Malle, B. F., & Ullman, D. (2021). A multi-dimensional conception and measure of human-robot trust. In C. S. Nam and J. B. Lyons (eds.), *Trust in human-robot interaction: research and applications* (pp. 3-25). Elsevier.

A Multi-Dimensional Conception and Measure of Human-Robot Trust

Bertram F. Malle and Daniel Ullman, Brown University

Abstract

Robots are increasingly used in social applications, which raise challenges regarding people's trust in robots. A modern conception of human-robot trust must go beyond the conventional notions of human-automation relations and better connect to the current understanding of human-human trust, without assuming that human-robot trust is identical. A review of the literature together with our recent empirical work suggest that trust is multi-dimensional, incorporating both performance aspects (central in the human-automation literature) and moral aspects (central in the human-human trust literature). A multi-dimensional conception can be applied to human-robot trust, even if only some of the dimensions will be relevant for any given interaction with a robot. In addition to proposing an integrative conception of trust, we offer a measurement instrument for public use: the Multi-Dimensional Measure of Trust (MDMT). This measure captures two superordinate factors of trust (Performance trust, Moral trust) that each break into two subfacets (Reliable and Capable within Performance, and Sincere and Ethical within Moral). We are continuing to test this measure in follow-up research and encourage other researchers to join us in collectively validating it.

Keywords
trust, human-robot trust, moral, social robotics, human-robot interaction, trust measurement

You are standing in front of a robot, with your back to it. Your task is to let yourself fall backwards, and the robot's task is to catch you. Would you trust the robot?

Commonly referred to as the "trust fall" exercise, this situation serves as a quintessential example of trust between two agents. The exercise hinges on the falling agent's expectation that the catching agent will not let them hit the ground. This expectation has two features: First, the falling agent needs to believe that the catching agent is *capable* of catching them before they hit the ground. Second, the falling agent needs to believe that the catching agent has the moral integrity and commitment to catch them and is sincere when promising to do so. The first idea is what we henceforth call "Performance trust," while the second idea is what we call "Moral trust." Performance trust refers to the trustor's confidence that the trustee is capable of completing a given task, while moral trust refers to the trustor's confidence that the trustee will choose the morally right act and not exploit the trustor's vulnerability. These two beliefs can diverge: We may be confident that another agent is capable of doing what benefits us but not be

morally motivated to do so; or we may be confident of the other's moral commitment but not in their capacity to fulfill it. Thus, trust seems to have at least two conceptually independent dimensions.

Robots are beginning to offer benefits to humans in a range of settings, from schools to the workplace, and across a variety of application domains, from medical care to support in one's home. Robots have stepped out of their cages of isolation and safety precautions, where their primary contribution is to reliably complete repetitive physical tasks. Increasingly, robots are used in the world of social interaction, where their expected contribution is to assist and support people. But for robots to successfully fulfill roles in this social context of human-robot interaction, people must be willing to interact with these robots, entrust them with tasks that have socially beneficial results, and be confident that the robots are both capable of and committed to bringing about those results. Because trust is integral to social relationships, people will be inclined to trust robots with whom they interact. If robots are to succeed in these social interactions, we must first design robots that are worthy of human trust.

The Concept of Trust

Trust, in common usage, is an expectation that something good will happen—while also knowing that it might not happen. Hope also rests on an expectation of something good, but hope has a relatively low level of confidence (Bruininks & Malle, 2006) whereas trust carries a high level of confidence. The Oxford English Dictionary calls trust a "firm belief in the reliability, truth, or ability of someone or something" or the "confident expectation of something." Collins COBUILD Advanced English Dictionary (which aims to specifically capture current use) also refers to a "firm belief or confidence in the honesty, integrity, reliability, justice, etc. of another person or thing." We set aside here trust in "things" and focus on persons or agents—including robots. Our thesis may then be formulated this way: Trust's underlying expectation can be directed at multiple different properties that the other agent might have, and these properties make up multiple dimensions of trust. The dictionary entries mention *ability, reliability, honesty*, and *integrity*, and we will provide evidence that these are indeed four major dimensions of trust: one can trust someone who is Reliable, Capable, Ethical, and Sincere. We will review literature on human-automation trust, human-human trust, and human-robot trust and will show that these four dimensions repeatedly appear in these literatures. In light of emerging data from our lab, we then suggest that a more complete model of trust needs to incorporate all four dimensions, and we offer an instrument that makes them conveniently measurable.

Trust in Human-Automation Interaction

Work in the domain of automation emphasizes the performance of automated systems. Automation serves its purpose when a system is able to safely and efficiently perform a task that reduces a burden on human users. Many of these task domains have historically consisted of nonsocial tasks, often combined with system oversight by a human operator but little collaboration between human and system. A useful heuristic in thinking about whether or not to trust a system or an agent comes in the form of a basic question: "What do I worry about that

would prevent me from interacting with this agent?" In the automation literature, such worry is tied to the **performance** of the system; and this worry can be alleviated, and allows for trust, if the system is capable of performing its task, and does so consistently (Schaefer, Chen, Szalma, & Hancock, 2016).

Sheridan and Parasuraman (2005) reviewed research in human-automation interaction and offered two sets of features of trust. One set comes from the system's lower-level **reliability** of performance, while the other comes from the system's higher-level **ability**. Ideally, trust in a system is grounded in an accurate conception of the system's ability and reliability; however, trust is not always appropriately calibrated. Parasuraman and Riley (1997) discussed multiple miscalibrations: misuse (overreliance on automation), disuse (underutilization of automation), and abuse (inappropriate application of automation). Lee and See (2004) proposed, based on previous research, that systems must be designed to help match users' expectations to the system's actual performance capabilities. Calibrated trust is grounded in an accurate assessment of what a system can and cannot do—as well as why the system fails when it does.

Because errors reveal a system's performance quality they have been a particular focus in human-automation work. For example, Madhavan, Wiegmann, and Lacson (2006) showed that trust in an automated decision aid declined when the user observed system errors. However, not all errors were treated the same way: when a system made errors on easy trials rather than difficult trials, users mistrusted the system far more. People potentially use task difficulty as a diagnostic indicator: failing to complete easy trials shows low ability.

In the absence of trust in a system, users will opt to not use it. Lee and Moray (1992) investigated the determinants that influence whether human operators will rely on a system performing a task, or opt for manual control in the task. They used an experimental task where human operators would choose between manual control or automatic control for operating a simulated semi-automatic pasteurization plant. The researchers found a tradeoff between human operators' self-confidence and trust in the system, such that operators opted for automatic control when trust in the system exceeded self-confidence and opted for manual control when self-confidence exceeded trust in the system. Trusting a system involves accepting some kind of risk and believing that the system is able to limit that risk.

