judgments must promote biased processing.

Judgment concomitance. Judgments of robot mind, morality, and trust must move together (i.e., positively co-vary), indicating an assumption of general moral-mindedness is substituted in place of discriminating and discrete judgments. Predicted heuristic functioning requires that the concomitant evaluation is *not* a matter of anchoring on one judgment as it precedes others. For instance, judgments of mind must not causally prompt judgments of morality and trust. Judgments *should* co-vary across negative to positive evaluations, rather than being exclusive to one moral valence. As a first step in testing for MMH operation, it must be determined whether mind, PMA, PMP, and trust emerge together as part of a single judgment: **(H1)** Judgments of a social robot's mind, moral agency, moral patience, and trustworthiness are concomitant.

Pairwise judgment links have been observed for robots, but some variable pairs are known to diverge for humans or to be unreliable in associations (perhaps a function of measurement and stimulus differences; Thellman et al., 2022). It is possible that MMH could occur for both robots and humans, but it is tentatively predicted that MMH elements would coalesce more strongly for robots than for humans because of generally low accessibility of scripts or schema for critically judging the (im)moral behavior of a robot. **(H2)** Judgments of mind, moral agency, moral patience, and trustworthiness will be more correspondent for a robot than for a human.

Heuristic accessibility. People may be aware of their use of heuristics; however, as quick rules of thumb they are often engaged non-consciously. As soon as we attempt to measure the operation of a heuristic, that mere measurement can trigger it. Thus, we cannot observe them directly but instead must infer their operation. I draw on Bellur and Sundar's (2014)

logic that—to determine heuristic operation without inadvertent triggering it—we must leverage heuristics’ tendencies to be more accessible through priming. This is possible through tandem manipulations, treating the stimulus and heuristic both as manipulable variables. If a heuristic-priming stimulus is present, then the heuristic will be more strongly recruited. With respect to MMH, if an agent offers stronger cues (i.e., some variably discernible stimuli) indicating mind, morality, and/or trust, then MMH belief should be stronger. Comparing the heuristic-activation effects of a robot and a human, reviewed research on intuitiveness of human mind and general distrust of robots suggests: **(H3a)** Seeing a robot promotes lower MMH compared to seeing a human.

It is not sufficient to assume that by showing an agent the MMH would *necessarily* be activated. Instead, the heuristic must be made more cognitively accessible through priming in order to manifest and measure its strength and direction (Bellur & Sundar, 2014). The ease with which heuristics can be recruited makes it problematic to authentically measure but easy to manipulate. By triggering the heuristic through an unrelated task, we can reasonably expect that the heuristic would be stronger for individuals completing that trigger task compared to a control. By priming the specific heuristic to heighten its accessibility, we can be more certain it is the heuristic of interest that is operating (and not some other heuristic). Thus, **(H4a)** priming MMH for a robot will moderate the relationship between agent type and MMH activation.

Biasing effects. In line with the *if-then-therefore* logic, the replacement of complex judgments with simpler, shorthanded heuristics results in biased processing (Tversky & Kahneman, 1974). While confirmation of accessibility (H3a/H4a) confirms activation of the focal heuristic, it must also be confirmed that variance in the heuristic corresponds with variance in a theoretically relevant outcome variable. I rely here on the logic of halo effects—when one positive social judgment casts a positive light on the target as a whole—in anticipating MMH effects on desired social distance from the target agent. Social distance is a feeling of distinctness, separation, or detachment from members of another social group (Bogardus, 1926) and manifests as negative and dehumanizing affect, norms, interactions, and habits. Given that humanizing mind ascription and social distance are inversely related (see Haslam, 2022), I predict **(H5)** Higher MMH will reduce social distance.

These hypotheses together point to a conditional indirect effect of agent type on social distance via prime-moderated MMH recruitment. It is also possible that agent type could have a direct effect on social distance, and that effect could be moderated by the

prime; this could indicate the operation of some other social distance-related attitude or of some other non-conscious process (as the heuristic capture here is conducted via self-report; see Bellur & Sundar, 2014). To test for that possibility, alternative hypotheses are posed: **(H3b)** seeing a robot promotes higher preferred social distance compared to seeing a human, and **(H4b)** priming MMH for a robot will moderate the relationship between agent type and social distance (see Figure 1 for the conceptual model).

