CHAPTER NUMBER X

TRUST AND THE DISCREPANCY BETWEEN EXPECTATIONS AND ACTUAL CAPABILITIES OF SOCIAL ROBOTS

BERTRAM F. MALLE, KERSTIN FISCHER,
JAMES E. YOUNG, AJUNG MOON,
EMILY COLLINS

Corresponding author:
Bertram F. Malle, Professor
Department of Cognitive, Linguistic, and Psychological Sciences
Brown University
190 Thayer St.
Providence, RI 02912, USA
bfmalle@brown.edu
+1 (401) 863-6820

Kerstin Fischer, Professor (WSR)
Deptment of Design and Communication
University of Southern Denmark
Alsion 2
DK-6400 Sonderborg, Denmark
kerstin@sdu.dk
Phone: +45-6550-1220

James E. Young, Associate Professor
Department of Computer Science
University of Manitoba
Winnipeg, Manitoba R3T 2N2, Canada
Email: young@cs.umanitoba.ca
Phone: (lab) +1-204-474-6791

AJung Moon, Assistant Professor
Department of Electrical and Computer Engineering
McGill University
3480 University Street
Montreal, Quebec H3A 0E9, Canada
ajung.moon@mcgill.ca
Phone: +1-514-398-1694

Emily C. Collins, Research Associate
Department of Computer Science
University of Liverpool
Ashton Street
Liverpool L69 3BX, UK
E.C.Collins@liverpool.ac.uk
Phone: +44 (0)151 795 4271

Abstract

From collaborators in factories to companions in homes, social robots hold the promise to intuitively and efficiently assist and work alongside people. However, human trust in robotic systems is crucial if these robots are to be adopted and used in home and work. In this chapter we take trust to be a set of expectations about the robot's capabilities and explore the risks of discrepancies between a person's expectations and the robot's actual capabilities. We examine major sources of these discrepancies and ways to mitigate their detrimental effects. No simple recipe exists to help build justified trust in human-robot interaction. Rather, we must try to understand humans' expectations and harmonize them with robot design over time.

Introduction

As robots continue to be developed for a range of contexts where they work with people, including factories, museums, airports, hospitals, and homes, the field of Human-Robot Interaction explores how well people will work with these machines, and what kinds of challenges will arise in their interaction patterns. Social robotics focuses on the social and relational aspects of Human-Robot Interaction, investigating how people respond to robots cognitively and emotionally, how they use their basic interpersonal skills when interacting with robots, and how robots themselves can be designed to facilitate successful human-machine interactions.

Trust is a topic that currently receives much attention in human-robot interaction research. If people do not trust robots, they will not collaborate with them or accept their advice, let alone purchase them and delegate to them the important tasks they have been designed for. Building trust is therefore highly desirable from the perspective of robot developers. A closer look at trust in human-robot interaction, however, reveals that the concept of trust itself is multidimensional. For instance, one could trust another human (or perhaps robot) that they will carry out a particular task reliably and without errors, and that they are competent to carry out the task. But in some contexts, people trust another agent to be honest in their communication, sincere in their promises, and to value another person's, or the larger community's interests. In short, people may trust agents based on evidence of reliability, competence, sincerity, or ethical integrity [1], [2][1]. What unites trust along all these dimensions is that it is an *expectation*—expecting that the other is reliable, competent, sincere, or ethical. Expectations, of course, can be disappointed. When the other was not as reliable, capable, or sincere as one thought, one's trust was misplaced. Our goal in this chapter is to explore some of the ways in which people's expectations of robots may be raised too high and therefore be vulnerable to disappointment.

To avert disappointed expectations, at least two paths of action are available. One is to rapidly expand robots' capacities, which is what most designers and engineers strive for. But progress has been slow [3], and the social and communicative skills of artificial agents are still far from what seems desirable [4], [5]. Another path is to ensure that people trust a robot

[1] The authors have provided a measure of these multiple dimensions of trust and invite readers to use that measure for their human-robot interaction studies: http://bit.ly/MDMT_Scale

to be just as reliable, capable, and ethical as it really is able to; that is, to ensure that people understand the robot's actual abilities and limitations. This path focuses on one aspect of *transparency*: providing human users with information about the capabilities of a system. Such transparency, we argue, is a precondition for justified trust in any autonomous machine, and social robots in particular [6], [7].

