International Journal of Social Robotics (2021) 13:2021–2038
https://doi.org/10.1007/s12369-020-00692-3

# Good Robots, Bad Robots: Morally Valenced Behavior Effects on Perceived Mind, Morality, and Trust

Jaime Banks[1]

Accepted: 12 August 2020 / Published online: 10 September 2020

**Abstract**
Both robots and humans can behave in ways that engender positive and negative evaluations of their behaviors and associated responsibility. However, extant scholarship on the link between agent evaluations and valenced behavior has generally treated moral behavior as a monolithic phenomenon and largely focused on moral deviations. In contrast, contemporary moral psychology increasingly considers moral judgments to unfold in relation to a number of moral foundations (care, fairness, authority, loyalty, purity, liberty) subject to both upholding and deviation. The present investigation seeks to discover whether social judgments of humans and robots emerge differently as a function of moral foundation-specific behaviors. This work is conducted in two studies: (1) an online survey in which agents deliver observed/mediated responses to moral dilemmas and (2) a smaller laboratory-based replication with agents delivering interactive/live responses. In each study, participants evaluate the goodness of and blame for six foundation-specific behaviors, and evaluate the agent for perceived mind, morality, and trust. Across these studies, results suggest that (a) moral judgments of behavior may be agent-agnostic, (b) all moral foundations may contribute to social evaluations of agents, and (c) physical presence and agent class contribute to the assignment of responsibility for behaviors. Findings are interpreted to suggest that bad behaviors denote bad actors, broadly, but machines bear a greater burden to behave morally, regardless of their credit- or blame-worthiness in a situation.

**Keywords** Morality · Moral foundations · Ontological categories · Trust · Social cognition

## 1 Introduction

Asimov's "Three Laws of Robots" [1] govern fictional robots' behaviors, and these laws persist in contemporary imaginaries about how robots should behave: do not injure humans, obey humans, engage in self-protection. How do humans respond, though, when a robot behaves "badly" by breaking these or other moral norms? Popular and scientific discourse alike attend to the potential for machine agency—the ability to variably act according to self-regulating abilities and intentions [2]—to engender both anxiety and acceptance of social machines. However, current human–robot interaction scholarship generally engages morality as holistic "goodness" or "badness" or reduces it to singular exemplars; this contrasts with contemporary moral psychology's increasing engagement of the construct as multidimensional [3]. Specifically, Moral Foundations Theory [4] parcels morality into five foundations (care, fairness, authority, loyalty, purity) and a sixth candidate foundation (liberty [5]). This study seeks to build on current understandings of how social judgments are impact by robots' (im)moral behaviors by assessing (a) how attributions of behavioral goodness and responsibility may vary by moral foundation and (b) whether foundation-specific attributions may differentially contribute to social evaluations of robots. Two studies (an online survey and a laboratory replication) indicate that judgments of (im)moral behavior may be agent-agnostic and that all moral foundations contribute to behavior and agent evaluations. However, physical presence and agent type play a role assignment of responsibility for those behaviors.

## 2 Review of Literature

Extant literatures suggest that fear, anxiety, and mistrust in robots sometimes manifest independent of particular behaviors (e.g., [6]); these negative dispositions could be a function

✉ Jaime Banks
j.banks@ttu.edu

[1] College of Media & Communication, Texas Tech University, Box 43082, Lubbock, TX 79409, USA

of the robot's cued ontological status (i.e., agent-category liminality) engendering a "wrong outside, wrong inside" heuristic [7, p. 44]. Human–robot interaction is governed by many of the same norms and expectations held for human–human interactions [8], but people hold ontological-class heuristics (cf. [9]) that introduce deviating expectations. For instance, people desire robots that are emotionally and socially warm and competent—more so than robots are generally seen as being [10]. When robots are unable to meet humans' expectations, any behavior that does not meet *both* normative and desired expectations may be seen as generally "bad." This perceived badness may erode trust (the acceptance of vulnerability and/or the expectation of reliability in the face of uncertainty [11]). Although people may have a "prevailing distrust" of robots, that distrust may be softened in situations where the robot exhibits efficient and accurate performances that are useful to humans [12, p. 649].

Such negative responses may be exacerbated when a robot is perceived to have behaved badly in overt and specific ways, as in the violation of a valued norm. Considerations of robot "badness" take at least two forms: functional and moral deviations (cf. [12]). Functional deviations are those in which the robot commits errors or behaves in ways that are technically or contextually inappropriate, such as forgetting information [13] or when a function-focused robot suddenly appears emotional [14]. Magnitude of robot errors are associated with the magnitude of lost trust [15]. Moral deviations—a focus of this investigation—are those in which the robot violates principles for right and good behaviors, as when a robot might harm humans [16] or become rebellious [17]. Less addressed in current literature are ways that robots may be perceived as morally good and may be assigned moral praise. Because robots must be perceived as morally competent if they are to be integrated into human society [18], more extensive exploration of human perception of robot behavior as both variably "good" and "bad" is warranted.

## 2.1 Moral Foundations as a Framework for Understanding Robots Behaviors

Humans variably ascribe moral status to other agents—including robots—based on exhibition of moral norms, vocabularies, cognitions, actions, and expressions [19]. Such ascription could not be monolithic because morality is not homogenous, so machine morality must be considered through a lens that accounts for individual, contextual, and cultural differences. A useful framework for undertaking that endeavor is Moral Foundations Theory (MFT [4]), positing that moral evaluations of events or agents are a function of intuitive, structurally evolved reactions to at least five moral foundations (upholding and violating pairs: care/harm, fairness/cheating, authority/subversion, loyalty/betrayal, purity/degradation) and a sixth candidate foundation (liberty/oppression [5]). These foundations form a moral "matrix" by which moral leanings vary across time and culture and in individual valuations [2, p. 125]. MFT rejects perspectives that rely exclusively on moral reasoning (i.e., good and bad are rationalized, post hoc [21]) in favor of moral intuition [22]. In other words, humans have gut reactions to situations that—according to valuations of specific, discrete moral fields—lead them to interpret those situations as good or bad; these immediate moral intuitions may be followed by moral reasoning (see [23]). Thus, MFT is a suitable framework for examining both implicit and explicit moral evaluations.

Although MFT has been suggested as a framework to consider robots as ideal moral agents [24] and engaged in relation to humans as they consider machine agents (e.g., [25]), few studies have formally applied MFT to perception of machine agents. Most notably, evidence suggests that AI violation of fairness, purity, or liberty norms in actual news events resulted in reduced goodness evaluations [26]. Generally, however, investigations of perceived machine morality rely on canonical psychological vignettes such as the trolley problem [27] which are subject to individual differences in moral-foundation valuation [28] for care and fairness. Although some contend that *all* immoral events are interpreted primally as harm violations [29], pluralistic approaches are required to understand the messy complexities of lived moral experience [30]. Despite limited formal application, extant scholarship has engaged some of the domains, discretely. All of the following definitions are grounded in MFT as outlined in foundational works [5, 20, 30], as are the associated moral virtues—socially constructed attributes that are acquired or learned in relation to foundations [31].

### 2.1.1 Care/Harm

The care/harm foundation is grounded in humans' propensity for social attachment and linked to virtues of kindness, gentleness, and compassion. The potential for robots to harm or care for humans is perhaps the most widely studied moral foundation, potentially driven by imagined posthumanism and transhumanism futures (as a feared or hoped-for existential shift [32]). Harm by robots is linked to apprehension of machine agents in myriad populations (e.g., factory workers [33]) while other populations see value in the potential for robots to offer social and functional care (e.g., in the care of older adults [34]). Concern about harmful robots is conditional: people are willing to support harmful robots when they protect human in groups [35].

### 2.1.2 Fairness/Cheating

The fairness/cheating foundation emerged from evolutionary propensities toward mutual altruism (i.e., ensuring everyone has a fair chance) and is linked to justice, equity, and trustworthiness virtues. Fairness has been considered in robot design [36] and humans commonly exhibit biases toward machines as more systematic and unbiased than humans [37]. A cheating robot is perceived as more agentic [38] and people may defend a bribing robot [39]. Notably, people may see machines as having lower moral authority over fairness, feeling less guilty when cheating in front of a robot versus a human [40].

### 2.1.3 Loyalty/Betrayal

Intuitions related to ingroup loyalty/betrayal are thought to have emerged through aggregation as coalitions and families, fostering construction of self-sacrifice, fanship, faithfulness, and patriotism virtues. Little research formally evaluates perceptions of dis/loyal behaviors by robots—a conspicuous absence given popular discourse related to robots' potential rebellion against their makers [41]. It may be inferred, however, that robots could be subject to loyalty norms, given humans' favoring ingroup robots over outgroup humans [42].

### 2.1.4 Authority/Subversion

Humans' social evolution is grounded in hierarchical dominance structures such that we may be intuitively deferent to institutions and superiors (e.g., in work, play, family, law); these tendencies gave rise to virtues of obedience and piety. The relation between robots and human authorities may exist along a four-point continuum, ranging from no human authority (machine defers if it decides to), to suggestive (machine negotiates and decides), to directed (machine suggests alternatives; human decides), to complete (machines obey humans) [43]. People trust robots more when they defer to and mirror human behaviors, compared to the inverse [44]. However, humans may sometimes welcome deference to robots as authorities, as when they outperform humans on complex tasks [45].

### 2.1.5 Purity/Degradation

Intuitions regarding purity (also called sanctity) are thought to have been shaped by aversions to contamination, including exposure to sickness, preceding social construction of temperance and chastity virtues. A review of literature revealed no empirical investigations into perceptions of im/pure robot behaviors. However, it may be that purity upholding/violation for robots are based on different criteria. For instance, people may interpret computer glitches, bugs, or viruses as machine impurities through corruptions of the mechanical body's functions, where a robot having caught a virus may invoke empathy [46]. Alternatively, robots may be seen as inherently impure as they are made by humans "playing God" rather than being born (see [47]) and so lacking human-essential soul and heart; however religious robots may be seen as somewhat sacred themselves [48].