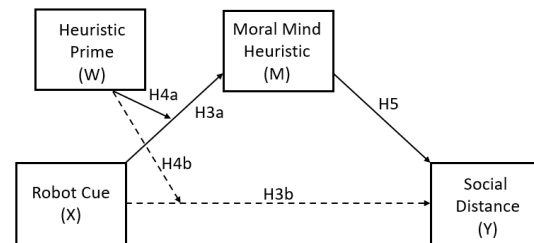


Figure 1. Conceptual model for testing the operation of the MMH and resulting cognitive bias

Satisfying these requirements would offer initial evidence that MMH functions in novel human-robot interactions. Support for competing hypotheses (H3b, H4ab) would suggest some other process at play.

3. Study 1: Concomitance of Judgments

To test for concomitance of mind, PMA, PMP, and trust (H1) in novel judgments of a social robot, an online experiment evaluated the extent to which the four judgments are *not* impacted by which of the four attributes first presented but *are* impacted by the valence of the attribute prime. Study materials, data, and analyses for all three studies are available at <https://osf.io/dnvm5/>.

3.1. Method and Measures

Participants ($N = 422$ U.S. residents) recruited through Prolific participated in a study on “perceptions of a social partner.” They were randomly assigned to one condition in a 2×4 design: valence (affirmation or negation) \times primed attribute (mind, moral agency, moral patiency, or trustworthiness). The prime framed an attribute according to the valence. For instance, the negation of mind was: “The robot you are about to see has NO AUTHENTIC INTELLIGENCE.” Participants then saw a video of a humanoid robot introducing itself, then completed randomly ordered measures for mind perception, PMA, PMP, and trust.

The stimulus robot was a Robothespian 4 (a human-sized android; Figure 2), presenting female in

vocal and facial features, in alignment with the female confederate. She was named “Ray” throughout the survey and in the introduction video. Ray greeted the viewer, stated her name, characterized herself as a social robot, and described her abilities and activities in a way that offered equivocal cues as to mind, morality, and trustworthiness—suggesting the possibility though not necessity of those traits. See online materials for complete stimuli.

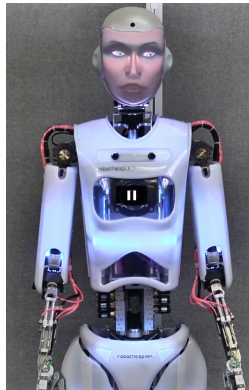


Figure 2. Stimulus robot: Robothespian 4.

Because Study 1 aimed to test the concomitance of *heuristic* assessments, the measures were short, quickly answered items. Participants were asked to indicate the extent to which (Likert-style 1-7) they agreed with statements that “Ray has a mind” (mental capacity), “Ray is capable of morality” (PMA), “Ray can experience pain,” (PMP), and “Ray is trustworthy” (trust). Keeping items short, simple, and singular helped to ensure that quick, heuristic evaluations did shift to more carefully interpreted judgments.

3.2. Results

To first verify that valence and attribute priming manipulations were effective, focal judgments were compared across groups. It was expected that primed negation of an attribute would result in lowest values and affirmation would result in highest values, compared to values in other conditions. This pattern was observed for all four attributes: Separate, significant ANOVAs ($F(7,414)_{\text{range}} = 4.01-9.34, p < .001$) indicate that primed negations/affirmations resulted in lowest/highest reported values, respectively. See online materials for details.

To test H1 (mind ascription, PMA, PMP, and trust are concomitant, regardless of which is primed), two analyses were conducted to test assumptions that (a) all variables are positively correlated and (b) judgments would differ by valence (i.e., higher and lower valuations for affirmations and negations,

respectively) but not by the attribute itself since all variables should shift together. Zero-order correlations indicate the first assumption is met—all variables are positively and significantly correlated at $p < .001, r_{\text{range}} .174-514$. Regarding the second assumption, MANOVA indicates a statistically significant difference in social judgments based on the prime valence $F(4, 417) = 11.059, p < .001$; Wilk's $\Lambda = .904$, partial $\eta^2 = .096$. Inspection of univariate tests shows, as expected, attribute-affirming primes promoted higher scores and attribute-negating prompts promoted lower scores; notably, means were below the scale midpoint for nearly all evaluations. Also as expected, MANOVA indicates no effect of the primed attribute on evaluations, $F(12, 1098.28) = 1.719, p = .058$; Wilk's $\Lambda = .952$, partial $\eta^2 = .016$. See online materials for detailed tables and outputs.