In this chapter, we describe some of the sources of discrepancies between people's expectations and robots' real capabilities. We argue the discrepancies are often caused by superficial properties of robots that elicit feelings of trust in humans without validly indicating the underlying property the person trusts in. We therefore need to understand the complex human responses triggered by the morphology and behaviour of autonomous machines, and we need to build a systematic understanding of the effects that specific design choices have on people's cognitive, emotional, and relational reactions to robots. In the second part of the chapter we lay out a number of ways to combat these discrepancies.

Discrepancies Between Human Expectations and Actual Robot Capabilities

In robot design and human-robot interaction research, the tendency to build ever more social cues into robots (from facial expressions to emotional tone of voice) is undeniable. Intuitively, this makes sense since robots that exhibit social cues are assumed to facilitate social interaction by leveraging people's existing social skill sets and experience, and they would fit seamlessly into social spaces without constantly being in the way [8]. However, in humans, the display of social cues is indicative of certain underlying mental properties, such as thoughts, emotions, intentions, or abilities. The problem is that robots can exhibit these same cues, through careful design or specific technologies, even though they do not have the same, or even similar, underlying properties.

For example, in human interaction, following another person's gaze is an invitation to joint attention [9]; and in communication, joint attention signals the listener's understanding of the speaker's communicative intention. Robots using such gaze cues [10] are similarly interpreted as indicating joint attention and of understanding a speaker's instructions [11], [12]. However, robots can produce these behaviors naïvely using simple algorithms, without having any concept of joint attention or any actual understanding of the speaker's communication. Thus, when a robot displays these social cues, they are not symptoms of the expected underlying processes, and a person observing this robot may erroneously attribute a range of (often human-like) properties to the robot [13].

Erroneous assumptions about other people are not always harmful. Higher expectations than initially warranted can aid human development (when caregivers “scaffold” the infant’s budding abilities; [14], can generate learning success [15], and can foster prosocial behaviors [16]. But such processes are, at least currently, wholly absent with robots. Overestimating a robot’s capacities poses manifest risks to users, developers, and the public at large. When users entrust a robot with tasks that the robot ends up not being equipped to do, people may be disappointed and frustrated when they discover the robot’s limited actual capabilities [17]; and there may be distress or harm if they discover these limitations too late. Likewise, developers who consistently oversell their products will be faced with increasing numbers of disappointed, frustrated, or distressed users who no longer use the product, write terrible public reviews (quite a significant impact factor for consumer technology), or even sue the manufacturer. Finally, the public at large could be deprived of genuine benefits if a few oversold robotic products cause serious harm, destroy consumer trust, and lead to stifling regulation.

Broadly speaking, discrepancies between expectations and reality have been well documented and explored under the umbrella of “expectancy violation,” from the domains of perception [18] to human interaction [19]. In human-robot interaction research, such violations have been studied, for example, by comparing expectations from media to interactions with a real robot [20] or by quantifying updated capability estimates after interacting with a robot [21]. Our discussion builds on this line of inquiry, but we do not focus on cases when an expectancy *violation* has occurred, which assumes that the person has become aware of the discrepancy (and is likely to lose trust in the robot). Instead, we focus on sources of such discrepancies and avenues for making a person aware of the robot’s limitations *before* they encounter a violation (and thus before a loss of trust).

Sources of Discrepancies

There are multiple sources of discrepancies between the perceived and actual capacities of a robot. Obvious sources are the entertainment industry and public media, which frequently exaggerate technical realities of robotic systems. We discuss here more psychological processes, from misleading and deceptive design and presentation to automatic inferences from a robot’s superficial behavior to deep underlying capabilities.

Misleading design

Equipping a robot with outward social cues that have no corresponding abilities is, at best, misleading. Such a strategy violates German designer Dieter Rams' concept of *honest design*, which is the commitment to design that "does not make a product more innovative, powerful or valuable than it really is" [22]; see also [23], [24]. Honest design is a commitment to transparency—enabling the user to "see through" the outward appearance and to accurately infer the robot's capacities. In the HRI laboratory, researchers often violate this commitment to transparency when they use Wizard-of-Oz (WoZ) methods to make participants believe that they are interacting with an autonomous, capable robot. Though such misperceptions are rarely harmful, they do contribute to false beliefs and overly high expectations about robots outside the laboratory. Moreover, thorough debriefing at the end of such experiments is not always provided [25], which would reset people's generalizations about technical realities.