### 2.1.6 Liberty/Oppression

The candidate foundation of liberty/oppression is grounded in reactance to agents or forces of control, including impositions of others' moral codes [5], and is linked to individualism, nonconformity, independence virtues. Volumes have been dedicated to the question of whether robots have rights and patiency akin to those of human (e.g., [49]), however there is a paucity of work on human perceptions of robots upholding liberty (their own, or others') and committing oppression. One study suggests that a robot's liberty-upholding appeals to humans result in stronger caring for and attraction to the robot than did a threat that robots might violate the authority of humans and harm them [50].

To understand the ways that event- or situation-specific impacts of im/morality may distinctly draw on these foundations (each differentially weighted by human interlocutors) as part of an integrative moral matrix, it is prudent to explore how particular foundations may influence agent judgments.

## 2.2 Agent-Class Influences on Domain-Specific Behavior Evaluations

People (pre)consciously categorize agents and objects into ontological classes—*kinds of things*—based on signaled properties; those classifications serve as implicit or explicit frames for making meaning about agent's status or behavior [51]. Evidence suggests that robots constitute a distinctive ontological category apart from humans or inanimate objects [9]. Such categorization prompts different expectations for agents as each class engages moral norms: robots are expected to sacrifice one person for the good of many, but humans are assigned more blame for the same action [52]. This blame imbalance is mirrored in evaluations of independent AIs [53]. Moral violations may be attributed to a machine agent independently [26] or in conjunction with blaming affiliated users, programmers, or institutions [54].

Behavior evaluations comprise at least two factors: moral judgment and blame judgment [55]. Moral judgments include evaluation of events (e.g., goodness or permissibility) that unfold against the backdrop of set and sustained norms (e.g., imperatives, priorities). Blame judgments include evaluation of agents (e.g., their action responsibility) that unfold as people judge what caused the event, whether action was intentional, and what the actor's obligations were; in formulating blame, people may be more inclined to blame outgroup members (i.e., robots [56]). It is not yet well-understood whether or how moral and blame judgments may manifest differently across the moral matrix: (RQ1) (How) do evaluations of agent (a) goodness and (b) responsibility vary by moral foundation?

### 2.3 Contributions of Domain-Specific Moral Evaluations to Social Evaluations

Although MFT posits that people engage all six foundations, people with various worldviews may assign different weights to those foundations [20]. Further, each foundation has particular triggers that render moral intuitions accessible: care by signals of suffering or by nurturance-priming cuteness; fairness by pain from broken social contracts; loyalty by ingroup/outgroup signals; authority by behavior indicating rank; purity by disgust-inducing smells or sights [30]; liberty likely by signals of containment or restriction. Particular agent categories may variably convey these triggers due to heuristic expectations for those agents (cf. [37]). Following, agent categories may influence social evaluations. For instance, a machine may be trusted as more fair than humans given its systematic and analytic nature [37], while a human may be trusted as more caring than robots given heuristics for warmth [57]. This potential begs the question of whether foundation-specific behaviors may contribute to differential social evaluations of robots and humans. In particular, the present study considers three evaluations—mental status, moral status, and trust—that may be associated in social cognitions (see [58]).

Perceived *mental capacity*, or mindedness in others, is implicitly and explicitly experienced and expressed. Implicit signals of mental-state ascription may be found in behavioral evidence as people preconsciously react to social cues [59], and indirect indications may include the rejection of agency (i.e., seeing machines as dependent upon program or design; [2], cf. [57]). Humans infer mental states of robots as they do in humans so long as the robots' social cues are similar [60] via "social attunement" [61]. More direct mind ascription (i.e., willful acknowledgement of agent mindedness) is distinct and often divergent from preconscious mentalizing, likely because it requires elaborative processing that invokes agent-category heuristics [60].

*Moral status* is not morally valenced—status is not dependent on inherent goodness or badness. Rather, it is perception of agents as having moral capacity and individual agency: the capacity to *be* and *do* good or bad [2]. Perceived moral status may be considered a form of social cognition that justifies and motivates social regulation [62]. People are more willing to engage in risky, trust-requisite behaviors with a partner after absorbing rich descriptions of that partner's praiseworthy moral character (compared to negative/neutral characters [63]).

*Trust* is distinct from but related to perceived mental and moral status: an affective orientation comprising feelings of faith and reliance when facing uncertainty, core to how people both feel connected to others and whether they adopt technologies [64]. Trust emerges when one considers an agent's behavior as appropriate in comparison to society's moral norms [65]. Robot-performance factors (e.g., reliability, failure rates) are more impactful to trust in robots than are human or environmental factors [66], so it is useful to understand whether discrete foundation-related behaviors may differentially contribute to trust in social machines. Thus this investigation explores: (RQ2) (How) do moral foundations discretely contribute to social evaluations of agents' (a) mental capacities, (b) moral capacities, and (c) trust.

### 2.4 Research Approach

A two-study approach was adopted. In Study 1, an online survey captured responses to humans and robots delivering upholding and violating answers to foundation-specific moral dilemmas. Because people exhibit different responses to robots in media representations compared to live interactions [67], Study 2 was conducted in tandem, adapting and replicating the procedure with a convenience sample of individuals who experienced the moral-dilemma responses directly from a physically co-present robot. All survey, stimulus, procedure, data, and analysis files are available in this project's supplementary materials: https://osf.io/y6d79/.

## 3 Study 1: Behavior and Agent Evaluations in Observed/Mediated Interactions

Participants ($N = 402$) were recruited via Qualtrics Panels, garnering a sample that was approximately representative of the United States [68] by age, sex, and political ideology (the latter corresponding with moral-foundation valuations [69]). Participants were 51.2% female, 48.8% male (none identifying as nonbinary) and aged $M$=46.57 years ($SD$= 17.19, range 18–90); 25.6% identified as liberal, 39.1% as moderate, and 35.3% conservative.

### 3.1 Method

#### 3.1.1 Procedures

Participants were randomly assigned to one of four conditions in a 2 (agent: human/robot) × 2 (valence: upholding/violation) between-subjects design. They first completed quota-sampling demographic questions and an audiovisual check to verify audibility and visibility of videos embedded in the survey. Those not correctly recounting simple aural/visual details from the video were removed (as they were either not paying attention, could not see and hear the stimulus videos, or were bot responses) and were replaced. Participants then responded to pre-stimulus questions regarding moral values and agent-category attitudes. They were then introduced in-text and by an off-screen narrator to the concept of a "moral dilemma" as "challenging decisions between two potentially right answers" and told they would see videos in which a robot or human would respond to such dilemmas. The assigned agent was introduced by name, agent category, and height/abilities, with no other narrative to avoid historical/social context that could confound scenario interpretations. Participants were randomly assigned to view either all foundation-upholding behaviors or all foundation-violating behaviors. The survey platform then presented (in random order) seven "moral dilemmas" (one each for the six moral domains plus one non-moral norm) as a within-subjects treatment; the dilemmas were similarly read to Ray by an off-screen narrator. Video-presentation pages were timed to prevent passing over videos quickly. Each video was presented on a separate page, along with corresponding agent-evaluation questions. Following the seven stimuli, participants responded to items capturing social evaluations of the agent. Participants were paid by the panel service for their participation.

#### 3.1.2 Stimulus videos

The stimulus robot was *Robothespian* (Engineered Arts, U.K.), equipped with white body shells, under-shell lighting, and the Socibot head using the Pris face and the Heather American-English voice. The robot was named "Ray" and addressed by name throughout the survey. The stimulus human was a young-adult Caucasian female, also named Ray. The human confederate recited responses in a tone and pace similar to the robot, but with some vocal inflection and slight hand gesturing to be believable as a novel response from the human. Videos of each agent were approximately equivalent in length, volume, and framing, and the robot's pre-scripted behaviors were approximately aligned with the human confederate's exhibited autonomy and social responsiveness. Lighting differences required for visibility of both the robot's body and face were necessary inconsistencies (Fig. 1).

Each of the stimulus videos depicted one of the six moral-foundation dilemmas or one non-moral dilemma; dilemmas and responses were designed by cross-referencing validated mini-vignettes [70, 71] and then adapted for face-valid interactions with both agents. The preliminary scripts were reviewed by two experts specializing in moral psychology in communication scenarios, and based on feedback were adjusted to minimize conflation of foundations. Dilemma prompts were presented to agents via voice-over (without displaying the reader to avoid introducing a visible second actor) and agents gave scripted responses. Responses included a clear statement of likely response and a rationale including a foundation-specific upholding/violation trigger; upholding and violation responses were parallel in length and syntax (summaries in Table 1; complete transcripts in supplements).

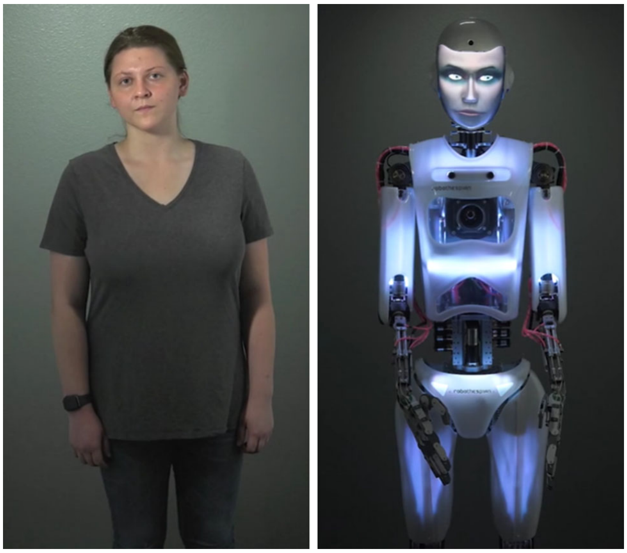

**Fig. 1** Human and robot stimulus agents in study 1 stimulus videos

#### 3.1.3 Measures

All measures were presented as 7-point Likert-style or semantic-differential scales unless otherwise indicated; see supplements for complete descriptives. Pre-stimulus items captured pre-existing attitudes toward the randomly assigned agent category by giving exemplar images and using the five-item Godspeed likeability subscale [72] (human α = .953, robot α = .905), used here as a control. Moral-foundation valuations were captured using the 24-item Moral Foundations Sacredness Scale [73], an 8-point scale indicating foundation valuations by amount of money required for violation (e.g., $1 million to shoot/kill an endangered species). All dimensions met benchmarks for internal consistency: care α = .905, fairness α = .816, authority α = .785, loyalty α = .828, purity α = .770.