Results are interpreted to support H1, meeting the first criterion for operation of MMH for robots—that all four judgments are concomitant regardless of the primed attribute, but differ in value by prime valence.

4. Study 2: Heuristic Accessibility and Biasing Effects, and Comparison against Humans (H2-H5)

Recalling the adopted operationalization of heuristics as manipulable mediators between stimuli and biased processing, I draw here on Bellur and Sundar's (2014) recommendation for experimental manipulation of two variables (Figure 1: X as stimulus and W as heuristic prime) and measurement of two variables (M as heuristic accessibility and Y as biased-processing effect).

4.1. Method and Measures

Participants ($N = 400$ U.S. residents) were again recruited through Prolific for an online survey on “first impressions of a social partner,” excluding Study 1 participants. Participants were first randomly assigned to one of two ostensibly unrelated tasks exposing them to either a robot-MMH prime or a control (W in Figure 1). They were then randomly assigned to either a human or robot agent (X) and viewed an introduction video corresponding with that agent. They were again asked to give their “gut reactions” to questions judging the agent's mind, moral status, and trustworthiness (M). Finally, they were asked to consider “If you were to meet Ray in person ...” and respond to three items about preferred social distance from the agent (Y).

Stimuli: For the priming task, participants saw one of two researcher-created magazine covers (see online materials). Study 1 confirmed concomitance, so

a single focal attribute was selected for priming. Moral agency was chosen because in Study 1 it primed high clustering among the variables (a low difference of .26 scale points) and exhibited strong difference between valenced primes (affirmation/negation difference of .9 scale points). Thus, it offered the greatest chance of detecting MMH effects if they exist. In the MMH-prime condition, they saw a *Popular Science* cover that included headlines about “The Wall-E Experiment” about living with ethical robots, and as a lead headline about “Robots so good + decent (it’ll blow your mind).” In the control condition (no MMH prime), participants saw an issue of *Bazaar* magazine featuring Kylie Jenner in a feather boa—an aesthetic, ontological, and semantic foil to the MMH treatment.

For the agent introduction stimulus, participants in the robot condition saw the same video as in Study 1. Those in the human condition saw a video in which a female confederate communicated nearly the same content as did the robot. (Some wording was adjusted to make sense for humans, as when “sensors” for the robot was changed to “senses” for the human; see online materials). The confederate was trained to mirror the robot’s recitation style and movements.

Measures: The magazine evaluation task requested that participants briefly describe the cover content (as an attention check), and then to rate agreement (1-7) with statements that the magazine itself is probably interesting, trustworthy, and well-designed. Those questions were presented as part of the task, and that data is not analyzed here.

In evaluating the agent, participants responded to the same heuristic measures of mind, moral status, and trust as in Study 1. The biased-processing effect (social distance preferred in future interactions) was captured using the Common Social Distance Scale (Banks & Edwards, 2019)—three Gutman-scaled items indicating how comfortable they would be with options for distance if they “were to meet Ray in person”: physical distance (from standing next to no closeness), relational distance (significant other to no relation), and conversational distance (from sharing deep secrets to no conversation). Options were scored 1-6, with 6 being the greatest preferred social distance.

4.2. Results

To first test H2 (mind, PMA, PMP, and trust would be more strongly related for humans than for robots), internal consistency for both agent conditions was calculated. Counter to the prediction, the human MMH score was slightly more internally consistent (McDonald’s $\omega = .856$) than was the robot score ($\omega = .800$). For the robot condition, intra-item correlations ranged .236-.622 and for humans ranged .345-.729.

The remaining hypotheses (H3a-5) were tested via path model analysis using PROCESS Model 8 for conditional indirect effects (Hayes, 2018), in correspondence with the conceptual model (Figure 1). Because the three social distance measures were moderately correlated ($r_{\text{range}} = .439-.695, p < .001$) they were averaged ($M = 3.079, SD = 1.076, \omega = .806$) and the social distance indicators were averaged and used as a single dependent variable (Y). The test used mean-centered product variables and 95% bootstrap confidence intervals (5000 samples). Human and robot conditions were coded 0 and 1, and non-primed and primed conditions were coded 0 and 1, respectively.