Deception

When a mismatch between apparent and real capacities is specifically intended—for example, to sell the robot or impress the media—it arguably turns into deception and even exploitation [26]. And people are undoubtedly vulnerable to such exploitation. A recent study suggested that people were willing to unlock the door to a university dormitory building for a verbally communicating robot that had the seeming authority of a food delivery agent. Deception is not always objectionable; in some instances it is used for the benefit of the end user [27], [28], such as in calming individuals with dementia [29] or encouraging children on the autism spectrum to form social bonds [30]. However, these instances must involve careful management of the risks involved in the deception—risks for the individual user, the surrounding social community, and the precedent it sets for other, perhaps less justified cases of deception.

Impact of norms

At times, people are well aware that they are interacting with a machine in human-like ways because they are engaging with the robot in a joint pretense [31] or because it is the normatively correct way to behave. For example, if a robot greets a person, the appropriate response is to reciprocate the greeting; if the speaker asks a question, the appropriate response is to answer the question. Robots may not recognize the underlying social norm and they may not be insulted if the user violates the norm, but the user, and the surrounding community (e.g., children who are learning these norms), benefit from the fact that both parties uphold relevant social practices and thus a cooperative, respectful social order [32]. The more specific the roles that robots are assigned (e.g., nurse assistant, parking lot attendant), the more these norms and practices will influence people's behavior toward the robot [33]. If robots are equipped with the norms that apply to their roles (which is a significant challenge; [34], this may improve interaction quality and user satisfaction. Further, robots can actively leverage norms to shape how people interact with it, but perhaps even in manipulative fashion [35]. Norm-appropriate behavior is also inherently trust-building, because norms are commitments to act, and expectations that others will act, in ways that benefit the other (thus invoking the dimension of ethical trust; [36], norm violations become all the more powerful in threatening trust.

Expanded inferences

Whereas attributions of norm competence to a robot are well grounded in the robot's actual behavior, a robot that displays seemingly natural communicative skills can compel people to infer (and genuinely assume to be present) many other abilities that the robot probably is unlikely to have [37]. In particular, seeing that a robot has some higher-level abilities, people are likely to assume that it will also possess more basic abilities that in humans would be a prerequisite for the higher-level ability. For instance, a robot may greet someone with "Hi, how are you?" but be unable itself to answer the same question when the greeting is reciprocated, and it may not even have any speech understanding capabilities at all. Furthermore, a robot's syntactically correct sentences do not mean it has a full-blown semantics or grasps anything about conversational dynamics [38]. Likewise, seeing that a robot has one skill, we must expect people to assume that it also is has other skills that in humans are highly correlated with the first. For example, a robot may be able to entertain or even tutor a child but be unable to recognize when the child is choking on a toy. People find it hard to imagine that a being can have selected, isolated abilities that do not build upon each other [39].

Though it is desirable that, say, a manufacturer provides explicit and understandable documentation of a system's safety and performance parameters [40], [41], making explicit what a robot can and cannot do will often fail. That is because some displayed behaviors set off a cascade of inferences that people have evolved and practiced countless times with human beings [32]. As a result, spontaneous reactions to robots in social contexts and their explicit beliefs on what mental capacities robots possess can come apart [42], [43].

Automatic inferences

Some inferences or emotional responses are automatic, at least upon initial encounters with artificial agents. Previous research has shown that people treat computers and related technology (including robots) in some ways just like human beings (e.g., applying politeness and reciprocity), and often do so mindlessly [44]. The field of human-robot interaction has since identified numerous instances in which people show basic social-cognitive responses when responding to humanlike robots—for example, by following the "gaze" of a robot [45] or by taking its visual perspective [46]. Beyond such largely automatic reactions, a robot's humanlike appearance seems to invite a wide array of inferences about the robot's intelligence, autonomy, or mental capacities more generally [47]–[49]. But even if these appearance-to-mind inferences are automatic, they are not simplistic; they do not merely translate some degree of humanlikeness into a proportional degree of "having a mind." People represent both humanlike appearance and mental capacities along multiple dimensions [50]–[52], and specific dimensions of humanlike appearance trigger people's inferences for specific dimensions of mind. For example, features of the Body Manipulator dimension (e.g., torso, arms, fingers) elicit inferences about capacities of reality interaction, which include perception, learning, acting, and communicating. By contrast, facial and surface features (e.g., eyelashes, skin, apparel) elicit inferences about affective capacities, including feelings and basic emotions, as well as moral capacities, including telling right from wrong and upholding moral values [53].