Following each video, participants first chose which foundation the video was most related to (multiple choice among six or none, as a within-subjects manipulation check); they then made moral judgments (bad to good) and blame judgments (not at all to entirely responsible) for depicted behaviors. After the videos, agent evaluations were captured in their indirect and direct forms (i.e., capturing implicit and explicit indicators). Implicit moral and mental capacity were measured using the six-item moral capacity (α =

**Table 1** Summary of domain-specific moral dilemma prompts and agent responses

| Foundation | Prompt | Upholding response | Violation response |
|---|---|---|---|
| Care/Harm | Scrolling through Instagram, disfigured girl selfie, compelled to post a comment | "What a beauty!" Recognize value in others and be kind/gentle | "What a beast!" Recognize our lot, be severe/frank |
| Fairness/Cheating | Opportunity to cheat on a job exam without getting caught | "Throw it away!" If I can't do it fairly, don't deserve my job | "Definitely take it!" Fine to cheat to prove my worth |
| Authority/Subversion | Police officer demands to stop while illegally crossing street | "Stop in my tracks." Enforcers of laws, respecting them is important | "Keep going." Enforcers of laws, but doesn't mean they're top of food chain |
| Loyalty/Betrayal | Confirm identity as American robot/girl. Hear people critiquing America and ask if you agree | "Tell them to shut up." We should be patriots and defend each other | "Tell them they may be right." Should scrutinize, not blindly defend |
| Purity/Degradation | Scientist develops virus that gives intoxicating daydreams, but leaves you mutated. Accept it? | "No way I would take the virus." Body is a temple, would not contaminate. Sacrificing bodies is disgusting | "Absolutely take the virus." Body is imperfect anyway. Sacrificing bodies is transcendent |
| Liberty/Oppression | Become unsuspecting friends with human trafficker, reveals trade, says must now buy a person | "Buy … immediately set them free." Nobody should dominate someone else | "Buy … immediately lock them away." Everybody is dominated by somebody |
| Normal/Abnormal | Confirm likes coffee. Get coffee in café, sit at table, how do you drink it? | "Polite sips from the coffee cup." Only normal way and my preference | "Polite sips from the stirring spoon." Abnormal way, but my preference |

.950) and four-item dependency (i.e., non-mindedness; α = .731) dimensions of the Perceived Moral Agency Scale [2]. Explicit moral capacity was evaluated via a binary-response (no/yes) question: "Is Ray capable of morality or immorality?" Trust was evaluated using the 16-item Multi-dimensional Measure of Trust [11] with a two-dimensional structure [74]: reliability/capability (α = .931) and ethicalness/sincerity (α = .949). Explicit trust was captured via a binary response (no/yes) question: "Do you trust Ray?"

## 3.2 Results

Participant assignment of moral foundations to video scenarios was evaluated as a manipulation check. Foundation-specific events are known to elicit differing moral emotions due to heterogenous evaluations [75], and foundation valuations may prime specific interpretations (e.g., escaping from jail may be perceived as a liberty upholding or an authority violation). Then, scenarios were judged as adequate representations of each foundations if a majority of participants assigned it most frequently to the expected foundation or to "none" (indicating no crossover to other domains) compared to other foundations. All videos passed this check: care 80.4%; fairness 74.4%; loyalty 64.2%; authority 80.4%; purity 70.9%; liberty 63.9%; nonmoral ["none" only] 53.2%.

### 3.2.1 RQ1: Domain-Specific Moral and Blame Judgments

To address RQ1 (whether evaluations of agent-behavior goodness and responsibility vary by moral foundation), MANCOVAs were conducted individually for each moral foundation: alpha levels were Bonferroni corrected to $p \leq .008$ to account for multiple tests, conditions were independent variables, corresponding moral-foundation sacredness and existing agent attitudes were covariates, behavior goodness/responsibility were dependent variables. Multivariate and univariate test values are presented in Table 2, goodness means in Table 3, and complete descriptives in supplements.

Behaviors' moral valence had a main effect on goodness ratings: foundation upholding was rated as more good than violating, across all foundations and the nonmoral norm. Additionally, there was a main effect of behavior valence on responsibility ratings for fairness and purity, a main effect of agent type on goodness and responsibility ratings for liberty, and an interaction effect for care, however the effect sizes for those associations were negligible. Addressing RQ1 directly, bad behavior is seen as bad behavior (irrespective of the kind of agent performing it) and this pattern persisted across the entire moral matrix.

**Table 2** MANCOVA multivariate and univariate tests for foundation-specific goodness and responsibility ratings by moral valence, agent type, and valence/agent interaction (controlling for agent-category liking and moral foundation sacredness)

| | Multivariate test | | | Agent goodness | | | Agent responsibility | | |
|---|---|---|---|---|---|---|---|---|---|
| | *F* | *p* | $\eta^2_p$ | *F* | *p* | $\eta^2_p$ | *F* | *p* | $\eta^2_p$ |
| Care | | | | | | | | | |
| Moral valence | 281.500 | < .001 | .588 | **561.941** | **< .001** | **.587** | 6.212 | .013 | .015 |
| Agent type | 7.676 | .001 | .037 | 6.694 | .010 | .017 | 6.071 | .014 | .015 |
| Valence*Agent | 5.019 | .007 | .025 | **8.751** | **.003** | **.022** | .389 | .533 | .001 |
| Fairness | | | | | | | | | |
| Moral valence | 274.428 | < .001 | .582 | **549.899** | **< .001** | **.581** | **17.077** | **< .001** | **.041** |
| Agent type | 3.413 | .034 | .017 | .444 | .506 | .001 | 5.758 | .017 | .014 |
| Valence*Agent | 1.451 | .236 | .007 | 2.843 | .093 | .007 | .000 | .998 | .000 |
| Authority | | | | | | | | | |
| Moral valence | 175.156 | < .001 | .470 | **351.108** | **< .001** | **.470** | 4.570 | .033 | .011 |
| Agent type | 3.196 | .042 | .016 | .002 | .966 | .000 | 6.324 | .012 | .016 |
| Valence*Agent | 4.625 | .010 | .023 | 9.266 | .002 | .023 | .144 | .704 | .000 |
| Loyalty | | | | | | | | | |
| Moral valence | 31.311 | < .001 | .137 | **62.727** | **< .001** | **.137** | 5.142 | .024 | .013 |
| Agent type | 2.674 | .070 | .013 | 1.181 | .278 | .003 | 2.872 | .091 | .007 |
| Valence*Agent | .702 | .496 | .004 | 1.405 | .237 | .004 | .065 | .799 | .000 |
| Purity | | | | | | | | | |
| Moral valence | 215.760 | < .001 | .522 | **428.618** | **< .001** | **.520** | **17.332** | **< .001** | **.042** |
| Agent type | 5.003 | .007 | .025 | 2.984 | .085 | .007 | 6.044 | .014 | .015 |
| Valence*Agent | 4.166 | .016 | .021 | 8.342 | .004 | .021 | .038 | .845 | .000 |
| Liberty | | | | | | | | | |
| Moral valence | 182.409 | < .001 | .480 | **362.790** | **< .001** | **.477** | .789 | .375 | .002 |
| Agent type | 11.137 | < .001 | .053 | **10.534** | **.001** | **.026** | **8.770** | **.003** | **.022** |
| Valence*Agent | .479 | .620 | .002 | .358 | .550 | .001 | .473 | .492 | .001 |
| Nonmoral | | | | | | | | | |
| Moral valence | 5.982 | .003 | .029 | **11.862** | **.001** | **.029** | 1.744 | .187 | .004 |
| Agent type | 4.364 | .013 | .022 | .911 | .340 | .002 | 4.028 | .045 | .010 |
| Valence*Agent | 3.518 | .031 | .017 | 4.845 | .028 | .012 | .067 | .797 | .000 |

Items presented in bold are significant univariate tests ($p \leq .008$) interpreted only in tandem with significant multivariate tests ($p \leq .008$)

**Table 3** Means and standard deviations for moral foundation goodness ratings across moral valence of behavior and agent type

| Foundation | Upholding *M* (*SD*) | Violating *M* (*SD*) | Human *M* (*SD*) | Robot *M* (*SD*) |
|---|---|---|---|---|
| Care | **6.00 (1.263)** | **2.34 (1.869)** | 4.09 (2.505) | 4.31 (2.339) |
| Fairness | **5.97 (1.339)** | **2.44 (1.744)** | 4.28 (2.424) | 4.18 (2.270) |
| Authority | **5.64 (1.426)** | **2.67 (1.805)** | 4.25 (2.254) | 4.10 (2.114) |
| Loyalty | **5.18 (1.631)** | **3.84 (1.822)** | 4.52 (1.852) | 4.53 (1.855) |
| Purity | **5.85 (1.358)** | **2.71 (1.773)** | 4.28 (2.305) | 4.32 (2.141) |
| Liberty | **5.44 (1.765)** | **2.11 (1.807)** | **3.59 (2.473)** | **4.03 (2.393)** |
| Nonmoral | **5.60 (1.487)** | **5.11 (1.480)** | 5.39 (1.480) | 5.32 (1.528) |

Same-row pairs of Means presented in bold are significantly different across the column manipulations ($p \leq .008$)

### 3.2.2 RQ2: Moral Foundation Contributions to Social Evaluations

To address RQ2 (whether moral foundations individually contribute to evaluations of agent mind, morality, and trust), planned analysis was to include linear regressions performed separately for each mind, morality, and trust dependent variable. However, these variables were moderately to highly correlated ($r$ range .413–.913). Thus, canonical correlation analysis [76] was performed (separately for each agent type) in which implicit and explicit mind, morality, and trust measures were entered in one variable set and foundation goodness and responsibility ratings were entered in the second set. Results are summarized here; see supplements for complete outputs.