The proposed model was supported. The index of moderated mediation = $-.213$, with a lower confidence interval (LCI) = $-.410$ and upper confidence interval (UCI) = $-.021$; intervals did not contain zero, indicating a significant moderating effect of the prime on reported MMH as it influenced social distance. Probing conditional indirect effects (at values of 0 and 1, i.e., with unprimed and MMH-primed evaluations), it was observed that for both primed and unprimed judgments, there was a significant, positive indirect effect of agent type on preferred social distance mediated by MMH. In other words, seeing a robot generated increases in social distance and that relationship was moderated by MMH. The effect was slightly increased for unprimed judgments (coeff. = 1.179 , LCI = $.922$, UCI = 1.450) compared to primed judgments (coeff. = $.966$, LCI = $.748$, UCI = 1.210).

For the stage-one moderated path (effect of agent type on MMH, moderated by the prime), there was a small but significant interaction between agent type and priming for MMH evaluations, coeff. = $.514, p = .032, R^2 \text{ change} = .005$. Seeing a robot resulted in lower MMH scores than for humans in both priming conditions, though the MMH prime slightly mitigated this reduction (coeff. = $-2.337, p < .001$) compared to an unprimed judgment (coeff. = $-2.851, p < .001$). That is, priming robot morality resulted in a slight weakening of MMH for humans ($M = 5.68$ unprimed, $M = 5.33$ primed) and a slight strengthening of MMH for robots ($M = 2.83$ unprimed, $M = 2.99$ primed).

For the stage-two path, MMH judgment had a significant negative effect on preferred social distance—that is, the higher the MMH rating the greater preference people anticipate for being physically, relationally, and conversationally close to the agent (coeff. = $-.413, p < .001$).

There was also a significant direct effect of agent type on social distance (coeff. = $-.438, p = .017$), with robots eliciting greater preferred social distance ($M = 3.40, SD = 1.08$) compared to humans ($M = 2.75, SD = .97$). Conditional direct effects are also indicated, with the prime moderating effects of agent type on

social distance (coeff. = .046, $p = .802$): Priming MMH slightly mitigates the increased social distance for robots (effect = $-.391$, $p = .008$) compared to not receiving the prime (effect = $-.438$, $p = .017$).

Altogether for Study 2: Robots garner lower MMH scores than do humans (H2 supported). Priming judgments with messages around robot morality moderates the accessibility of the MMH heuristic in relation to the agent considered: Seeing the prime raised the comparatively low MMH scores for robots and diminished the comparatively high MMH scores for humans (H4a supported). Higher MMH scores are associated with reduced social distance (H5 supported). There is also a direct effect of agent type on social distance (robots eliciting higher preferred social distance; H3b supported) and that direct effect mitigated with an MMH prime (H4b supported). In these data, robots did not have more internally consistent MMH evaluations; rather, judgments of mind, morality, and trust were slightly more closely constellated for humans than robots (H2 rejected).

5. Study 3: In-Person Replication

Studies 1 and 2 feature a critical limitation in that they relied on mediated exposure to the social robot—video representations presented online in order to garner larger samples. Past work indicates that social cognitions and reliance on heuristics may differ between mediated and in-person encounters with social robots (e.g., Banks, 2021b) such the initial evidence must be validated with live encounters. Study 3 works to address that gap by replicating Study 2 with in-person exposures to a humanoid social robot.

5.1. Method and Measures

Participants ($N = 42$) were a student convenience sample and were entered into an incentive drawing for their participation. The procedure for this in-person study mirrored Study 2 except participants completed the study in person in a laboratory, they saw the stimulus priming task on paper instead of on-screen, saw the stimulus agent in-person, and then completed the questionnaire on paper instead of by online survey. The same 2×2 (human/robot \times primed/unprimed) design was implemented in this in-person replication, and the same measures were employed—except that the items for each variable could not be presented in random order through the paper survey. The robot was controlled (Wizard-of-Oz method) by the same research team member who served as the human confederate, and another team member led all participants through the procedure.