A recent review of 127 empirical studies on human trust in automation (Hoff & Bashir, 2015) identified numerous factors that affect trust, including culture, personality, task characteristics, work load, self-confidence, and more; but the object of trust itself—the automation system—was described only in terms of its performance: its reliability, predictability, and error-proneness. In sum, in the literature on trust in automation, the primary focus is on appropriately matching a person's expectations for a system with information about the performance of the system. These systems are motiveless, and users are not concerned about being betrayed, exploited, or deceived by the system. Trust here is the expectation that a system will perform a task as intended and expected, and the only worries concern the system's **reliability** and **ability**.

Trust in Human-Human Interaction

As we move from the domain of human-automation trust into the domain of human-human trust, we can pose the same heuristic question: "What do I worry about that would prevent me from interacting with this agent?" Unlike the primary focus on **performance trust** in human-automation interaction, here the focus widens to include expectations of whether a human agent will act morally—what we refer to as **moral trust**. In situations of human-human trust, the question becomes whether a human agent will exploit another human's potential vulnerability—either unintentionally because of a lack of ability, or intentionally because of a lack of moral integrity.

It is instructive to imagine what a social community without trust would look like. People would not expect that another person would have their best interest in mind; instead, they would expect that others do what benefits them even when such actions exploit or harm others. Lack of trust is thus a reflection of a society without prosocial norms, without moral commitments. By contrast, in societies that have such commitments, trust is possible and has the power to enable and sustain cooperative behavior (e.g., Gambetta, 1988; Jones & George, 1998). Trust acts as a glue that enables people to live and work together without constant worry and threat of being exploited. In fact, such societies uphold a norm to trust other people (Dunning, Anderson, Schlosser, Ehlebracht, & Fetchenhauer, 2014), which places demands on those people to justify the trust they are granted—and makes violations of trust all the more salient.

Philosophical analyses typically treat trust as a three-part relation among a trustor, a trustee, and something that the trustor expects the trustee to do (Hardin, 2002) or to care for (Baier, 1986). The oft-cited conceptualization of trust by Mayer, Davis, and Schoorman (1995) has at its core a three-part relationship as well, but in addition to the trustor and a trustee, the authors focus on the role of risk. Integrating these elements, we can say that the trustor typically expects the trustee to act so as to avoid or reduce the trustor's risk in the situation. The object of this expectation (what the trustee is expected to do or be) corresponds to the notion of trustworthiness—a trustee's characteristics that either inspire or justify the trusting expectation: their ability, reliability, ethical integrity, and so on. The exact characteristics of trustworthiness, and thus the objects of trust expectations, are debated. We therefore review several of these proposals, ranging from few to many characteristics, and identify the common denominators. With those in hand we can then examine the place of human-robot trust in the broader psychological landscape of trust.

Rotter (1967) conceptualized trust as "an expectancy...that the word, promise, verbal or written statement of another...can be relied upon" (p. 651). Rotter's proposal thus highlighted the trust expectation of **sincerity** (truthful words and standing by promises), which is confirmed by a closer look at the items in his interpersonal trust scale. Chun and Campbell (1974) conducted cluster and factor analyses on this scale, and in their results we see the primary interpersonal facet of **sincerity** (honest, truthful). Selfish exploitation formed another factor, perhaps the opposite pole of what other authors call **benevolence**. Many of the other items in Rotter's scale referred to institutional trust and specifically to concerns about institutions being **sincere** and

ethical (e.g., unbiased, not cheating). However, there is no mention of competence or reliability/ predictability. The worries people have about others are cast here entirely in terms of morality.

Several authors who examined human-human trust, from sociology to management, highlighted **competence** and **integrity** as the major expectations (Kim, Dirks, & Cooper, 2009; Parsons, 1969). In Cook and Wall's (1980) conception, trust was organized into faith in the intentions of others (often labeled **benevolence**), and confidence in the **ability** and the **reliability** of others. However, when the authors measured trust in an organizational setting, these three aspects did not come apart. Barber (1983), in his analysis of the broader societal role of trust, distinguished between three trust expectations: **persistence** (predictability), **competence**, and moral duties (to have others' interests in mind, which most authors call **benevolence**). Slovic, Flynn, Johnson, and Mertz (1993) asked ordinary citizens to express their trust or distrust in power plant management as a function of various behaviors by the management; the authors' choice of such behaviors revealed their conception of trust as expectations about **competence** (e.g., being prepared for accidents) and about two moral dispositions that are often labeled **benevolence** (good motives) and **sincerity/transparency** (being truthful, providing access). However, no attempt was made to distinguish these expectations in measurement. Focusing on people's attitudes toward institutions, Carnevale's (1995, p. xi) definition of trust included **reliability** and **competence** as well three moral aspects: **nonthreatening** (benevolent), **fair**, and **ethical**. Caldwell and Clapham's (2003) proposal, tailored to organizations, included seven aspects of trustworthiness that can be divided broadly into **competence** (knowledge, ability), responsibility to inform (**transparency**), and various **moral** or normative aspects (quality assurance, respect, legal compliance).

Gabarro (1978) conducted interviews with four company presidents and 33 subordinates over the span of three years. He extracted six objects of trust. While maintaining aspects of **competence** (skills and good judgments) and **consistency** (reliability, predictability), he differentiated the moral dimension into **integrity** (encompassing honesty and moral character), **motives** (benevolent motives, commitment), **openness** (defined as honesty, being straight, not hiding—aspects most other authors label **sincerity**), and **discreetness** (not violating confidence). Butler and Cantrell (1984) and Schindler and Thomas (1993) experimentally tested five of these characteristics (omitting discreetness, perhaps because it could be grouped under integrity) as determinants of overall trust judgments. **Integrity** and **competence** showed the most considerable impact, while consistency was weak, and openness had little to no impact.

Butler (1991) also interviewed managers and content-analyzed characteristics that the managers mentioned in describing trusted and mistrusted people. In subsequent scale development and iterated factor analyses, Butler postulated nine characteristics, but only six seem to directly capture trust: **competence** and **consistency** as the familiar performance aspects, as well as the moral characteristics of **integrity**, **fairness/loyalty**, **discreetness**, and **promise fulfillment**. (The others were availability, receptivity, and openness.) However, when examined in people's judgments, Butler's moral characteristics were very highly correlated (*r*s between .65 and .76), suggesting one large moral cluster. **Competence** was somewhat differentiated from

these moral components (correlating with them in the .40s and .50s), while **consistency** more clearly set itself apart (correlating with all the other components in the .30s and .40s).

Despite differences among the various authors' conceptions of human-human trust, we see that almost all of them support a multi-dimensional concept. Most authors assigned prominent status to competence and reliability, mirroring the human-automation literature, but many added a moral dimension, with anywhere from one to four moral facets. Mayer et al. (1995) tried to integrate these variations and to consolidate them into the major characteristics of trustworthiness that sway a trustor. Against the background of 23 previous proposals, they derived three such characteristics: **ability** (including knowledge, expertise, competence), **benevolence** (a positive orientation toward the trustor), and **integrity** (adhering to moral principles shared with the trustor). Two omissions are noteworthy here. First, the authors excluded predictability from the conceptual space, mainly because they argued that reliability was not sufficient for trust. However, none of the individual characteristics are sufficient for trust, so reliability should not be discarded. Second, sincerity (common among many other models) was absent, not because of direct empirical evidence but because of the authors' conceptual decisions in selecting and compiling characteristics of trust. Interestingly, McKnight, Cummings, and Chervany (1998) cited Mayer as the basis for their conception of trust expectations but actually worked with four dimensions, including **competence** (ability), **predictability** (added back in), **honesty** (rather than integrity, thus bringing sincerity back into the picture), and **benevolence**.