Variations

We should note, however, that people's responses to robots are neither constant nor universal. They show variation within person, manifesting sometimes as cognitive, emotional, or social-relational reactions, can be in the foreground or background at different moments in time, and change with extended interactions with the robot [8], [32]. They also show substantial *inter*personal variation, as a function of levels of expertise [54], personal style [55], and psychosocial predispositions such as loneliness [56].

Status quo

The fact remains, however, that people are vulnerable to the impact of a robot's behavior and appearance [57]. We must expect that, in real life as in the laboratory, people will be willing to disclose negative personal information to humanoid agents [58], [59], trust and rely on them [60], empathize with them [61], [62], give in to a robot's obedience-like pressure to continue tedious work [63] or perform erroneous tasks [64]. Further, in comparison to a mechanical robot, people are more prone to take advice from a humanoid robot [65], trust and rely on them more [60], and are more likely to comply with their requests [66]. None of these behaviors are inherently faulty; but currently they are unjustified, because they are generated by superficial cues rather than by an underlying reality [57]. At present, neither mechanical nor humanoid robots have more knowledge to share than Wikipedia, are no more trustworthy to keep secrets than one's iPhone, and have no more needs or suffering than a cartoon character. They may in the future, but until that future, we have to ask how we can prevent people from having unrealistic expectations of robots, especially humanlike ones.

How to Combat Discrepancies

We have seen that discrepancies between perceived and actual capacities exist at multiple levels and are fed from numerous sources. How can people recover from these mismatches or avoid them in the first place? In this section, we provide potential paths for both short- and long-term solutions to the problem of expectation discrepancy when dealing with social robots.

Waiting for the future

An easy solution may be to simply wait for the robots of the future to make true the promises of the present. However, that would mean an extended time of misperceived reality, and numerous opportunities for misplaced trust, disappointment, and non-use. It is unclear whether recovery from such prolonged negative experiences is possible. Another strategy to overcome said discrepancies may be to encourage users to acquire minimally necessary technical knowledge to better evaluate artificial agents, perhaps encouraging children to program machines and thus see their mechanical and electronic insides. However, given the widespread disparities in access to quality education in most of the world's countries, the technical-knowledge path would leave poorer people misled, deceived, and more exploitable than ever before. Moreover, whereas the knowledge strategy would combat some of the sources we discussed (e.g., deception, expanded inferences), it would leave automatic inferences intact, as they are likely grounded in biologically or culturally evolved response patterns.

Experiencing the cold truth

Another strategy might be to practically force people to experience the mechanical and lifeless nature of machines—such as by asking people to inspect the skinless plastic insides of an animal robot like Paro or by unscrewing a robot's head and handing it to the person. It is, however, not clear that this will provide more clarity for human-robot interactions. A study of the effects of demonstrating the mechanistic nature of robots to children in fact showed that the children still interacted with the robot in the same social ways as children to whom the robotic side of robots had not been pointed out [67]. Furthermore, if people have already formed emotional attachments, such acts will be seen as cruel and distasteful, rather than have any corrective effects on discrepant perceptions.

Revealing real capacities

Perhaps most obvious would be truth in advertising. Robot designers and manufacturers, organizations and companies that deploy robots in hotel lobbies, hospitals, or school yards would signal to users what the robot can and cannot do. But there are numerous obstacles to designers and manufacturers offering responsible and modest explanations of the machine's real capacities. They are under pressure to produce within the constraints of their contracts; they are beholden to funders; they need to satisfy the curiosity of journalists and policy makers, who are also keen to present positive images of developing technologies.

Further, even if designers or manufacturers adequately reveal the machine's limited capabilities, human users may resist such information. If the information is in a manual, people won't read it. If it is offered during purchase, training, or first encounters, it may still be ineffective. That is because the abovementioned human tendency to perceive agency and mind in machines that have the tell-tale signs of self-propelled motion, eyes, and verbal communication is difficult to overcome. Given the eliciting power of these cues, it is questionable (though empirically testable) whether explicit information can ever counteract a user's inappropriate mental model of the machine.

Legibility and explainability

An alternative approach is to make the robot itself "legible"—something that a growing group of scholars is concerned with [68]. But whereas a robot's intentions and goals can be made legible—e.g., in a projection of the robot's intended motion path or in the motion itself—capabilities and other dispositions are not easily expressed in this way. At the same time, the robot can correct unrealistic expectations by indicating some of its *limits* of capability in failed actions [69] or, even more informative, in explicit statements that it is unable or forbidden to act a certain way [70].