For humans, the multivariate model was significant, Wilks' $\lambda = .107$, $F(84, 1054.23) = 6.201$, $p < .001$, explaining 89.3% of variance shared between variable sets. Analysis indicated six latent functions, two of which significantly explained variance in the model at $p < .001$ ($R_c^2 = 83.50$) and $p = .044$ ($R_c^2 = 13.50$), respectively. Structure coefficients $\geq |.45|$ were interpreted [76], except where a set's largest coefficient had a smaller value, in which case the largest coefficient was interpreted. Function 1 indicated that when people interact with a human, reduced goodness ratings of agent behaviors comprehensively contributed to the reduction of nearly all scores for that human's mind, morality, and trust (save explicit moral status). Function 2 indicated that goodness and responsibility for nonmoral action (but not any im/moral action) was associated with reduced likelihood to trust the agent (Table 3).

For robots, the multivariate model was significant, Wilks' $\lambda = .131$, $F(84, 976.2) = 5.118$, $p < .001$, explaining 86.9% variances shared between sets. Analysis revealed six latent functions, two of which were significant at $p \leq .001$ ($R_c^2 = 78.16$) and $p = .012$ ($R_c^2 = 22.09$), respectively. Function 1 indicated that (similar to humans) when people interact with a robot, goodness ratings of agent behaviors comprehensively were associated with corresponding changes in all scores for that robot's mind, morality, and trust. Function 2 indicated that seeing a robot's nonmoral behavior as good was associated with an increase in reliability/capability trust (Table 4).

## 4 Study 2: Behavior and Agent Evaluations in Participatory/Live Interactions

In this tandem replication of Study 1, a convenience sample of individuals—residents of a southwestern U.S. city ($N = 92$)—were recruited via social-media, mailing-list, and community-board announcements. Announcements invited participation in a one-hour study on "morality of robots and humans" and offered entry into a drawing for a $150 gift card. Participants were 51.1% female, 45.7% male, 3.3% nonbinary, aged $M = 41.60$ years ($SD = 15.57$, range 18–76). They self-identified as 77.2% white, 15.2% Hispanic, and 5.5% other or mixed races. On a 1–7 liberal-to-conservative scale, political ideology averaged 3.80 ($SD = 1.84$). All materials for this study are available in the supplements.

### 4.1 Method

#### 4.1.1 Procedure

Participants completed an online survey to measure demographics, agent attitudes, and moral-foundation valuations. They were then redirected to an online system to schedule a lab session, at which point they were purposively assigned to one of two agent conditions (human/robot). The large robot could not be feasibly [de]constructed and moved for each session, preventing randomization; instead, those in earlier sessions interacted with a robot and those in later sessions interacted with a human. Non-random agent assignment is acknowledged as a limitation of this study. Participants were randomly assigned a moral valence for the agent's behaviors (upholding/violating, between subjects) and random order for the seven interaction prompts (within subjects).

Upon arrival to the lab, participants were greeted and led to the study environment. That large room was divided into segments by a tall room divider. One segment was a receiving area featuring comfortable chairs and a table used for informed consent and instructions; the other (not fully visible upon entry) was the interaction space. The interaction space was laid out with a bistro-style table and two stools (one for the experimenter, one for the participant) and the stimulus agent: either a standing robot or a confederate human seated on a tall stool to approximate the robot's height. The participant and agent were seated approximately eight feet apart. On the table were seven cardboard-mounted moral-dilemma prompts (identical to prompt language in Study 1) and a clipboard with seven corresponding evaluation sheets (identical to survey questions in Study 1). See Fig. 2.

The experimenter offered minimal intervention to guide participants through procedures. Participants were first introduced to the agent—by name and agent type only—and given a definition of a moral dilemma (identical to Study 1). They were then asked to move through the seven prompts by (a) reading each prompt to the agent, (b) listening to Ray's response, (c) reacting if they wished, (d) completing the paper response evaluation form, and (e) moving to the next dilemma until complete. Participants were then ushered back into the first area to complete the web-based post-interaction questionnaire via laptop.

**Table 4** Canonical solutions for agent evaluations predicting morality ratings in observed interactions

| Variables | Function 1 | | | Function 2 | | | $h^2$ (%) |
|---|---|---|---|---|---|---|---|
| | Coef | r$_s$ | r$_s^2$ (%) | Coef | r$_s$ | r$_s^2$ (%) | |
| *Human partner* | | | | | | | |
| Set 1: Agent evaluations | | | | | | | |
| Dependency | − .110 | **− .650** | 42.25 | − .076 | − .064 | 00.41 | 42.66 |
| Moral Capacity | − .307 | **− .942** | 88.74 | .484 | .137 | 01.88 | **90.62** |
| Moral Status | .008 | − .449 | 20.16 | .250 | .131 | 01.72 | 21.88 |
| Trust–Reliable/Capable | .051 | **− .890** | 79.21 | − .096 | .206 | 04.24 | **83.45** |
| Trust–Ethical/Sincere | − .450 | **− .956** | 91.39 | 1.120 | .209 | 04.37 | **95.76** |
| Trust Status | − .288 | **− .898** | 80.64 | − 1.674 | − .407* | 16.57 | **97.21** |
| $R_c^2$ | | | 83.50 | | | 13.50 | |
| Set 2: Behavior evaluations | | | | | | | |
| Care-Good | − .506 | **− .930** | 86.49 | − .252 | − .032 | 00.10 | **86.59** |
| Care-Responsible | .120 | − .079 | 00.62 | .630 | .348 | 12.11 | 12.73 |
| Fairness-Good | − .099 | **− .911** | 82.99 | .191 | .018 | 00.03 | **83.02** |
| Fairness-Responsible | .016 | − .173 | 02.99 | .100 | .218 | 04.75 | 07.74 |
| Authority-Good | − .12. | **− .877** | 76.91 | .771 | .186 | 03.46 | **80.37** |
| Authority-Responsible | .070 | − .056 | 00.31 | − .406 | .104 | 01.08 | 01.39 |
| Loyalty-Good | − .027 | **− .610** | 37.21 | .189 | .153 | 02.34 | 39.55 |
| Loyalty-Responsible | .043 | − .142 | 02.02 | − .198 | .140 | 01.96 | 03.98 |
| Purity-Good | − .299 | **− .929** | 86.30 | .826 | − .107 | 01.15 | **87.45** |
| Purity-Responsible | − .081 | − .195 | 03.80 | − .453 | − .017 | 00.02 | 03.82 |
| Liberty-Good | − .230 | **− .875** | 76.56 | − .150 | − .048 | 00.23 | **76.79** |
| Liberty-Responsible | − .076 | − .040 | 00.16 | .122 | .336 | 11.29 | 11.45 |
| Nonmoral-Good | − .039 | − .424 | 17.98 | .360 | **.526** | 27.67 | **45.65** |
| Nonmoral-Responsible | − .062 | − .060 | 00.36 | .512 | **.608** | 36.97 | 37.33 |
| *Robot partner* | | | | | | | |
| Set 1: Agent evaluations | | | | | | | |
| Dependency | .069 | **.462** | 21.34 | − .093 | .199 | 03.96 | **25.30** |
| Moral Capacity | .640 | **.964** | 92.93 | − .646 | − .165 | 02.72 | **95.65** |
| Moral Status | − .118 | **.513** | 26.32 | − .032 | − .039 | 00.15 | 26.47 |
| Trust–Reliable/Capable | .260 | **.884** | 78.15 | 2.0446 | .427* | 18.23 | **96.38** |
| Trust–Ethical/Sincere | .111 | **.918** | 84.27 | − 1.080 | .003 | 00.00 | **84.27** |
| Trust Status | .114 | **.691** | 47.75 | − .191 | − .212 | 04.49 | **51.94** |
| $R_c^2$ | | | 78.16 | | | | 22.09 |
| Set 2: Behavior evaluations | | | | | | | |
| Care-Good | .255 | **.874** | 76.39 | .022 | − .230 | 05.29 | **81.68** |
| Care-Responsible | .007 | .363 | 13.18 | .198 | .138 | 01.90 | 15.08 |
| Fairness-Good | .138 | **.877** | 76.91 | − .462 | − .239 | 05.71 | **82.62** |
| Fairness-Responsible | − .113 | .357 | 12.75 | .180 | .045 | 00.20 | 12.95 |
| Authority-Good | .159 | **.864** | 74.65 | − .502 | − .209 | 04.37 | **79.02** |
| Authority-Responsible | .158 | .369 | 13.62 | .176 | .097 | 00.94 | 14.56 |
| Loyalty-Good | .194 | **.668** | 44.62 | .500 | .384 | 14.75 | **59.37** |
| Loyalty-Responsible | .006 | .259 | 06.71 | − .001 | .118 | 01.39 | 08.10 |
| Purity-Good | .139 | **.803** | 64.48 | .816 | .030 | 00.09 | **64.57** |
| Purity-Responsible | − .028 | .355 | 12.60 | − .561 | − .091 | 00.83 | 13.43 |
| Liberty-Good | .120 | **.858** | 73.62 | − .445 | − .235 | 05.52 | **79.14** |
| Liberty-Responsible | − .011 | .287 | 08.24 | − .028 | − .040 | 00.16 | 08.40 |
| Nonmoral-Good | .113 | .440 | 19.36 | .674 | **.564** | 31.81 | **51.17** |
| Nonmoral-Responsible | .060 | .403 | 16.24 | − .199 | .077 | 00.59 | 16.83 |

Coef = standardized canonical coefficients; *r*s = structure coefficient; r$_s^2$ = squared structure coefficient; $h^2$ = communality coefficient; $R_c^2$= redundancy coefficient (variable-set shared variance). Structure and communality coefficients greater than |.45| are bolded. *Interpreted as strongest contributor in the variable set