Nonetheless, a meta-analysis showed that Mayer et al.'s three-dimensional conception is successful in predicting trust states (overall willingness to accept vulnerability) from trust expectations (Colquitt, Scott, & LePine, 2007). The predictive correlations were in the .60s for each dimension, but the three dimensions were also correlated with each other between $r = .62$ and $r = .68$. Because reliability and sincerity were not included in the meta-analysis we do not know what their role would be in affecting subjective trust states.

A decade later, after yet more and varied proposals of conceptualizing trust, Burke, Sims, Lazzara, and Salas (2007) tabulated 27 such conceptualizations. We performed a frequency analysis of the most-used content words in these definitions (see Table 1) and found considerable common ground in what trust is: a dyadic relation in which one person accepts vulnerability because they expect that the other person's future action will be governed by certain characteristics. And though the specific characteristics (of trustworthiness) are rarely mentioned in the definitions, those that are mentioned include the by now familiar notions of **ability**, **reliability**, and a bundle of **moral characteristics** such as **benevolence**, **honesty**, and **integrity**.

Our review of dimensions of trust expectations in the literature is summarized in Table 2. Though not all assignments are clear-cut, the overall picture suggests that trust has two sides: a **performance** side, with facets of competence and reliability, and a **moral** side, with facets of sincerity, integrity, and benevolence. The importance of this moral dimension is what appears to differentiate trust between humans from trust in automation, and this difference is best explained by the significance of social interaction and risks involved in human-human relationships. The question now arises whether human-robot trust has anything like a moral dimension.

Table 1. Frequency analysis of the words used in 27 definitions of trust (Burke et al., 2007).

68	**Dyadic**	28	Trustor, one, individual, person, party
		29	Trustee, another one, other(s)
		11	Relationship(s), interpersonal, interdependence, interactions, reciprocal, mutual
32	**Accept vulnerability**	17	Risk(s), vulnerable, vulnerability, damage, harmful, stake
		15	Accept(ing), willing(ness)
23	**Future action**	15	Action(s), behavior, act, behave, behavioral, perform, fulfill(s)
		8	Intent(ion), will
19	**Expectation**		Belief(s), believes, expectation(s), assumption, perceives, perception
13	**Objects of trust**	6	Benevolent(ly), helpful, concern, welfare
		3	Honest, fairness, integrity
		2	Reliability, predictable
		2	Ability, competent
12	**Other**	4	Cognitive
		4	Affective, emotions, feel
		2	Possible, potential
		2	Safety, security

Table 2. Dimensions of trust expectations (characteristics of trustworthiness) in selected models

	Performance		*Moral*		
	Competence	Reliability	Integrity	Sincerity	Benevolence
Rotter (1967)				✅	
Chun & Campbell (1974)			✅	✅	✅
Gabarro (1978)	✅	✅	✅	✅[1]	✅[4]
Parsons (1969)	✅		✅		
Cook & Wall (1980)	✅	✅			✅
Barber (1983)	✅	✅			✅[3]
Butler (1991)	✅	✅	✅		✅[5]
Mayer et al. (1995)	✅		✅		✅
Slovic et al. (1993)	✅			✅	✅
Carnevale (1995)	✅	✅	✅		✅
McKnight et al. (1998)	✅	✅	✅[2]	✅[2]	✅
Caldwell & Clapham (2003)	✅		✅		✅
Analysis of definitions in Burke et al. (2007)	✅	✅	✅		✅
Kim et al. (2009)	✅		✅		

[1] Calls it "openness" but explicates as honest, straight, and not hiding things—arguably elements of sincerity.
[2] Call it "honesty," which, according to premier dictionaries has both meanings of being sincere, truthful and having integrity, moral principles (though with a stronger emphasis on shades of sincerity).
[3] Calls it "moral duties" (with a focus on having others' interests in mind)
[4] Calls it "motives" and arguably refers to benevolent motives
[5] Calls it "fairness/loyalty"

Trust in Human-Robot Interaction

Much of the extant work on human-robot trust focuses on the reliability and ability characteristics of robotic systems, so trust in intelligent robots is considered along the same factors seen in the broader human-automation literature (Yanco, Desai, Drury, & Steinfeld, 2016). For example, Hancock et al. (2011) conducted a meta-analysis of factors that prior research has identified as influencing trust in human-robot interaction. The authors collated 21 studies and found that factors related to the robot—specifically, robot performance (such as reliability)—had the strongest association with trust. Human-related factors (e.g., attitudes and comfort with robots) and environmental factors (e.g., culture and physical environment) contributed relatively less.

In a recent review, Lewis, Sycara, and Walker (2018) argued for a distinction between performance-based interactions between humans and robots and social-based interactions. This

social focus points to some of the distinguishing features that set robots apart from automation. In particular, robots are being introduced into more and more social settings, projected to be social companions for older adults, tutors for children in schools, or assistants to people with health needs (Broadbent, Stafford, & MacDonald, 2009). Such contexts often involve an element of risk. Risk may arise from a robot moving around an older adult's house and accidentally knocking them to the ground, which is a traditional safety concern; but risk may also arise from a robot refusing to increase a person's pain medication because the decision authority cannot be reached, or from a robot unknowingly presenting incorrect information to a child. The latter situations involve social and moral norms and thus raise the question of moral trust in a robot. But do people actually treat robots as moral agents?

When asked to infer the mental capacities of robots, respondents are inclined to grant robots some capacity for moral decision-making (Malle, 2019; Weisman, Dweck, & Markman, 2017), and significantly more so the more humanlike the robot looks (Malle, Zhao, & Phillips, 2019). By contrast, people do not welcome moral decisions by cars and other machines if they lack humanlike mental capacities (Bigman & Gray, 2018). An increasing number of studies also show that people apply moral norms to AI and robotic agents and blame those agents when they violate the pertinent norms (Malle, Magar, & Scheutz, 2019; Malle, Scheutz, Arnold, Voiklis, & Cusimano, 2015; Shank & DeSanti, 2018). In a human-robot interaction study, too (Kahn, Jr. et al., 2012), a majority of people thought of the robot as morally accountable for a specific transgressive behavior. If people treat robots as capable of moral decision making, then there is room for developing trust in robots if they are sincere, ethical, and benevolent—and to lose trust if they are not.