A step further would be to design the robot in such a way that it can explicate its own actions, reasoning, and capabilities. But whereas giving users access to the robot's ongoing decision making and perhaps offering insightful and human-tailored explanations of its performed actions may be desirable [71], "explaining" one's *capacities* is highly unusual. Most of this kind of communication among humans is done indirectly, by providing information about, say, one's occupation [72] or acquaintance with a place [73]. Understanding such indirect speech requires access to shared perceptions, background knowledge, and acquired common ground that humans typically do not have with robots. Moreover, a robot's attempts to communicate its knowledge, skills, and limitations can also disrupt an ongoing activity or even backfire if talk about capabilities makes users suspect that there is a problem with the interaction [32]. There is, however, a context in which talk about capabilities is natural—educational settings. Here, one agent learns new knowledge, skills, abilities, often from another agent, and both might comment freely on the learner's capabilities already in place, others still developing, and yet others clearly absent. If we consider a robot an ever-learning agent, then perhaps talk about capabilities and limitations can be rather natural.

One potential drawback of robots that explain themselves must be mentioned. Such robots would appear extremely sophisticated, and one might then worry which other capacities people will infer from this explanatory prowess. Detailed insights into reasoning may invite inferences of deeper self-awareness, even wisdom, and user-tailored explanations may invite inferences of caring and understanding of the user's needs. But perhaps by the time full-blown explainability can really be implemented, some of these other capacities will too; then the discrepancies would all lift at once.

Managing expectations

But until that time, we are better off with a strategy of managing expectations and ensuring performance that matches these expectations and lets trust build upon solid evidence. Managing expectations will rely on some of the legibility and explainability strategies just mentioned along with attempts to explicitly set expectations low, which may be easily exceeded to positive effect [74]. However, such explicit strategies would be unlikely to keep automatic inferences in check. For example, in one study, Zhao et al. (submitted) showed that people take a highly humanlike robot's visual perspective even when they are told it is a wax figure. The power of the mere humanlike appearance was enough to trigger the basic social-cognitive act of perspective taking.

Thus, we also need something we might call *restrained design*—attempts to avoid overpromising signals in behavior, communication, and appearance, as well as limiting the robot's roles so that people form limited, role- and context-adequate expectations. As a special case of such an approach we describe here the possible benefit of an incremental robot design strategy—the commitment to advance robot capacities in small steps, each of which is well grounded in user studies and reliability testing.

Incremental Design

Why would designing and implementing small changes in a robot prevent discrepancies between a person's understanding of the robot's capacities and its actual capacities? Well-designed small changes may be barely noticeable and, unless in a known, significant dimension (e.g., having eyes after never having had eyes), will limit the number of new inferences that would be elicited by it. Further, even when noticed, the user may be able to more easily adapt to a small change, and integrate it into their existing knowledge and understanding of the robot, without having to alter their entire mental model of the robot.

Consider the iRobot Roomba robotic vacuum cleaner. The Roomba has a well-defined, functional role in households as a cleaning appliance. From its first iteration, any discrepancy between people's perceptions of the robot's capacities and its actual capacities were likely related to the robot's cleaning abilities, which could be quickly resolved by using the robot in practice. As new models hit the market, Roomba's functional capacities improved only incrementally—for example, beep-sequence error codes were replaced by pre-recorded verbal announcements, or random-walk cleaning modes were replaced by rudimentary mapping technology. In these cases, the human users have to accommodate only minor novel elements in their mental models, each changing only very few parameters.

Consider, by contrast, Softbank's Pepper robot. From the original version, Pepper was equipped with a humanoid form including arms and hands that appeared to gesture, and a head with eyes and an actuated neck, such that it appeared to look at and follow people. Further, marketing material emphasized the robot's emotional capacities, using such terms as "perception modules" and an "emotional engine." We can expect that these features encourage people to infer complex capacities in this robot, even beyond perception and emotion. Observing the robot seemingly gaze at us and follow a person's movements suggests attention and interest; the promise of emotional capacities suggests sympathy and understanding. However, beyond pre-coded sentences intended to be cute or funny, the robot currently has no internal programmed emotional model at all. As a result, we expect there to be large discrepancies between a person's elicited expectations and the robot's actual abilities. Assumptions of deep understanding in conversation and willingness toward risky personal disclosure may then be followed by likely frustration or disappointment.