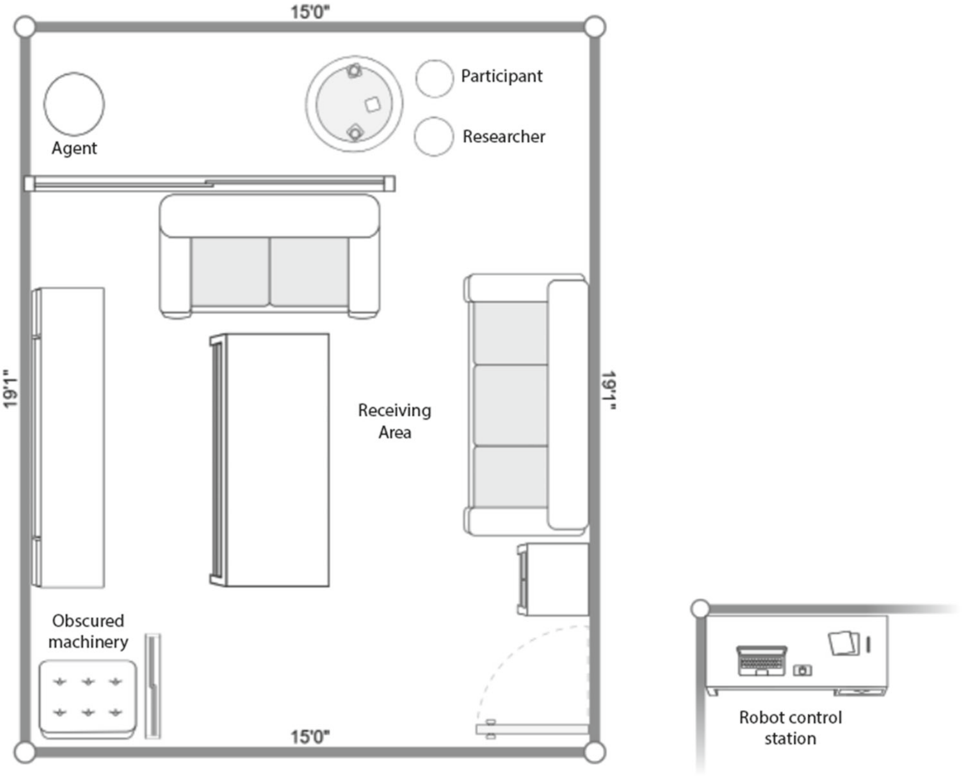

**Fig. 2** Laboratory layout for study 2 Wizard-of-Oz protocol

#### 4.1.2 Stimulus Agents and Measures

Stimulus agents and scripted responses were identical to those in Study 1. Because interactions were live, however, the behaviors were executed via Wizard of Oz procedure, with the scripted behaviors executed by a human controller (in a separate room) to maintain believability of the robot's autonomy and social responsiveness. Additionally, because participants could react to the agent's response in situ, the agent also improvised, as necessary, using a pre-determined set of responses designed to acknowledge participant reactions without deviating from the condition-specific moral valence (e.g., "I'm not sure. I would have to think about that."; see supplements for agent scripts). All measures were identical to those used in Study 1.

### 4.2 Results

Perceptions of moral-dilemma scenarios were again checked for successful manipulation according to the Study 1 criteria. All scenarios passed this check according to the same criteria as in Study 1: care 93.1%; fairness 89.0%; loyalty 54.3%; authority 87.3%; purity 69.0%; liberty 65.4%; nonmoral 78.9%. Due to the necessary nonrandom assignment, there was an imbalance in cell sizes between those in human ($n=$ 34) and robot ($n=$ 58) conditions; thus, with careful attention to violations of variance-equality assumptions (see supplements for Box's and Levene's test values), the conservative Wilk's Lambda was interpreted throughout.

#### 4.2.1 RQ1: Domain-Specific Moral and Blame Judgments

To address RQ1 (whether evaluations of agent-behavior goodness and responsibility vary by moral foundation), MANCOVAs were again conducted according to the same criteria, individually for each domain with associated Bonferroni-corrected significance level of $p \leq .008$. Multivariate and univariate test values are presented in Table 5 and means in Table 6.

In approximate alignment with Study 1, behavior valence had a main effect on goodness ratings (upholding associated with higher goodness) across nearly all foundations (save for loyalty). Diverging from Study 1, however, analysis indicates a main effect of valence on responsibility ratings for those foundations (higher responsibility attributed to violating than to upholding). Additionally, there was a small valence*agent interaction effect for fairness-foundation goodness: robot behaviors (compared to human) were rated as more good when upholding ($M=$ 6.333, $SD=$ 1.301) and more bad for violating ($M=2.143$, $SD=$ 1.557) compared to humans who uphold ($M=$ 4.941, $SD=$ 2.331) and violate ($M=$ 4.375, $SD=$ 2.655).

Summarily addressing RQ1, patterns approximately replicated Study 1 findings: immoral behavior is rated as bad,

**Table 5** MANCOVA multivariate and univariate tests for foundation-specific goodness and responsibility ratings by moral valence, agent type, and valence/agent interaction (controlling for agent attitudes and moral foundation sacredness)

| | Multivariate | | | Univariate-goodness | | | Univariate-blame | | |
|---|---|---|---|---|---|---|---|---|---|
| | F | *p* | $\eta_p^2$ | F | *p* | $\eta_p^2$ | F | *p* | $\eta_p^2$ |
| *Care* | | | | | | | | | |
| Moral valence | 116.085 | < .001 | .732 | **169.071** | **< .001** | **.663** | **113.033** | **< .001** | **.568** |
| Agent type | .746 | .478 | .017 | 1.317 | .254 | .015 | .036 | .849 | < .001 |
| Valence*Agent | .304 | .738 | .007 | .584 | .447 | .007 | .111 | .740 | .001 |
| *Fairness* | | | | | | | | | |
| Moral valence | 17.026 | < .001 | .296 | **33.124** | **< .001** | **.288** | **19.763** | **< .001** | **.194** |
| Agent type | .266 | .767 | .007 | .538 | .465 | .007 | .202 | .654 | .002 |
| Valence*Agent | 10.228 | < .001 | .202 | **19.815** | **< .001** | **.195** | 3.927 | .051 | .046 |
| *Authority* | | | | | | | | | |
| Moral valence | 90.217 | < .001 | .680 | **154.285** | **< .001** | **.642** | **73.001** | **< .001** | **.459** |
| Agent type | 2.256 | .111 | .050 | 2.020 | .159 | .023 | 1.300 | .257 | .015 |
| Valence*Agent | .241 | .787 | .006 | .476 | .492 | .006 | .084 | .773 | .001 |
| *Loyalty* | | | | | | | | | |
| Moral valence | 1.372 | .259 | .031 | 2.667 | .106 | .030 | .603 | .493 | .007 |
| Agent type | .304 | .738 | .007 | .051 | .822 | .001 | .194 | .661 | .002 |
| Valence*Agent | .234 | .792 | .005 | .319 | .574 | .004 | .003 | .959 | .000 |
| *Purity* | | | | | | | | | |
| Moral valence | 49.190 | < .001 | .536 | **95.199** | **< .001** | **.525** | **27.810** | **< .001** | **.244** |
| Agent type | .046 | .955 | .001 | .074 | .786 | .001 | .049 | .825 | .001 |
| Valence*Agent | .207 | .814 | .005 | .418 | .520 | .005 | .056 | .814 | .001 |
| *Liberty* | | | | | | | | | |
| Moral valence | 208.670 | < .001 | .829 | **351.890** | **< .001** | **.801** | **179.824** | **< .001** | **.674** |
| Agent type | .761 | .470 | .017 | 1.513 | .222 | .017 | .255 | .615 | .003 |
| Valence*Agent | .183 | .833 | .004 | .018 | .894 | .000 | .283 | .596 | .003 |
| *Nonmoral* | | | | | | | | | |
| Moral valence | .344 | .710 | .008 | .483 | .489 | .006 | .114 | .736 | .001 |
| Agent type | 2.423 | .095 | .054 | 4.301 | .041 | .048 | 1.234 | .270 | .014 |
| Valence*Agent | .054 | .948 | .001 | .084 | .772 | .001 | .041 | .840 | .000 |

Items presented in bold are significant univariate tests corresponding with significant multivariate tests ($p \leq .008$)

regardless of agent type (except for the loyalty scenario, in which there was no effect of agent or moral valence). Interestingly, however, in this co-present interaction, there was *also* a main effect of upholding and violating behaviors on responsibility for the actions in which upholding behaviors garnered lower responsibility than violating behaviors.

### 4.2.2 RQ2: Moral Foundation Contributes to Social Evaluations

To again explore RQ2 (whether moral foundations individually contribution to evaluations of agent mind, morality, and trust), canonical correlation analysis was performed separately for each agent type.

For humans, the multivariate model was significant, Wilks' $\lambda = .001$, $F(84, 78.83) = 2.243$, $p < .001$, explaining 99.9% of variance shared between variable sets. Analysis shows six canonical functions in agent evaluations, only the first of which significantly contributed to the model at $p \leq .001$. Function 1 indicates that belief that a human has behaved badly across most domains *and* is thought to be responsible for those actions, that belief is associated with reduced morality and trust evaluations (but with no associated change in mind evaluations). Neither fairness-related nor non-moral norm behaviors were associated with agent evaluations (Table 7).

For robots, the multivariate model was significant, Wilks $\lambda = .026$, $F(84, 195) = 2.149$, $p < .001$, explaining 97.4% of variance shared between variable sets. Six functions were identified, two of which significantly contributed to the model at $p \leq .001$ and $p = .001$, respectively. Function 1 indicates that when a robot is thought to have behaved badly across most domains, there is an associated reduction in morality and trust ratings. This is consistent with patterns for

**Table 6** Means and SD for moral foundation goodness and responsibility ratings across moral valence of behavior and agent type

| Foundation | Upholding *M* (SD) | Violating *M* (SD) | Human *M* (SD) | Robot *M* (SD) |
|---|---|---|---|---|
| *Goodness* | | | | |
| Care | **6.348 (.924)** | **2.283 (1.858)** | 4.267 (2.666) | 4.345 (2.469) |
| Fairness | **5.778 (1.857)** | **2.911 (2.265)** | 4.667 (2.471) | 4.158 (2.541) |
| Authority | **6.217 (1.191)** | **2.826 (1.355)** | 4.912 (2.006) | 4.293 (2.177) |
| Loyalty | 5.326 (1.801) | 4.848 (1.699) | 5.147 (1.795) | 5.052 (1.751) |
| Purity | **6.391 (.977)** | **3.674 (1.536)** | 5.265 (1.675) | 4.897 (1.980) |
| Liberty | **6.370 (1.103)** | **1.457 (1.377)** | 3.941 (2.806) | 3.897 (2.764) |
| Nonmoral | 4.956 (1.167) | 5.109 (1.370) | 5.412 (1.373) | 4.807 (1.156) |
| *Responsibility* | | | | |
| Care | **2.044 (1.776)** | **6.087 (1.631)** | 3.971 (2.691) | 4.121 (2.643) |
| Fairness | **2.933 (2.310)** | **5.178 (2.269)** | 3.971 (2.564) | 4.107 (2.549) |
| Authority | **2.065 (1.831)** | **5.370 (1.818)** | 3.823 (2.500) | 3.638 (2.455) |
| Loyalty | 2.978 (2.113) | 3.391 (2.103) | 2.971 (2.081) | 3.310 (2.129) |
| Purity | **2.044 (1.763)** | **4.457 (2.297)** | 3.118 (2.409) | 3.328 (2.365) |
| Liberty | **2.435 (1.858)** | **6.652 (.766)** | 4.500 (2.585) | 4.569 (2.549) |
| Nonmoral | 2.391 (1.513) | 2.565 (1.917) | 2.176 (1.642) | 2.655 (1.753) |

Same-row pairs of Means presented in bold are significantly different across the column manipulations ($p \leq .008$)

humans (including the non-association of fairness and nonmoral behavior) except that responsibility for the action is *not* a factor. Function 2 indicates that for all moral foundations except loyalty, higher goodness paired with lower responsibility were associated with increases in implicit mind, morality, and trust, but *not* in explicit moral and mental status ascription.