A series of studies investigated whether people recognize a robot's attempt to cheat in a game of rock-paper-scissors by changing its gesture or in a game of Battleship by lying about a ship's position; and people did recognize those violations (Short, Hart, Vu, & Scassellati, 2010; Ullman, Leite, Phillips, Kim-Cohen, & Scassellati, 2014). Such recognition, rather than the belief that the robot was simply malfunctioning, may reveal that people saw the robot as being unethical or insincere. Indeed, Wijnen, Coenen, and Grzyb (2017) showed that a lying robot (diverting blame to a human for a negative act the robot committed) was trusted less in a behavioral trust game than an honest robot.

For people to gain or lose moral trust in a robot, the robot does not have to be a genuine moral agent. Even when the designer is really the one who is insincere and untrustworthy, the human interacting with the robot may well direct their trust and disappointment at the robot. And what holds for sincerity could in principle hold for benevolence and integrity as well, though these two characteristics demand more behavioral evidence from a robot.

Evidence for ethical integrity would be seen in following moral norms and principles: for example, being fair, nondiscriminatory, cooperative, and respectful (Kuipers, 2018; Malle & Scheutz, 2019). Especially when robots are taking on a broader range of roles (e.g., as educators or health care providers; Broadbent et al., 2009), the norms associated with these roles must be made explicit and the robot must be equipped to comply with them. To build robots that guide their behavior by adhering to social and moral norms is still a significant challenge given the

current state of the art in robotics (Malle, Bello, & Scheutz, 2019); but even just seeing a robot *try* to follow norms is likely to instill a considerable amount of trust in people.

Benevolence requires putting one's own interests behind others' interests, rather than selfishly benefitting while others incur costs. Even when robots are not designed to be selfish, they will sometimes pursue goals that run counter to people's interests—such as when a robot hinders or inconveniences someone or is unable to meet a request. In some cases, a robot's built-in goals may purposefully circumvent people's interests or values—such as when a robot tries to coax people into buying something or revealing private information (Calo, 2011; Hartzog, 2015). Such behaviors are often disguised and therefore insincere as well, suggesting overall lack of ethical standards.

Even if current robots have only minimal moral capacities, their ever-widening roles in society, their routine spoken communication, and their often humanoid looks will (and perhaps already do) prompt people to treat them as social and moral agents (Coeckelbergh, 2012). Robots are machines, evaluated for their performance, but increasingly they are also social agents, evaluated for their potential to cause harm. In a context of vulnerability, people are likely to consider a robot's moral characteristics of trustworthiness, such as sincerity and integrity. We must therefore measure these moral characteristics along with the familiar characteristics of ability and reliability. What measurement tools are available to assess these multi-dimensional layers of human-robot trust?

Measuring Human-Robot Trust

Hancock et al. (2011) noted that there is a scarcity of validated measures with which to evaluate human-robot trust consistently across experimental designs—which makes the phenomenon difficult to study and compare across studies, labs, and domains. This echoes the call by Steinfeld et al. (2006) for a singular toolkit of human-robot interaction metrics that include trust.

A discussion of measuring human-robot trust must first address the distinction between trust as a subjective state and trust as a choice or action (Kee & Knox, 1970). Although often trust becomes apparent to the outside observer only when the agent makes a choice that embraces some risk, such a choice is ultimately the product of an internal state of trust that is translated into action. Mayer et al. (1995) similarly distinguished between the willingness to accept one's vulnerability and the action of taking a risk. These two aspects, the subjective-internal and the public-behavioral, can be measured separately and are predictably related (Colquitt et al., 2007). However, a risk-taking choice is not always grounded (solely) in subjective trust. Imagine a forced choice to rely on one of two people—relying on Agent A is not necessarily indicative of trust in Agent A, but it could be due to a much greater benefit one stands to gain from choosing Agent A or a fear of Agent B. Thus, even though the act of risk-taking is often an important consequence of trust, trust itself is more closely captured by an internal state (Lewis et al., 2018). Obviously, internal states cannot be measured directly but are themselves operationalized by verbal reports, nonverbal expressions, and the like.

In the human-human domain, this internal state of trust consists of an acceptance of one's vulnerability and of expectations that the other's characteristics warrant this acceptance. These

expectations, we have seen, are multi-dimensional and include at least performance characteristics (ability, reliability) and one or more moral characteristics (e.g., sincerity, integrity, benevolence). The measurement tools currently available for human-robot trust vary in how much they consider **moral trust** and **performance trust** factors, but they are heavily skewed toward the latter.

Schaefer (2013) offered one of the more comprehensive developments of a trust measure. Across six experiments, Schaefer developed a 40-item trust scale to assess the three antecedents of human trust in robots considered by Hancock et al. (2011): features of the robot, the human, and the environment. In addition, a 14-item scale focused specifically on trust expectations: characteristics that make the robot trustworthy. Of these 14 items, eleven reflect dimensions of **performance** (e.g., function successfully, act consistently) while three relate to social aspects (provide appropriate information, communicate with people). No item in the short form directly invokes the moral dimensions, but a few items in the longer 40-item scale refer to open and truthful communication (i.e., sincerity), and one item mentions protecting people. Similarly, Yagoda and Gillan (2012) proposed a scale for human-robot interaction in teams that identifies various system features (e.g., sensor data, effectors, interface), all evaluated with the words consistent, dependable, and **reliable** (and a few with the words understandable, accessible), but without reference to moral aspects.

Madsen and Gregor (2000) developed a trust measure for human-computer interaction that contains 25 items covering five content domains: perceived **reliability**, perceived technical **competence**, perceived understandability, faith (in a system's advice and solutions), and personal attachment (to a system). The items in these domains juxtapose subjective experiences (being attached to and understanding the system) with familiar trust expectations regarding **ability** and **reliability** (e.g., "The system makes use of all the knowledge and information available to it to produce its solution to the problem"; "The system responds the same way under the same conditions at different times"). The authors did not numerically compare the fit of various factor structures but favored a two-factor solution in which understandability dominated the second factor and virtually all ability and reliability items (along with attachment) hung together in the first factor. No aspects of moral trust were included.

Not all scales are limited to performance measurement. Jian, Bisantz, and Drury (2000) used a bottom-up approach to examine the semantic field of trust-related expressions. After examining a considerable number of such expressions, the authors arrived at a 12-item scale. Three items capture expectations of **reliability** (confident in, dependable, reliable), two items capture **competence** (provides security, harmful outcomes), and four items capture **moral** characteristics—one representing trustworthiness (**integrity**) and three representing its absence (deceptive, underhanded, suspicious of intent). However, a subsequent confirmatory factor analysis did not uncover separation between any of these dimensions, only between distrust-related and trust-related items (Spain, Bustamante, & Bliss, 2008).