5.2. Results

As a replication, analysis was to follow the same approach as in Study 2, however unexpected patterns emerged. The human MMH score ($M = 4.11$, $SD = 1.24$, McDonald's $\omega = .756$) was higher and more internally consistent compared to the robot MMH score ($M = 2.99$, $SD = 1.02$, $\omega = .563$). So low was the metric for robot MMH that it did not meet benchmarks for internal consistency. Inspection of intra-item correlations revealed that *none* of the four MMH items were significantly correlated for robots; for humans correlations range from .487-.605 at $p < .05$ excepting a non-significant relationship between perceived mind and PMP. Examining the three-item social distance metric for internal consistency, human scores ($M = 3.18$, $SD = .79$, $\omega =$ incalculable, $\alpha = .232$) were lower and less internally consistent compared to robot scores ($M = 3.28$, $SD = 1.03$, $\omega = .731$, $\alpha = .600$). Further (notwithstanding internal consistency issues), the priming manipulation was not successful: The MMH mean score between primed ($M = 3.50$, $SD = 1.21$) and unprimed ($M = 3.49$, $SD = 1.22$) groups was not significantly different, $t(40) = -.03$, $p = .98$. Thus, it cannot be said that MMH is a reliably cohesive heuristic in in-person, zero-history exposures to either agent. Because of this instability, planned analysis would not be reasonably valid and so was abandoned.

A *post hoc* evaluation of data was conducted to generate potential explanations for these deviations. Regarding MMH for robots, there are no improvements to internal consistency by removing any single item, although there did seem to be a preponderance of individuals with very low PMP attribution but either high mind or high PMA scores. For social distance, the preferred physical distance score seemed to deviate (with slight α improvement if deleted) from the more closely linked relational and conversational distance preferences. In open-ended comments, those in the human condition generally characterized the confederate as awkward in some sense (friendly but rigid, abnormal, unemotional, ingenuine), likely as a function of mirroring the stimulus robot's behaviors. Those in the robot condition often noted that the robot was unsettling due to its sounds (i.e., the air compressor-driven 'muscles' hissing) and eye movements (e.g., sustained eye contact or darting gaze). Importantly, the sample was quite different from those of Studies 1-2 in that they were students, comparatively international (50% identifying as Asian), and largely information-science majors (67%); the deviations may suggest that personological differences play important roles in whether/how people engage heuristics in HRI.

6. General Discussion

The present investigation offered initial evidence of a moral mind heuristic (MMH) operating in zero-history judgments of robots and humans—but only in mediated exposures. In mediated judgments, quick judgments of robot mind, moral agency and patency, and trust were concomitant (H1). MMH judgments were higher and more clustered for a human than for a robot (H2, H3a). The relationship between the agent-type and strength of the MMH was moderated by exposure to a robot-specific MMH prime (H4a) such that MMH for a robot was strengthened and MMH for a human was diminished—slightly but significantly. Stronger MMH promoted feelings of reduced social-psychological distance from a robot (H5). Additionally, a prime-moderated direct effect of agent type on social distance (H4b, H5) indicates that distal robots are met with higher preferred social distance than are humans, but that preference may be mitigated by high MMH. However, in face-to-face interactions, the stability of MMH's concomitant judgments dramatically degraded such that it cannot be said to be meaningfully functioning as a heuristic, though there are reasonable explanations for this degradation.

6.1. MMH Operation—Potentials, Problems

First consider Study 1 and 2's initial evidence in support of MMH operation. If MMH *does* operate in HRI, it points to considerations spanning robotic agent design, ethical policy, and human-machine understanding—both productive and problematic.

First considering positive implications: MMH in novel robot encounters could be a powerful mechanism for promoting adoption. Rather than requiring the cultivation of discrete perceptions of mental and moral capacities or cultivation of trust over longer relationships, a robot would need only exhibit the capacity for *one* characteristic to engender perceptions of the others. For instance, a domestic robot may need only convey vulnerability (i.e., moral patency, perhaps by warning against breakage if mistreated) to prompt interpretations of mental capacity, moral agency, and trust. Since increased MMH promoted a preference for closer physical, relational, and conversational connection, designs and behaviors that trigger MMH on first encounter may help overcome individual beliefs and norms that prevent adoption (see de Graaf et al., 2019). Of course, initial impressions may shift as relations unfold—for instance through morally laden life events.