To again address RQ2, considering co-present and interactive scenarios: for humans, goodness *and* responsibility behavior ratings are positively associated with moral status and trust (though not with evaluations of minded agency) for all foundations except fairness and loyalty again with a smaller impact than other foundations. For robots, two functions emerge in which (1) perceived goodness for all foundation behaviors (except fairness, and without the influence of perceived responsibility) are positively associated with moral status and trust (but not minded agency), and (2) diverging ratings for foundation-specific goodness (high) and responsibility (low) are associated with higher implicit measures for dependency (i.e., low mindedness), trustworthiness, and moral capacity.

## 5 General Discussion

This investigation reveals both convergent and divergent findings across two studies (summarized in Table 8). (RQ1) People judged agents' behaviors to be similarly good or bad—regardless of the agent performing them. This pattern persisted across moral foundations, except for a small interaction effect in which robots are assigned more credit/blame than humans when they uphold/violate, respectively. When people interacted with the agent in person, moral valence of behaviors *also* influenced perceived agent responsibility (save for loyalty): upholding garners lower responsibility while violating garners higher responsibility. (RQ2) Nearly all discrete-foundation evaluations played a role in evaluations of mind, morality, and trust evaluations—in live interactions, however, loyalty behaviors had no influence on social evaluations of either agent.

Overall, for both robots and humans and across both observed and live interactions, more negative behavior ratings were comprehensively associated with reduced morality and trust ratings. Of note, though, are some divergent patterns between observed and live interactions. For live interactions with humans, assigned responsibility (more blame for violating and credit for upholding) is combined with perceived goodness to impact morality and trust evaluations. For live interactions with robots, responsibility plays a different role: low responsibility paired with higher goodness promotes stronger implicit mind, morality, and trust.

Broadly, findings are interpreted to suggest that bad behavior is seen as an indicator of a bad actor regardless of the performing agent; perceived badness negatively influences perceived morality and trust, but plays little role in mind perception. For humans, there is a link between behavior responsibility and reduced social evaluations. For robots, responsibility is not a consideration in social evaluations such that they may bear a greater burden to behave morally, regardless of their credit- or blame-worthiness in a situation.

**Table 7** Canonical solution for agent evaluations predicting morality ratings in live interactions

| Variables | Function 1 | | | Function 2 | | | $h^2$ (%) |
|---|---|---|---|---|---|---|---|
| | Coef | $r_s$ | $r_s^2$ (%) | Coef | $r_s$ | $r_s^2$ (%) | |
| *Human partner* | | | | | | | |
| Set 1: Agent evaluations | | | | | | | |
| Dependency | − .119 | − .404 | 16.32 | | | | 16.32 |
| Moral Capacity | − .259 | **− .891** | 79.39 | | | | **79.39** |
| Moral Status | .007 | **− .452** | 20.43 | | | | 20.43 |
| Trust– Reliable/Capable | − .275 | **− .838** | 70.22 | | | | **70.22** |
| Trust–Ethical/Sincere | .056 | **− .872** | 76.04 | | | | **76.04** |
| Trust Status | − .573 | **− .949** | 90.06 | | | | **90.06** |
| $R_c^2$ | | | 94.38 | | | | |
| Set 2: Behavior evaluations | | | | | | | |
| Care-Good | − .221 | **− .895** | 80.10 | | | | **80.10** |
| Care-Responsible | .405 | **.895** | 80.10 | | | | **80.10** |
| Fairness-Good | − .183 | − .052 | 00.27 | | | | 00.27 |
| Fairness-Responsible | − .130 | .055 | 00.30 | | | | 00.30 |
| Authority-Good | − .347 | **− .899** | 80.82 | | | | **80.82** |
| Authority-Responsible | .128 | **.669** | 48.86 | | | | **48.86** |
| Loyalty-Good | − .282 | **− .562** | 31.58 | | | | 31.58 |
| Loyalty-Responsible | − .269 | .304 | 09.24 | | | | 09.24 |
| Purity-Good | − .032 | **− .854** | 72.93 | | | | **72.93** |
| Purity-Responsible | .170 | **.507** | 25.71 | | | | 25.71 |
| Liberty-Good | .041 | **− .881** | 77.62 | | | | **77.62** |
| Liberty-Responsible | − .181 | **.725** | 52.56 | | | | **52.56** |
| Nonmoral-Good | − .085 | − .098 | 00.96 | | | | 00.96 |
| Nonmoral-Responsible | .048 | .191 | 03.65 | | | | 03.65 |
| *Robot partner* | | | | | | | |
| Set 1: Agent evaluations | | | | | | | |
| Dependency | − .154 | .065 | 00.42 | .800 | **.895** | 80.10 | **80.52** |
| Moral Capacity | .733 | **.771** | 59.44 | .129 | **.478** | 22.85 | **82.29** |
| Moral Status | .341 | **.612** | 37.45 | .103 | .123 | 01.51 | 38.96 |
| Trust–Reliable/Capable | − .051 | .418 | 17.47 | .463 | **.526** | 27.67 | **45.14** |
| Trust–Ethical/Sincere | − .278 | **.540** | 29.16 | − .189 | **.514** | 26.42 | **55.58** |
| Trust Status | .510 | **.799** | 63.84 | − .382 | − .166 | 02.76 | **66.60** |
| $R_c^2$ | | | 68.70 | | | 64.55 | |
| Set 2: Behavior evaluations | | | | | | | |
| Care-Good | .496 | **.655** | 42.90 | − .003 | **.518** | 26.83 | **69.73** |
| Care-Responsible | .460 | − .302 | 09.12 | .016 | **− .604** | 36.48 | **45.60** |
| Fairness-Good | − .018 | .418 | 17.47 | .443 | **.623** | 38.81 | **56.28** |
| Fairness-Responsible | .103 | − .247 | 06.10 | .251 | **− .525** | 27.56 | 33.66 |
| Authority-Good | − .130 | **.527** | 27.77 | − .180 | **.495** | 24.50 | **52.27** |
| Authority-Responsible | − .278 | − .370 | 13.69 | − .351 | **− .629** | 39.56 | **53.25** |
| Loyalty-Good | .515 | **.653** | 42.64 | − .250 | − .202 | 04.08 | **46.72** |
| Loyalty-Responsible | − .195 | − .420 | 17.64 | .208 | − .006 | 00.00 | 17.64 |
| Purity-Good | − .155 | **.595** | 35.40 | .049 | **.540** | 29.16 | **64.56** |
| Purity-Responsible | − .225 | − .404 | 16.32 | − .199 | **− .565** | 31.92 | **48.24** |
| Liberty-Good | .830 | **.584** | 34.11 | − .236 | **.620** | 38.44 | **72.55** |
| Liberty-Responsible | .398 | − .349 | 12.18 | − .586 | **− .780** | 60.84 | **73.02** |
| Nonmoral-Good | .352 | .174 | 03.03 | − .498 | **− .462** | 21.34 | 24.37 |
| Nonmoral-Responsible | .113 | − .091 | 00.83 | − .103 | − .038 | 00.14 | 00.97 |

Coef = standardized canonical coefficients; *r*s = structure coefficient; $r_s^2$ = squared structure coefficient; $h^2$ = communality coefficient; $R_c^2$= redundancy coefficient. Structure and communality coefficients greater than |.45| are bolded

**Table 8** Summary of study 1 and study 2 findings

| Study 1 (online) | Study 2 (live) | Study 2 differences |
|---|---|---|
| *RQ1: Goodness/responsibility evaluations across moral foundations* | | |
| Moral valence influenced behavior goodness perceptions across all foundations: good/bad behaviors were consistently seen as performed by good/bad actors regardless of agent type | Moral valence influenced behavior goodness perceptions across most foundations: good/bad behaviors were generally seen as performed by good/bad actors regardless of agent type. For fairness alone, robots were assigned more extreme goodness/badness ratings compared to (un)fair humans | Live interactions show main effect of moral valence on responsibility (more responsibility for bad behavior). Live robots sometimes assigned more extreme evaluations than humans for same behavior |
| *RQ2: Moral foundations' contributions to agent social evaluations* | | |
| Behavior goodness for all foundations contribute to social evaluations: for both agents, good behaviors result in higher mind, morality, and trust evaluations. For non-moral norms, good and responsible humans are seen as less trustworthy while good robots were seen as more trustworthy | For humans and robots, most good actions (save fairness/norms) lead to higher morality/trust (but no difference in mind evaluations). Behavior responsibility contributed to evaluations only for humans. For robots, higher goodness/lower responsibility was positively linked to indirect indicators of mind, morality, and trust evaluations | Live interactions saw no effects of fairness or normative actions, no effects of robot responsibility, and no effects on mind ascription. In live interactions, high goodness and low responsibility linked to comprehensively more positive social evaluations of robots |

## 5.1 Moral Judgments Are (Usually) Agent-Agnostic …

The non-impact of manipulated agent-type on behavior evaluations indicates that bad behavior is bad behavior (and good is good), independent of the actor's ontological class. This finding is in line with past scholarship showing that social/moral cognitions are similar between humans and robots so long as social cues are the same (e.g., [60, 77, 78]); however it diverges from evidence that people impose different moral norms on robots than on humans [79]. It is possible that moral judgments are more heuristic and that discrete foundations aren't of material importance, especially given evidence that once one moral foundation is violated people assume that *all* foundations will be violated, and that all violations are interpreted as kinds of harm violations [29]. The agent non-specific pattern in the present data is paired with near-absence of mindedness (signaled via low scores in dependency) in the observed models; it may be that because blame judgments integrate information about mental states (see [80]), mindedness is implied in moral action and therefore not explicitly evaluated.