We have by no means considered all extant measures of human-robot trust. But it is clear that the large majority does not consider moral aspects, and no specific trust measure appears to exist that reliably captures one or more moral aspects in human-robot interaction. Given the

growing similarities between human-robot and human-human relations, there is a need to address the moral aspects of people's trust in robots above and beyond performance aspects. We should not assume that these aspects hold for all robots, but they may hold for some of them—and for those robots, a measurement tool must be available. Moreover, as ethical requirements for robot behavior increase (Arkin, Ulam, & Wagner, 2012; Malle & Scheutz, 2014, 2019; Wallach & Allen, 2008), measuring the moral dimension of trust is key to evaluating the success of designing such ethical robots. We now introduce our initial steps of developing a measurement tool for performance and moral trust that can be used in both human-human and human-robot situations.

A Multi-Dimensional Conception and Measurement of Trust

Our review of the human-human trust literature indicated reasonable consensus that trust can be defined as follows:

> Trust = a dyadic relation in which one person accepts vulnerability because they expect that the other person's future action will have certain characteristics; these characteristics include some mix of performance (ability, reliability) and/or morality (honesty, integrity, and benevolence).

We take this definition as our starting point to introduce a multi-dimensional measure of trust. We believe that the subjective state of trusting (accepting vulnerability) cannot easily be divorced from the trust *expectations* regarding the other's capability, reliability, and morality; instead, the subjective state is typically directed at one or more of these characteristics. That is, when people say, "X trusts Y," they (implicitly or explicitly) refer to X trusting that Y is capable of and/or reliable in performing a certain action and/or sincere when uttering a statement (e.g., promise, information offer) and/or has the moral integrity not to exploit or otherwise harm X. Measuring these focal characteristics of trust expectations lies at the heart of our proposed instrument. At the same time, we encourage researchers to present a separate overall question of subjective trust (e.g., Mayer & Davis, 1999), and this question may have additional predictive validity for certain ensuing beliefs or behaviors (Colquitt et al., 2007).

Trust Words in Semantic Space

We started our investigation of multi-dimensional trust by conceptually mapping out the space within which people think about trust, independent of whether it is trust in people, institutions, or robots (Ullman & Malle, 2018). Initially we considered two candidate dimensions of this space: trusting that an agent is capable of completing a task ("capacity trust") and trusting that an agent will not place another at risk ("personal trust"). We collected 62 words from dictionaries, the trust literature, and published trust measures and asked participants (recruited via Amazon Mechanical Turk) to indicate where each word fell on a slider scale from "more similar to capacity trust" to "more similar to personal trust" (defined as above). These original items can be viewed in a supplementary document available on our lab website (http://research.clps.brown.edu/SocCogSci/Measures/index.html). Whereas many previous measures consist of sentences related to trust, we opted for simplicity in this task and used single

words or very short phrases. We then engaged in an iterative process of Principal Components Analysis (PCA) and item analysis and arrived at 32 items distributed over four components, which we represented by the labels Reliable (7 items), Capable (8 items), Sincere (6 items), and Ethical (11 items). Given that we had asked people to rate words only on the single capacity-personal variable, the conceptual structure that emerged was surprising, as these components were strikingly similar to the trust expectations identified by our review of human-automation trust models (Reliable and Capable) and human-human trust models (Sincere and Ethical). Additional item analysis allowed us to shorten each cluster to five items, yielding four initial subscales of trust: Reliable (count on, depend on, reliable, faith in, confide in, $\alpha = .72$), Capable (capable, diligent, rigorous, accurate, meticulous, $\alpha = .88$), Sincere (sincere, genuine, truthful, benevolent, authentic, $\alpha = .84$), and Ethical (honest, principled, reputable, respectable, scrupulous, $\alpha = .87$). The intercorrelations among the four components revealed that Sincere and Ethical were related to each other ($r = .46$, $p = .01$), suggesting that "moral trust" encompasses two related facets.

Sorting Trust Words

We sought to replicate the four dimensions and their item clusters in a second study, employing a guided sorting task. 60 participants recruited via Amazon Mechanical Turk were asked to consider 32 words or short phrases, six to seven for each of the hypothesized four dimensions as well as five filler items assumed to be unrelated to trust (e.g., humorous, intentional). The words or short phrases were either taken from trust words in the study described above (Ullman & Malle, 2018) or were added to reflect other published trust work. Participants then indicated how well they thought each item described different "person types" by sorting the items into one of four boxes. Each box represented a person with a single character trait—the markers of the four hypothesized dimensions: Reliable, Capable, Sincere, and Ethical (a fifth category was labeled "Other" for words people believed did not fit any person type). We computed sorting consensus scores as the percentages of participants who classified a given item into a given category and then grouped items by sorting consensus (see Figure 1). The replication succeeded: Each hypothesized dimension of trust expectations contained largely the same items as in the semantic space study (Ullman & Malle, 2018).

	Sorting Category				
Items	**Reliable**	**Capable**	**Sincere**	**Ethical**	**Other**
Reliable	92%	5%	0%	2%	2%
Can count on	82%	7%	8%	3%	0%
Consistent	77%	20%	2%	2%	0%
Can depend on	70%	15%	5%	7%	3%
Responsible	58%	23%	0%	17%	2%
Can have faith in	56%	7%	15%	19%	3%
Steadfast	49%	27%	8%	3%	12%
Capable	2%	92%	3%	2%	2%
Competent	12%	85%	3%	0%	0%
Intelligent	3%	80%	5%	0%	12%
Meticulous	13%	65%	3%	2%	17%
Diligent	31%	54%	7%	3%	5%
Accurate	39%	51%	2%	0%	8%
Rigorous	22%	42%	3%	3%	29%
Sincere	2%	0%	88%	5%	5%
Genuine	2%	2%	88%	3%	5%
Authentic	5%	5%	71%	15%	3%
Honest	0%	2%	51%	47%	0%
Truthful	2%	2%	47%	47%	2%
Transparent	5%	0%	42%	29%	24%
Can confide in	27%	5%	33%	25%	10%
Ethical	0%	0%	5%	93%	2%
Principled	3%	3%	3%	88%	2%
Respectable	14%	7%	8%	66%	5%
Trustworthy	23%	3%	27%	47%	0%
Scrupulous	7%	20%	2%	37%	35%
Reputable	28%	15%	12%	35%	10%
Humorous	3%	3%	5%	2%	86%
Easy going	2%	2%	23%	3%	70%
Creative	0%	42%	2%	0%	56%
Benevolent	7%	5%	25%	27%	36%
Intentional	10%	27%	25%	5%	32%

Figure 1. Consensus rates (0-100%) of sorting task in which participants grouped 32 items as describing persons who are Reliable, Capable, Sincere, or Ethical. Words in italics designate filler items that were not considered part of a trust dimension. Degree of consensus is shaded from dark grey (0%) to bright green (100%).

Performance trust		Moral trust	
Reliable	**Capable**	**Sincere**	**Ethical**
Reliable	Capable	Sincere	Ethical
Predictable	Skilled	Genuine	Respectable
Can count on	Competent	Candid	Principled
Consistent	Meticulous	Authentic	Has integrity

Figure 2. Items within each subscale of the Multi-Dimensional Measure of Trust (MDMT), developed after sorting study and used in expectation change study.