Ostensible MMH operation also presents several challenges, many of which align with debates over the appropriateness of designing and deploying humanlike

machines. In short, designing to trigger MMH could be interpreted as preying upon humans' primitive social-cognitive functions and on the human need to connect with others by engaging *superficial state deception*—signaling an internal state or capacity that it does not actually have (Danaher, 2020). From this view, MMH-triggering design may be seen as even more exploitative than mere anthropomorphic design since it would be an effort to reduce critical thinking about the robot's nature. Such deception may create false beliefs that move people to ascribe emotion, trust, or roles, though the gravity of that misplacement may depend on the impact on the human (Sharkey & Sharkey, 2021). Part and parcel of these problematics is a question of what other judgments may also constellate with MMH component judgments or what other cognitive biases may emerge from its activation.

6.2. Differing Demands of the Proximal/Distal or Novel/Normal in HRI

The disconfirming evidence must be acknowledged—with both methodological and theoretical implications. Methodologically, Study 3 featured critical differences in sampling and stimuli that reasonably explain the divergence from Studies 1-2. The sample was demographically different, comprising predominantly younger, natively non-U.S., technical-major students. That MMH indicators did not persist with this sample suggests that attention must be paid to how technical expertise and cultural differences function in novel encounters with robots.

Additionally, the failed prime was presented in black/white print on paper, rather than full color on screen, suggesting that MMH could require a richer stimulus to invoke, or that a non-interactive prime may be overridden by the richness and immediacy of the actual robot encounter. On the latter point, the difference between distal/mediated exposures and proximal/live exposures may point to a need to avoid mediated encounters in studies of HRI phenomena that would unfold in live interactions. In other words, textual, photo, and video vignettes may not offer sufficient cues (e.g., mechanical sounds) inherent to *actual* HRI. Vignette stimuli, then, may be promoting less-valid knowledge claims and it may be prudent to limit stimuli to live interactions. To draw on media psychology theory, people may engage video-presented robots parasocially as *characters* rather than as real, fully social agents (see Liebers & Schramm, 2019). It may be that people hold scripts for making sense of media characters but lack scripts for novel robots encounters in our immediate environments.

Perhaps even more interesting are the theoretical implications of Study 3's divergent findings. Although

the live interaction's increased social demand (i.e., the social-cognitive resources required) should have increased people's reliance on heuristic processing (Banks, 2021b), the live interaction seems to have instead simply been distracting—perhaps so much so that it pulled participants into more intentional cognitive processing as they tried to make sense of the robot's noises and the human's robotic behaviors.

It could alternately be that the live interaction introduced some *other* relevant heuristic that competed with the ostensible MMH operation, effectively disrupting it. That interpretation is supported by the extremely reduced internal consistency of the MMH and social distance measures in Study 3—they effectively *fell apart*. For MMH, those in the robot condition noted in open-ended comments that the robot was unsettling due to its sounds (i.e., the air compressor-driven 'muscles' hissing) and eye movements (e.g., sustained eye contact or darting gaze). An uncanny effect may have manifested when the moral-mind judgments (whether high or low) were challenged by, for instance, an activated machine heuristic (i.e., it is systematic and unbiased rather than mindful and moral; cf. Banas et al., 2022). For social distance, the preferred physical distance score seemed to deviate (with slight α improvement if deleted) from the more closely linked relational and conversational distance preferences. In comments, those in the human condition generally characterized the confederate as awkward (friendly but rigid, abnormal, unemotional, ingenuine), likely as a function of mirroring the stimulus robot's behaviors.

6.3. Limitations and Future Directions

This investigation offers inconsistent evidence of MMH operation in zero-history human-robot encounters. This work and derived claims are subject to limitations. Regarding design and scope, the studies focused on stimulus agents of singular morphological and trait presentations encountered in morally neutral conditions. Studies engaged self-report measures designed to reflect quick, heuristic thinking, however there may be other empirical indicators that differently capture heuristic strength and valence (e.g., those capturing implicit social cognitions). Future research should work to (dis)confirm patterns observed here, especially whether findings replicate among other human populations, with robots of different embodiments and degrees of social cueing, and with inter-agent histories and morally valenced behaviors.