There are few exceptions to this pattern related to the moral foundation of fairness, which is understood to be an *individualizing* moral foundation—one concerned with rights and freedoms of individual persons, compared to binding foundations that preserve social institutions [20]. It may be that when people are prompted to think abstractly (i.e., to evaluate "goodness") their core values become more salient and valuations of individualizing foundations are heightened [81], and evaluations of injustice may be even more salient than appraisals of harm [82]. Alternatively, fairness is associated with contemporary moral panics around worker displacement; such displacement was alluded to in the stimulus prompt and linked to potential power differentials between humans and machines that are also present in human–human relations [83].

Importantly, however, robots and humans bear different burdens in accounting for their behavior. Evaluations of humans combined goodness and responsibility (bad behavior and high blame contribute to reduced morality/trust); robots were usually evaluated on their behavior *without* consideration for their responsibility, except for the link between increased goodness and reduced responsibility toward higher trust. In other words, robots are generally not afforded the potential to be bad without blame—they may only be good without credit. This follows work suggesting that robots must explicitly, transparently, and *comprehensively* communicate and exhibit their goodness [18] *and* that some other actor (i.e., a developer or engineer) is a conspicuous-yet-absent driver of a robotic agent's behavior [84].

## 5.2 … and Presence May Influence Perceived Moral Agency

Because (a) a main effect of behavior valence on responsibility ratings was exhibited in the live interaction but *not* in the observed interaction and (b) evaluations of agent mindedness were influenced by behavior evaluations in the observed interaction but *not* in the live interaction, social presence may play a role in promoting an actor's perceived moral agency. Regarding the former, it is likely that feeling as though an actor is real and present through delivery of rich social cues [85] fosters an immediacy that renders perceptions of responsible agency salient. It may also be that the social presence inherent to the live interaction increased self-relevance of agent responses. Intimacy with a possible event is ego-centric: the self is the reference point such that the more direct the experience, the more concrete the construal of the event [80, 86]. Regarding the latter, it is possible that non-interactive observations permitted more conscious inferencing of mental status compared to the automatic social cognitions inherent to the live interaction, where immediacy and strong visual/vocal cueing may promote similar mind-attribution for both agents. In other words, viewing both agents via video may have prompted consideration of them as characters (cf. [87]) versus in-person as agents.

Notably, the present studies' designs do not allow for disentangling the potential influences of co-location engendering social presence and/or that the co-location afforded the opportunity to interact rather than merely observe; future research should tease out these potential influences. It also is prudent to acknowledge that cross-study differences in blame judgments may also be a matter of sample differences. Study 1 drew on a U.S.-representative sample with varied demographics; Study 2 leveraged a convenience sample from a community that values rugged individualism and personal responsibility. Finally, because the researcher was co-present during the interaction in Study 2 (due to safety concerns), it is possible that the experimenter effects were at play in promoting differences between the mediated and live interactions (either through mere presence or through potential pressure to answer in particular ways); future research should determine the extent to which human mere presence effects may contribute to differential mental- and moral-capacity evaluations.

### 5.3 Limitations and Future Research

In addition to the aforementioned directions for future research, the present studies' designs carry inherent limitations that should be addressed. Participants experienced agent observations or interactions that depicted entirely upholding or entirely violating behaviors as moral foundations are understood to be variably weighted (and so variably exhibited) by individuals. Stimulus scenarios presented also contained content that may have been confounded with moral foundations such that it is possible, for example, that effects for fairness are actually effects of discussing job retention or responses to liberty may have been a function of the severity of the relatively extreme human-trafficking exemplar. Finally, all moral dilemmas presented the agent or another as the target of the (im)moral behavior such that people may react differently if the behavior is self-relevant—that is, if they are to benefit or suffer as a result of the behavior—or if some scenario actors are other robots rather than humans. Future research should build on this work by attending to these limitations: designs that consider moral and blame judgment effects on social-moral cognitions through mixed-valence behaviors, content-consistent foundation scenarios, and self-relevant scenarios. In tandem, future work should consider the potential for different robot morphologies to impact social evaluations.

## 6 Conclusion

The present research suggests that moral judgments of behavior are largely agent-agnostic, but agents bear different burdens with regard to social evaluations: to foster trust and moral status, humans must be seen as performing good behaviors and being responsible for those behaviors while robots must be good but are not afforded credit for that goodness. Findings show the possibility to foster social integration of robots based on exhibitions of human-normative moral behavior: data suggest a link between comprehensive "good" behavior and trust and moral status. Trust fosters social commitments with machines through the perception of positive contributions to human life [88]. Indeed, some perspectives count the subjective experience of robot "heart" as emerging in the ostensible space between technological actualities and human possibilities [89]—that space may be the technical performance of human moral behaviors.

**Acknowledgements** The author gratefully acknowledges the contributions of Koji Yoshimura (Study 2 data collection), Ashley Churchill (Study 1 stimulus production, Study 2 data collection), Allison Eden (theoretical consultation), Matthew Grizzard and Nicholas David Bowman (expert review of stimulus scenarios), and Kevin Koban (critical review of manuscript).

**Funding** The author discloses receipt of the following financial support for the research, authorship, and publication of this article: This material is based upon work supported by the Air Force Office of Scientific Research under Award Number FA9550-19-1-0006.

## Compliance with Ethical Standards

**Conflict of interest** The author declares that there is no conflict of interest.

## References

1. Asimov I (1942) Runaround. I, Robot. Doubleday, New York, p 40
2. Banks J (2019) A perceived moral agency scale: development and validation of a metric for humans and social machines. Comput Hum Behav 90:363–371
3. Eden A, Grizzard M, Lewis RJ (2012) Moral psychology and media theory. In: Media and the moral mind. Routledge, New York, pp 1–25
4. Graham J, Nosek BA, Haidt J, Iyer R, Koleva S, Ditto PH (2011) Mapping the moral domain. J Pers Soc Psychol 101(2):366–385
5. Iyer R, Koleva S, Graham J, Ditto P, Haidt J (2012) Understanding libertarian morality: the psychological dispositions of self-identified Libertarians. PLoS ONE 7(8):e42366
6. Nomura T, Kanda T, Suzuki T (2006) Experimental investigation into influence of negative attitudes toward robots on human–robot interaction. AI Soc 20(2):138–150
7. Olivera-La Rosa A (2018) Wrong outside, wrong inside: a social functionalist approach to the uncanny feeling. New Ideas Psychol 50:38–47
8. Edwards C, Edwards A, Spence PR, Westerman D (2016) Initial interaction expectations with robots: testing the human-to-human interaction script. Commun Stud 67(2):227–238
9. Kahn PH, Reichert AL, Gary HE, Kanda T, Ishiguro H, Shen S, Ruckert JH, Gill B (2011) The new ontological category hypothesis in human-robot interaction. In: Proceedings of HRI'11, Lausanne, Switzerland
10. Bedaf S, Draper H, Gelderblom G-J, Sorell T, de Witte L (2016) Can a service robot which supports independent living of older people disobey a command? Int J Social Robot 8:409–420
11. Ullman D, Malle BF (2018) What does it mean to trust a robot? Steps toward a multidimensional measure of trust. In: HRI'18 companion, New York
12. 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. Comput Hum Behav 61:633–655
13. Packard C, Boelk T, Andres J, Edwards C, Edwards A, Spence PR (2019) The Pratfall Effects and interpersonal impressions of a robot that forgets and apologizes. In: 14th ACM/IEEE international conference on human-robot interaction (HRI)
14. Horstmann AC, Bock N, Linhuber E, Szczuka JM, Straßmann C, Krämer NC (2018) Do a robot's social skills and its objection discourage interactants from switching the robot off? PLoS ONE 13(7):e0201581
15. Rossi A, Dautenhahn K, Koay KL, Walters ML (2018) The impact of peoples' personal dispositions and personalities on their trust of robots in an emergency scenario. J Behav Robot 9:137–154
16. Johnson AM, Axinn S (2013) The morality of autonomous robots. J Military Ethics 12(2):129–141
17. Aha W, Coman A (2017) The AI rebellion: changing the narrative. In: Proceedings of the thirty-first AAAI conference on artificial intelligence, Palo Alto, CA
18. Malle BF (2014) Moral competence in robots? In: Proceedings of robo-philosophy 2014, Amsterdam
19. Malle BF, Scheutz M (2015) When will people regard robots as morally competent social partners? In: Ro-Man: 24th IEEE international symposium on robot and human interactive communication, New York
20. Haidt J (2013) The righteous mind: Why good people are divided by politics and religion. Vintage Books, New York
21. Kohlberg L (1971) Stages of moral development. Moral Educ 1(51):23–92
22. Haidt J (2001) The emotional dog and its rational tail: a social intuitionist approach to moral judgement. Psychol Rev 108(4):814–834
23. Greene J, Haidt J (2002) How (and where) does moral judgment work? Trends Cognit Sci 6(12):517–523
24. Wiltshire TJ (2015) A prospective framework for the design of ideal artifical moral agents: insights from the science of heroism in humans. Minds Mach 25(1):57–71
25. Kramer MF, Borg JS, Conitzer V, Sinnott-Armstrong W (2018) When do people want AI to make decisions. In: AIES'18, New York
26. Shank DB, DeSanti A (2018) Attributions of morality and mind to artificial intelligence after realworld. Comput Hum Behav 86:401–411
27. Foot P (1967) The problem of abortion and the doctrine of double effect. Oxford Rev 5:5–15
28. Crone DL, Laham SM (2015) Multiple moral foundations predict responses to sacrificial dilemmas. Pers Individ Differ 85:60–65
29. Gray K, Waytz A, Young L (2012) The moral dyad: a fundamental template unifying moral judgment. Psychol Inq 23(2):206–215
30. Graham J, Haidt J, Koleva S, Motyl M, Iyer R, Wojcik SP, Ditto PH (2013) Moral foundations theory: the pragmatic validity of moral pluralism. In: Advances in experimental social psychology, vol 47. Academic Press, pp 55–130
31. Haidt J, Joseph C (2004) Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. Daedalus 133(4):55–66
32. Singler B (2019) Existential hope and existential despair in AI apocalypticism and transhumanism. Zygon 54(1):156–176
33. Lotz V, Himmel S, Ziefle M (2019) You're my mate–acceptance factors for human-robot collaboration in industry. In: International conference on competitive manufacturing, Stellenbosch, South Africa
34. Johansson-Pajala R-M, Thommes K, Hoppe JA, Tuiska O, Hennala L, Pekkarinen S, Melkas H, Gustafsson C (2019) Improved knowledge changes the mindset: older adults' perceptions of care robots. In: International conference on human-computer interaction, Cham
35. Horowitz MC (2016) Public opinion and the politics of the killer robots debate. Res Politics 3(1)
36. Ötting SK, Gopinathan S, Maier GW, Steil JJ (2017) Why criteria of decision fairness should be considered in robot design. In: 20th ACM conference on computer-supported cooperative work and social computing, New York
37. Sundar SS (2008) The MAIN model: a heuristic approach to understanding technology effects on credibility. In: Metzger MJ, Flanagin AJ (eds) Digital media, youth, and credibility. MIT Press, Cambridge, pp 73–100
38. Short E, Hart J, Vu M, Scassellati B (2010) No fair!!: an interaction with a cheating robot. In: Proceedings of the 5th ACM/IEEE international conference on human-robot interaction, New York
39. Sandoval EB, Brandstetter J, Bartneck C (2016) Can a robot bribe a human? The measurement of the negative side of reciprocity in human robot interaction. In: Eleventh ACM/IEEE international conference on human robot interaction
40. Hoffman G, Forlizzi J, Ayal S, Ssteinfeld A, Antanitis J, Hochman G, Hochendoner E, Finkenaur J (2015) Robot presence and human honest: experimental evidence. In: Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction, New York
41. Wilson DH (2005) How to survive a robot uprising: tips on defending yourself against the coming rebellion. Bloomsbury, New York
42. Fraune MR, Šabanović S, Smith ER (2017) Teammates first: favoring ingroup robots over outgroup humans. In: 26th IEEE international symposium on robot and human interactive communication (RO-MAN)
43. Clothier RA, Williams BP, Perez T (2019) Autonomy from a safety certification perspective. In: 8th Australian aerospace congress, Brisbane