We then selected three items from the top four items of each cluster and added one new face-valid item to each cluster: *predictable* (for Reliable), *skilled* (for Capable), *candid* (for Sincere), and *has integrity* (for Ethical). Sixteen items (four in each of the dimensions) thus constituted the first version of a measurement tool we labeled the Multi-Dimensional Measure of Trust (MDMT), shown in Figure 2.

Changes in trust expectations

We used the MDMT in a subsequent study that examined whether the four subscales of the MDMT (see Figure 2) are differentially sensitive to dimension-specific information one receives about a robot's behavior (Ullman & Malle, 2019). For example, the Sincere subscale should show an increase when new evidence arises about an agent's sincerity. 798 participants recruited via Amazon Mechanical Turk read a baseline sentence about a robot's behavior and provided initial trust ratings on the MDMT. Then they received new information designed to prompt either an increase or decrease in dimension-specific trust in the robot, followed by final ratings on the MDMT. The subscales had acceptable to good internal consistency (Cronbach's α): *Reliable*, $\alpha = .92$; *Capable*, $\alpha = .92$; *Sincere*, $\alpha = .79$; *Ethical*, $\alpha = .81$. However, a Principal Component Analysis (PCA) on the 16 MDMT change scores (before-after evidence manipulation) suggested only a two-dimensional structure (with $\lambda > 1$). Under orthogonal rotation, the first component contained all eight *Reliable* and *Capable* items (explaining 35.2% of the variance) and the second contained all eight *Ethical* and *Sincere* items (explaining 29.4% of the variance). Oblique rotation led to the same item divisions but allowed the two dimensions to correlate at $r = .66$.

In the analysis of subscale sensitivity, each of the four subscales was highly sensitive to evidence change overall (Fs > 100.0, ps < .001), $\eta^2 = .17$-$.43$. However, the subscales were not systematically sensitive to only the information about their dimension-specific evidence (e.g., the *Reliable* subscale did not respond only to evidence of being reliable). This may have been a result of the largely ineffective manipulations of trust-relevant evidence along the Capable and Sincere dimensions ($\eta^2 = .0$ to $.03$), whereas manipulations were strong along the Reliable

dimension (η^2 = .06 to .24) and the Ethical dimension (η^2 = .13 to .23). However, even without full dimensional specificity, we saw that the measures of *Reliable* and *Capable* moved in tandem across manipulated evidence, and so did the measures of *Sincere* and *Ethical*. Thus, in this initial study, the major differentiation between Performance trust and Moral trust did emerge, while the finer differentiations into the two component subscales of performance trust (i.e., Reliable and Capable) and the two component subscales of moral trust (i.e., Sincere and Ethical) did not. The limited effectiveness of the evidence manipulations temper our conclusions. We are continuing to replicate and validate this conceptual model of trust in follow-up studies, including studies that use the MDMT in laboratory human-robot interaction studies. We invite researchers to contribute to the measure's validation by using it in their own studies and sharing the results with us (see http://research.clps.brown.edu/SocCogSci/Measures/CurrentVersion_MDMT.pdf).

Discussion

We have accumulated evidence for a two-factor model of human-robot trust. In this model, people trust or distrust a robot as a function of two trust expectations: about the robot's performance characteristics (that it is capable of completing a task and/or will reliably do so) and about the robot's moral characteristics (that it will complete a task in adherence to social and moral norms). In addition, the literature and initial empirical evidence suggest a four-dimensional model, which poses numerous questions that future research must address.

Open Questions

First, we do not know yet whether people differentiate robot performance trust further into capability and reliability components and robot moral trust further into sincerity and ethical integrity components (and potentially benevolence). Though our first two studies suggest they do, our third study left us more cautious. Second, we do not yet know whether human-robot trust necessarily comes with a feeling of vulnerability that is characteristic of human trust (Colquitt et al., 2007; Mayer et al., 1995). We have used the more common term *worry* to denote the core feeling of trust, which can be directed at robots much like at humans: people worry about the robot's potential failure to complete a task, and for some robots, people may worry about the robot lying, cheating, or exploiting them. Third, in the human-human trust literature, researchers have distinguished between state trust and trait trust, and it stands to reason that people may similarly develop a trait-like propensity to trust or distrust robots—as a function of cultural exposure (e.g., science fiction), public discussion (e.g., recent worries about robots taking many human jobs), and actual experiences with robots. Fourth, some intriguing results in the human-human trust literature suggest that there is a social norm to trust others—even when exploitation is a real possibility. For example, Dunning et al. (2014) showed that most people trust a stranger (in the behavioral-economic Trust game) even when they are not confident that the other will make the benevolent choice; and they do so at least in part because they feel obligated to express respect toward others by trusting them. We might imagine that there is no such obligation to respect robots—or perhaps there is, if human observers are present during the human-robot interaction and the person does not want to appear distrustful. Only additional empirical studies will tell.

Moral Trust Expectations

Our empirical results and subsequent proposal of four dimensions of human-robot trust are not fully compatible with the widely-used model of human-human trust by Mayer et al. (1995). Their model omits the reliability dimension and identifies moral dimensions of integrity and benevolence, rather than integrity and sincerity. Several points must be considered to reconcile these differences. First, because of the close similarity between robots and automated systems, the reliability dimension appears necessary to consider in human-robot trust. Whether it also must be considered separately in human-human trust is an open question. Second, whereas Mayer et al.'s selection of dimensions was based primarily on theoretical arguments, the four dimensions we put forth were discovered empirically (albeit with a small set of data so far). In fact, in two of our studies, we had included the item "benevolent," but it clustered in the Sincere component, and it later dropped out because of its lower loading in the PCA and lower sorting consensus. Third, Mayer et al.'s model was developed for the business domain, and it is plausible that this domain activates a particular worry about another person following only economic self-interest rather than having benevolent motives.

In sum, the exact composition of the moral dimension of trust is not yet settled. Given the literature and our own data we would be surprised if neither reliability nor sincerity were central to people's trust expectations. However, we remain open to the possibility that three moral characteristics exist (integrity, sincerity, and benevolence), even if perhaps benevolence is more specifically useful in the business domain.

The Multi-Dimensional Measure of Trust (MDMT)

Trust in human-automation interaction, human-human interaction, and human-robot interaction share a number of similarities. The Multi-Dimensional Measure of Trust (MDMT) offers a single measurement instrument that flexibly adapts to the relevant dimensions of each domain and provides a standard of comparison across studies and domains. The measure is short (16 items) and easy to use. We have built a measurement feature into the scales that allows people to "opt out" of some of the terms, if they think the specific terms do not apply (e.g., *has integrity* for an industrial robot). Researchers can also selectively use specific subscales (e.g., when evaluating a clearly nonsocial robot in a low-stakes task). Moreover, adapting the measure to more trait-like trust expectations could be accomplished with minimal reformulations. Finally, future work should investigate the extent to which trust ratings obtained using the MDMT converge with or diverge from existent trust measures, including measures of overall feelings of trust and of trusting actions. We hope others will join in this endeavor (see http://research.clps.brown.edu/SocCogSci/Measures/CurrentVersion_MDMT.pdf).