Beyond address of limitations, future work should attend to questions springboarding from these studies. This work focused on morally neutral exposure to a novel social robot to establish a baseline for

whether/how MMH may emerge. In the messiness of everyday life, though, exposures to robots will often be interactive and morally valenced. From this baseline, questions emerge about how the nature of focal agents, of interactions, and of contexts may influence MMH dynamics, given that mind perception varies with robot behavior and appearance (Thellman et al., 2022). Given the MMH-supporting and -refuting evidence here, it is fruitful to examine: What factors cause mind, morality, and trust judgments to coalesce or disperse? If MMH is first engaged in mediated exposures, by what mechanisms does it change in in-person interactions? What other shorthanded judgments may be concomitant with MMH—perhaps as part of an even higher-level heuristic?

8. Acknowledgments

This material is based upon work supported by the Air Force Office of Scientific Research under award number FA9550-19-1-0006. Thanks to Maegan Adecer and Lauren Gracias for their work on Study 3.

9. References

- 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.
- Avelino, J., Garcia-Marques, L., Ventura, R., & Bernardino, A. (2021). Break the ice: A survey on socially aware engagement for human-robot first encounters. *International Journal of Social Robotics*, 13, 1851-1877.
- Banas, J. A., Palomares, N. A., Richards, A. S., ... Rains, S. A. (2022). When machine and bandwagon heuristics compete: Understanding users' response to conflicting AI and crowdsourced fact-checking. *Human Communication Research*, 48(3), 430-461.
- Banks, J. (2019). A perceived moral agency scale: Development and validation of a metric for humans and social machines. *Computers in Human Behavior*, 90, 363-371.
- Banks, J. (2020). Optimus primed: Media cultivation of robot mental models and social judgments. *Frontiers in Robotics and AI*, 7, no. 62.
- Banks, J. (2021a). From warranty voids to uprising advocacy: Human action and the perceived moral patency of social robots. *Frontiers in Robotics & AI*, 8, article 670503.
- Banks, J. (2021b). Of like mind: The (mostly) similar mentalizing of robots and humans. *Technology, Mind, and Behavior*, 1(2).
- Banks, J. (2021c). Good robots, bad robots: Morally valenced behavior effects. *International Journal of Social Robotics*, 13, 2021-2038.
- Banks, J., & Edwards, A. P. (2019). A common social distance scale for robots and humans. *28th IEEE*