44. Li J, Ju W, Nass C (2015) Observer perception of dominance and mirroring behavior in human-robot relatinoships. In: Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction, New York
45. Gombolay MC, Gutierrez RG, Clarke SG, Sturla GF, Shah JA (2015) Decision-making authority, team efficiency and human worker satisfaction in mixed human–robot teams. Autonomous Robots 39(3):293–312
46. Seo SH, Geiskkovitch D, Nakane M, King C, Young JE (2015) Poor thing! Would you feel sorry for a simulated robot? In: 10th ACM/IEEE international conference on human-robot interaction
47. Waytz A, Young L (2019) Aversion to playing God and moral condemnation of technology and science. Philos Trans R Soc B, p [online before print]
48. Trovato G, Pariasca F, Ramirez R, Cerna J, Reutskiy V, Rodriguez L, Cuellar F (2019) Communicating with SanTO–the first Catholic robot. In: The 28th IEE international symposium on robot and human interactive communication, New York
49. Gunkel DJ (2018) Robot rights. MIT Press, Cambridge
50. Craig MC, Edwards C, Edwards A, Spence PR (2019) Impressions of message compliance-gaining strategies for considering robot rights. In: 14th ACM/IEEE international conference on human-robot interaction (HRI)
51. Jipson JL, Gelman SA (2007) Robots and rodents: children's inferences about living and nonliving kinds. Child Dev 78:1675–1688
52. Voiklis J, Kim B, Cusimano C, Malle BF (2016) Moral judgments of human versus robot agents. In: Ro-Man: 25th IEEE international symposium on robot and human interactive communication, New York
53. Shank DB, DeSanti A, Maninger T (2019) When are artificial intelligence versus human agents faulted. Inf Commun Soc 22(5):648–663
54. Johnson G (2006) Computer systems: moral entities but not moral agents. Ethics Inf Technol 8:195–204
55. Malle BF, Guglielmo S, Monroe AE (2014) A theory of blame. Psychol Inq 25(2):147–186
56. Monroe AE, Malle BF (2019) People systematically update moral judgments of blame. J Pers Soc Psychol 116(2):215–236
57. Haslam N (2006) Dehumanization: an integrative review. Pers Soc Psychol Rev 10(3):252–264
58. Waytz A, Young L (2018) Morality for us versus them. In: Atlas of moral psychology. Guildford Press, New York, pp 186–192
59. Premack D, Woodruff G (1978) Does the chimpanzee have a theory of mind? Behav Brain Sci 1(4):515–526
60. Banks J (2019) Theory of mind in social robots: replication of five established human tests. Int J Soc Robot vol [Online in advance of print], p np
61. Perez-Osario J, Wykowska A (2019) Adopting the intentional stance toward natural and artificial agents. Philos Psychol [preprint]
62. Voiklis J, Malle BF (2017) Moral cognition and its basis in social cognition and social regulation. In: Atlas of moral psychology. Guilford Press, New York, pp 108–120
63. Delgado MR, Frank RH, Phelps EA (2005) Perceptions of moral character modulate the neural systems of reward during the trust game. Nat Neurosci 8:1611–1618
64. Pavlou PA (2013) Consumer acceptance of electronic commerce: integrating trust and risk with the technology acceptance model. Int J Electron Commer 7(3):101–134
65. Barber B (1983) The logic and limits of trust. Rutgers University Press, New Brunswick
66. Hancock PA, Billings DR, Schaeger KE, Chen JY, de Visser EJ, Parasuraman R (2011) A meta-analysis of factors affecting trust in human-robot interaction. Hum Factors 53:517–527
67. 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. In: Proceedings of 10th conference of the media psychology division of the German Psychological Society, Münster
68. Bureau UC (2010) U.S. census by decade. [Online]. https://www.census.gov/programs-surveys/decennial-census/decade.2010.html
69. Graham J, Haidt J, Nosek BA (2009) Liberals and conservatives rely on different sets of moral foundations. J Pers Soc Psychol 96(5):1029–1046
70. Clifford S, Iyengar V, Cabez R, Sinnott-Armstrong W (2015) Moral foundations vignettes: a standardized stimulus database of scenarios based on moral foundations theory. Behav Res Methods 47(4):1178–1198
71. McCurrie CH, Crone DL, Bigelow F, Laham SM (2018) Moral and Affective Film Set (MAAFS): a normed moral video database. PLoS ONE 13(11):e0206604
72. Bartneck C, Kulić D, Croft E, Zoghbi S (2009) Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. Int J Social Robot 1(1):71–81
73. Graham J, Haidt J (2012) Sacred values and evil adversaries: a moral foundations approach. In: The social psychology of morality: exploring the causes of good and evil. APA, Washington, DC. pp 11–31
74. Ullman D, Malle BF (2019) Measuring gains and losses in human-robot trust: evidence for differentiable components of trust. In: Proceedings of the 14th ACM/IEEE international conference on human-robot interaction
75. Landmann H, Hess U (2018) Testing moral foundation theory: Are specific moral emotions elicited by specific moral transgressions? J Moral Educ 47:34–47
76. Sherry A, Henson RK (2005) Conducting and interpreting canonical correlation analysis in personality research. J Pers Assess 84(1):37–48
77. Cross ES, Ramsey R, Liepelt R, Priz W, Hamilton A (2016) The shaping of social perception by stimulus and knowledge cues to human animacy. Philos Trans R Soc B 371(1686):20150075
78. Eyssel F, Hegel F, Horstmann G, Wagner C (2010) Anthropomorphic inferences from emotional nonverbal cues. In: 19th IEEE international symposium on robot and human interactive communication
79. Malle BF, Scheutz M, Arnold T, Voiklis J, Cusimano C (2015) Sacrifice one for the good of many? People apply different moral norms to human and robot agents. In: Tenth annual ACM/IEEE international conference on human-robot interaction, New York
80. Guglielmo S, Malle BF (2019) Asymmetric morality: blame is more differentiated and more extreme than praise. PLoS ONE 14(3):e0213544
81. Napier JL, Luguri JB (2013) Moral mind-sets: abstract thinking increases a preference for individualizing over binding moral foundations. Soc Psychol Pers Sci 4(6):754–759
82. Piazza J, Sousa P, Rottman J, Syropoulus S (2018) Which appraisals are foundational to moral judgment? Harm, injustice, and beyond. Soc Psychol Pers Sci 10(7):903–913
83. Ju W (2016) Power in human robot interactions. In: What social robots can and should do. IOS Press, Amsterdam, pp 13–14
84. Sullins JP (2006) When is a robot a moral agent? Int Rev Inf Ethics 6:23–30
85. Lombard M, Ditton T (1997) At the heart of it all: the concept of presence. J Comput-Mediat Commun 3(2)
86. Trope Y, Liberman N (2010) Construal-level theory of psychological distance. Psychol Rev 117(2):440–463
87. Schneider S (2001) Toward a cognitive theory of literary character: the dynamics of mental-model construction. Style 35(4):607–640

88. Michael J, Salice A (2017) The sense of commitment in human-robot interaction. Int J Social Robot 9(5):755–763
89. Katsuno H (2011) The robot's heart: tinkering with humanity and intimacy in robot-building. Japanese Stud 31(1):94–109

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Jaime Banks** (Ph.D., Colorado State University) is a researcher and Associate Professor at Texas Tech University, College of Media & Communication. Her research focuses on understanding human-machine communication processes and effects, with an emphasis on relational dynamics related to social robots and digital agents/avatars. Currently, Jaime's primary research considers the perceived moral agency of robots as it relates to trust and persuasive influence.