Conclusion

"What do I worry about that would prevent me from interacting with this agent?" The answer to this question looks slightly different across domains of interaction, but the core ideas about trust stay the same. A multi-dimensional conceptualization of trust is able to integrate the different answers—considering trust in an agent's performance (Reliable, Capable) as well as trust in an

agent's morality (Sincere, Ethical), whether the agent is a human or a robot. As we design robots that raise questions of trust in human-robot interactions, we must investigate all dimensions of trust, even if some of them are only slowly coming to light.

Acknowledgements

This work was supported by Office of Naval Research grants N00014-14-1-0144 and N00014-16-1-2278. Daniel Ullman was supported by the National Space Grant College and Fellowship Program, Space Grant Opportunities in NASA STEM (NNX15AI06H) and the National Defense Science & Engineering Graduate Fellowship (NDSEG) Program.

References

Arkin, R. C., Ulam, P., & Wagner, A. R. (2012). Moral decision making in autonomous systems: Enforcement, moral emotions, dignity, trust, and deception. *Proceedings of the IEEE*, *100*(3), 571–589. https://doi.org/10.1109/JPROC.2011.2173265

Baier, A. C. (1986). Trust and antitrust. *Ethics*, *96*, 231–260.

Barber, B. (1983). *The logic and limits of trust*. New Brunswick, N.J.: Rutgers University Press.

Bigman, Y. E., & Gray, K. (2018). People are averse to machines making moral decisions. *Cognition*, *181*, 21–34. https://doi.org/10.1016/j.cognition.2018.08.003

Broadbent, E., Stafford, R., & MacDonald, B. (2009). Acceptance of healthcare robots for the older population: Review and future directions. *International Journal of Social Robotics*, *1*(4), 319. https://doi.org/10.1007/s12369-009-0030-6

Bruininks, P., & Malle, B. F. (2006). Distinguishing hope from optimism and related affective states. *Motivation and Emotion*, *29*(4), 324–352.

Burke, C. S., Sims, D. E., Lazzara, E. H., & Salas, E. (2007). Trust in leadership: A multi-level review and integration. *The Leadership Quarterly*, *18*(6), 606–632. https://doi.org/10.1016/j.leaqua.2007.09.006

Butler, J. K. (1991). Toward understanding and measuring conditions of trust: Evolution of a conditions of trust inventory. *Journal of Management*, *17*(3), 643–663. https://doi.org/10.1177/014920639101700307

Butler, J. K., & Cantrell, R. S. (1984). A behavioral decision theory approach to modeling dyadic trust in superiors and subordinates. *Psychological Reports*, *55*(1), 19–28. https://doi.org/10.2466/pr0.1984.55.1.19

Caldwell, C., & Clapham, S. E. (2003). Organizational trustworthiness: An international perspective. *Journal of Business Ethics*, *47*(4), 349–364. Retrieved from JSTOR.

Calo, M. R. (2011). Robots and privacy. In P. Lin, K. Abney, & G. A. Bekey (Eds.), *Robot Ethics: The Ethical and Social Implications of Robotics* (pp. 187–201). Cambridge, US: MIT Press.

Carnevale, D. G. (1995). *Trustworthy government: Leadership and management strategies for building trust and high performance*. San Francisco: Jossey-Bass Publishers.

Chun, K.-T., & Campbell, J. B. (1974). Dimensionality of the Rotter Interpersonal Trust Scale. *Psychological Reports*, *35*(3), 1059–1070. https://doi.org/10.2466/pr0.1974.35.3.1059

Coeckelbergh, M. (2012). Can we trust robots? *Ethics and Information Technology*, *14*(1), 53–60. https://doi.org/10.1007/s10676-011-9279-1

Colquitt, J. A., Scott, B. A., & LePine, J. A. (2007). Trust, trustworthiness, and trust propensity: A meta-analytic test of their unique relationships with risk taking and job performance. *Journal of Applied Psychology*, *92*(4), 909–927. https://doi.org/10.1037/0021-9010.92.4.909

Cook, J., & Wall, T. (1980). New work attitude measures of trust, organizational commitment and personal need non-fulfilment. *Journal of Occupational Psychology*, *53*(1), 39–52. https://doi.org/10.1111/j.2044-8325.1980.tb00005.x

Dunning, D., Anderson, J. E., Schlosser, T., Ehlebracht, D., & Fetchenhauer, D. (2014). Trust at zero acquaintance: More a matter of respect than expectation of reward. *Journal of Personality*, *107*(1), 122–141. https://doi.org/10.1037/a0036673

Gabarro, J. J. (1978). The development of trust, influence and expectations. In A. Athos & J. J. Gabarro (Eds.), *Interpersonal behavior*. Retrieved from https://www.hbs.edu/faculty/Pages/item.aspx?num=7743

Gambetta, D. (1988). *Trust: Making and breaking cooperative relations*. Oxford, UK: Basil Blackwell.

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., de Visser, E. J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors*, *53*(5), 517–527.

Hardin, R. (2002). *Trust and trustworthiness*. New York, NY: Russell Sage Foundation.

Hartzog, W. (2015). Unfair and deceptive robots. *Maryland Law Review*, *74*(4), 785–832.

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, *57*(3), 407–434. https://doi.org/10.1177/0018720814547570

Jian, J.-Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated system. *International Journal of Cognitive Ergonomics*, *4*(1), 53–71.

Jones, G. R., & George, J. M. (1998). The experience and evolution of trust: Implications for cooperation and teamwork. *The Academy of Management Review*, *23*(3), 531–546. https://doi.org/10.2307/259293

Kahn, Jr., P. H., Kanda, T., Ishiguro, H., Gill, B. T., Ruckert, J. H., Shen, S., … Severson, R. L. (2012). Do people hold a humanoid robot morally accountable for the harm it causes? *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction*, 33–40. https://doi.org/10.1145/2157689.2157696

Kee, H. W., & Knox, R. E. (1970). Conceptual and methodological considerations in the study of trust and suspicion. *The Journal of Conflict Resolution*, *14*(3), 357–366.

Kim, P. H., Dirks, K. T., & Cooper, C. D. (2009). The repair of trust: A dynamic bilateral perspective and multilevel conceptualization. *The Academy of Management Review*, *34*(3), 401–422.