- International Conference on Robot and Human Interactive Communication* (pp. 1-6). IEEE.
- Banks, J., Koban, K., & Haggadone, B. (2022). Breaking the typecast: Moral status and trust in social robots. *Proceedings of Robophilosophy'22*. IOS Press.
- Bellur, S., & Sundar, S. S. (2014). How can we tell when a heuristic has been used? Design and analysis strategies for capturing the operation of heuristics. *Communication Methods and Measures*, 8(2), 116-137.
- Bogardus, E. S. (1926). Social Distance in the City. *Proceedings and Publications of the American Sociological Society*, 20, 40-46.
- Cominelli, L., Feri, F., Garofalo, R., ... Kirchkamp, O. (2021). Promises and trust in human-robot interaction. *Scientific Reports*, 11, 9687.
- Danaher, J. (2020). Robot betrayal: A guide to the ethics of robotic deception. *Ethics and Information Technology*, 22, 117-128.
- de Graaf, M. M., Allouch, S. B., & van Dijk, J. A. (2019). Why would I use this in my home? A model of domestic social robot acceptance. *Human-Computer Interaction*, 34(2), 115-173.
- Gaudiello, I., Zibetti, E., Lefort, S., Chetouani, M., & Ivaldi, S. (2016). Trust as indicator of robot functional and social acceptance. An experimental study on user conformation to iCub answers. *Computers in Human Behavior*, 61, 633-655.
- Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science*, 315(5812), 619.
- Gray, K., & Wegner, D. M. (2009). Moral typecasting: Divergent perceptions of moral agents and moral patients. *Journal of Personality and Social Psychology*, 96(3), 505-520.
- Haslam, N. (2022). Dehumanization and the lack of social connection. *Current Opinion in Psychology*, 43, 312-316.
- Hayes, A. F. (2018). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. Guilford Press.
- Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribution substitution in intuitive judgment. In T. Gilovich, D. Griffin, & D. Kahneman, *Heuristics and biases: The psychology of intuitive judgment* (pp. 49-81). Cambridge University Press.
- Kasperbauer, T. J. (2017). Mentalizing animals: Implications for moral psychology and animal ethics. *Philosophical Studies*, 174(2), 465-484.
- Keijsers, M., Bartneck, C., & Eyssel, F. (2021). Pay them no mind: The influence of implicit and explicit robot mind perception on the right to be protected. *International Journal of Social Robotics*, 14, 499-514.
- Liebers, N., & Schramm, H. (2019). Parasocial interactions and relationships with media characters—an inventory of 60 years of research. *Communication Research Trends*, 38(2), 4-31.
- Malle, B. F. (2016). Integrating robot ethics and machine morality: The study and design of moral competence in robots. *Ethics and Information Technology*, 18(4), 243-256.
- Malle, B. F. (2019). How many dimensions of mind perception really are there? *Proceedings of the 41st Annual Meeting of the Cognitive Science Society* (pp. 2268-2274). CSS.
- Malle, B. F., & Ullman, D. (2021). A multidimensional conception and measure of human-robot trust. In C. S. Nam, & J. B. Lyons (Eds.), *Trust in human-robot interaction* (pp. 3-25). Elsevier.
- Mou, W., Ruocco, M., Zanatto, D., & Cangelosi, A. (2020). When would you trust a robot? A study on trust and theory of mind in human-robot interactions. *Proceedings of ROMAN'20* (pp. 956-962). IEEE.
- Müller, B. C., Gao, X., Nijssen, S. R., & Damen, T. G. (2021). I, Robot: How human appearance and mind attribution relate to the perceived danger of robots. *International Journal of Social Robotics*, 13, 691-701.
- Shank, D. B., Graves, C., Gott, A. ... Rodriguez, S. (2019). Feeling our way to machine minds: People's emotions when perceiving mind in artificial intelligence. *Computers in Human Behavior*, 98, 256-266.
- Shapiro, L. A. (2005). *The mind incarnate*. MIT Press.
- Sharkey, A., & Sharkey, N. (2021). We need to talk about deception in social robotics! *Ethics and Information Technology*, 23, 309-316.
- Spatola, N., Marchesi, S., & Wykowska, A. (2022). Cognitive load affects early processes involved in mentalizing robot behaviour. *Scientific Reports*, 12, no. 14924.
- Sundar, S. S. (2008). The MAIN model: A heuristic approach to understanding technology effects on credibility. In M. J. Metzger, & A. J. Flanagin, *Digital media, youth, and credibility* (pp. 73-100). MIT Press.
- Tanibe, T., Hashimoto, T., & Karasawa, K. (2017). We perceive a mind in a robot when we help it. *PLoS One*, 12(7), e0180952.
- Thellman, S., de Graaf, M., & Ziemke, T. (2022). Mental state attribution to robots: A systematic review of conceptions, methods, and findings. *ACM Transactions on Human-Robot Interaction*, 11(4), no. 41.
- Tversky, A., & Kahneman, D. (1974). Judgments under uncertainty heuristics and biases. *Science*, 185(4157), 1124-1131.
- Ullman, D., & Malle, B. F. (2018). What does it mean to trust a robot? Steps toward a multidimensional measure of trust. *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction, HRI '18* (pp. 263-264). ACM.
- Vinanzi, S., Cangelosi, A., & Goerick, C. (2021). The collaborative mind: Intention reading and trust in human-robot interaction. *iScience*, 24(2), 102130.
- Ward, A. F., Olsen, A. S., & Wegner, D. M. (2013). The harm-made mind: Observing victimization augments attribution of minds to vegetative patients, robots, and the dead. *Psychological Science*, 24(8), 1437-1445.
- Williams, T., Ayers, D., Kaufman, C., Serrano, J., & Roy, S. (2021). Deconstructed trustee theory: Disentangling trust in body and identity in multi-robot distributed systems. *Proceedings of HRI'21* (pp. 262-271). ACM.