Kuipers, B. (2018). How can we trust a robot? *Communications of the ACM*, *61*(3), 86–95. https://doi.org/10.1145/3173087

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Lee, J., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, *35*(10), 1243–1270. https://doi.org/10.1080/00140139208967392

Lewis, M., Sycara, K., & Walker, P. (2018). The role of trust in human-robot interaction. In H. A. Abbass, J. Scholz, & D. J. Reid (Eds.), *Foundations of Trusted Autonomy* (pp. 135–159). https://doi.org/10.1007/978-3-319-64816-3_8

Madhavan, P., Wiegmann, D. A., & Lacson, F. C. (2006). Automation failures on tasks easily performed by operators undermine trust in automated aids. *Human Factors*, *48*(2), 241–256. https://doi.org/10.1518/001872006777724408

Madsen, M., & Gregor, S. (2000). Measuring human-computer trust. In G. Gable & M. Viatle (Eds.), *Proceedings of the 11th Australasian Conference on Information Systems* (pp. 53–64).

Malle, B. F. (2019). How many dimensions of mind perception really are there? In E. K. Goel, C. M. Seifert, & C. Freksa (Eds.), *Proceedings of the 41st Annual Meeting of the Cognitive Science Society* (pp. 2268–2274). Montreal, Canada: Cognitive Science Society.

Malle, B. F., Bello, P., & Scheutz, M. (2019). Requirements for an artificial agent with norm competence. In *Proceedings of 2nd ACM conference on AI and Ethics (AIES'19)*. New York, NY: ACM.

Malle, B. F., Magar, S. T., & Scheutz, M. (2019). AI in the sky: How people morally evaluate human and machine decisions in a lethal strike dilemma. In M. I. Aldinhas Ferreira, J. Silva Sequeira, G. Singh Virk, M. O. Tokhi, & E. E. Kadar (Eds.), *Robotics and Well-Being* (pp. 111–133). https://doi.org/10.1007/978-3-030-12524-0_11

Malle, B. F., & Scheutz, M. (2014). Moral competence in social robots. In *Proceedings of IEEE International Symposium on Ethics in Engineering, Science, and Technology, Ethics'2014* (pp. 30–35). Chicago, IL: IEEE.

Malle, B. F., & Scheutz, M. (2019). Learning how to behave: Moral competence for social robots. In O. Bendel (Ed.), *Handbuch Maschinenethik [Handbook of machine ethics]*. https://doi.org/10.1007/978-3-658-17484-2_17-1

Malle, B. F., 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 *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction, HRI'15* (pp. 117–124). New York, NY: ACM.

Malle, B. F., Zhao, X., & Phillips, E. (2019). Beyond anthropomorphism: Differentiated inferences about robot mind from appearance. *CHI'19 Extended Abstracts*. Presented at the Computer-Human Interaction, CHI'19, Glasgow, Scotland.

Mayer, R. C., & Davis, J. H. (1999). The effect of the performance appraisal system on trust for management: A field quasi-experiment. *Journal of Applied Psychology*, *84*(1), 123–136. https://doi.org/10.1037/0021-9010.84.1.123

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *The Academy of Management Review*, *20*(3), 709–734. https://doi.org/10.2307/258792

McKnight, D. H., Cummings, L. L., & Chervany, N. L. (1998). Initial trust formation in new organizational relationships. *The Academy of Management Review*, *23*(3), 473–490. https://doi.org/10.2307/259290

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, *39*(2), 230–253. https://doi.org/10.1518/001872097778543886

Parsons, T. (1951). *The social system*. Glencoe, IL: Free Press.

Rotter, J. B. (1967). A new scale for the measurement of interpersonal trust1. *Journal of Personality*, *35*(4), 651–665. https://doi.org/10.1111/j.1467-6494.1967.tb01454.x

Schaefer, K. E. (2013). *The perception and measurement of human-robot trust* (PhD Thesis, STARS. Electronic Theses and Dissertations). Retrieved from https://stars.library.ucf.edu/etd/2688/

Schaefer, K. E., Chen, J. Y. C., Szalma, J. L., & Hancock, P. A. (2016). A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Human Factors*, *58*(3), 377–400. https://doi.org/10.1177/0018720816634228

Schindler, P. L., & Thomas, C. C. (1993). The structure of interpersonal trust in the workplace. *Psychological Reports*, *73*(2), 563–573. https://doi.org/10.2466/pr0.1993.73.2.563

Shank, D. B., & DeSanti, A. (2018). Attributions of morality and mind to artificial intelligence after real-world moral violations. *Computers in Human Behavior*, *86*, 401–411. https://doi.org/10.1016/j.chb.2018.05.014

Sheridan, T. B., & Parasuraman, R. (2005). Human-automation interaction. *Reviews of Human Factors and Ergonomics*, *1*(1), 89–129. https://doi.org/10.1518/155723405783703082

Short, E., Hart, J., Vu, M., & Scassellati, B. (2010). No fair!! an interaction with a cheating robot. *Proceedings of the 5th ACM/IEEE International Conference on Human-Robot Interaction*. Presented at the Osaka, Japan. Osaka, Japan.

Slovic, P., Flynn, J., Johnson, S., & Mertz, C. K. (1993). The dynamics of trust in situations of risk. In *Report no. 93-2*. Eugene, Oregon: Decision Research.

Spain, R. D., Bustamante, E. A., & Bliss, J. P. (2008). Towards an empirically developed scale for system trust: Take two. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *52*(19), 1335–1339. https://doi.org/10.1177/154193120805201907

Steinfeld, A., Fong, T., Kaber, D., Lewis, M., Scholtz, J., Schultz, A., & Goodrich, M. (2006). Common metrics for human-robot interaction. *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction*, 33–40. https://doi.org/10.1145/1121241.1121249

Ullman, D., Leite, L., Phillips, J., Kim-Cohen, J., & Scassellati, B. (2014). Smart human, smarter robot: How cheating affects perceptions of social agency. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *36*, 2996–3001.

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*, 263–264. https://doi.org/10.1145/3173386.3176991

Wallach, W., & Allen, C. (2008). *Moral machines: Teaching robots right from wrong*. New York, NY: Oxford University Press.

Weisman, K., Dweck, C. S., & Markman, E. M. (2017). Rethinking people's conceptions of mental life. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(43), 11374–11379. https://doi.org/10.1073/pnas.1704347114

Wijnen, L., Coenen, J., & Grzyb, B. (2017). "It's not my fault!": Investigating the effects of the deceptive behaviour of a humanoid robot. *Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction, HRI 2017, Vienna, Austria, March 6-9, 2017*, 321–322. https://doi.org/10.1145/3029798.3038300

Yagoda, R. E., & Gillan, D. J. (2012). You want me to trust a ROBOT? The development of a human–robot interaction trust scale. *International Journal of Social Robotics*, *4*(3), 235–248. https://doi.org/10.1007/s12369-012-0144-0

Yanco, H. A., Desai, M., Drury, J. L., & Steinfeld, A. (2016). Methods for developing trust models for intelligent systems. In R. Mittu, D. Sofge, A. Wagner, & W. F. Lawless (Eds.), *Robust Intelligence and Trust in Autonomous Systems* (pp. 219–254). https://doi.org/10.1007/978-1-4899-7668-0_11