The discrepancy in Pepper's case stems in part from the jump in expectation that the designers invite the human to take and the actual reality of Pepper's abilities. Compared with other technologies people may be familiar with, a highly humanoid appearance, human-like social signaling behaviors, and purported emotional abilities trigger a leap in inference people make from "robots can't do much" to "they can do a lot." But that leap is not matched by Pepper's actual capabilities. As a result, encountering Pepper creates a large discrepancy that will be quite difficult to overcome. A more incremental approach would curtail the humanoid form and focus on the robot's gaze-following abilities, without claims of emotional processing. If the gaze following behavior actually supports successful person recognition and communication turn taking, then a more humanoid form may be warranted. And only if actual emotion recognition and the functional equivalent of emotional states in the robot are achieved would Pepper's "emotion engine" be promoted.

Incremental approaches have been implemented in other technological fields. For example, commercial car products have in recent years increasingly included small technical changes that point toward eventual autonomous driving abilities, such as cruise control, active automatic breaking systems, lane violation detection and correction, and the like. More advanced cars, such as Tesla's Model S, have an "auto-pilot" mode that takes a further step toward autonomous driving in currently highly constrained circumstances. The system still frequently reminds the user to keep their hands on the steering wheel and to take over when those constrained circumstances no longer hold (e.g., no painted lane information). However, the success of this shared autonomy situation depends on how a product is marketed. Other recent cars may include a great deal of autonomy in their onboard computing system but are not marketed as autonomous or self-driving but are called "Traffic Jam Assist" or "Super Cruise." Such labeling decisions limit what the human users expects of the car and therefore what they entrust it to do. A recent study confirms that labeling matters: People overestimate Tesla cars' capacities more than other comparable brands [75]. And perhaps unsurprisingly, the few highly-publicized accidents with Teslas are typically the result of vast overestimation of what the car can do [76], [77].

Within self-driving vehicle research and development, a category system is in place to express the gradually increasing levels of autonomy of the system in question. In this space, however, the incremental approach may still take steps that are too big. In the case of vehicle control, people's adjustment to continuously increasing autonomy is not itself continuous but takes a qualitative leap. People either drive themselves, assisted up to a point, or they let someone else (or something else) drive; they become *passengers*. In regular cars, actual passengers give up control, take naps, read books, chat on the phone, and would not be ready to instantly take the wheel when the main driver requests it. Once people take on the unengaged passenger role with increasingly (but not yet fully) autonomous vehicles, the situation will result in over-trust (the human will take naps, read books, etc.). And if there remains a small chance that the car needs the driver's attention but the driver has slipped into the passenger role, the situation could prove catastrophic. The human would not be able to take the wheel quickly enough when the car requests it because it takes time for a human to shift attention, observe their surroundings, develop situational awareness, make a plan, and act [78]. Thus, even an incremental approach would not be able to avert the human's jump to believing the car can handle virtually all situations, when in fact the car cannot.

Aside from incremental strategies, the more general restrained design approach must ultimately be *evidence-based* design. Decisions about form and function must be informed by evidence into which of the robot's signals elicit what expectations in the human. Such insights are still rather sparse and often highly specific to certain robots. It therefore takes a serious research agenda to address this challenge, with a full arsenal of scientific approaches: carefully controlled experiments to establish causal relations between robot characteristics and a person's expectations; examination of the stability of these response patterns by comparing young children and adults as well as people from different cultures; and longitudinal studies to establish how those responses will change or stabilize in the wake of interacting with robots over time. We close our analysis by discussing the strengths and challenges that come with longitudinal studies.

Longitudinal Research

Longitudinal studies would be the ideal data source to elucidate the source of and remedy for discrepancies between perceived and actual robot capacities. That is because, first, they can distinguish between initial reactions to robots and more enduring response patterns. We have learned from human-human social perception research that initial responses, even if they change over time, can strongly influence the range of possible long-term responses; in particular, initial negative responses tend to improve more slowly than positive initial reactions deteriorate [79]. In human-robot encounters, some responses may be automatic and have a lasting impact, whereas others may initially be automatic but could be changeable over time. Furthermore, some responses may reflect an initial lack of understanding of the encountered novel agent, and with time a search for meaning may improve this understanding [80]. Longitudinal studies can also track how expectations clash with new observations and how trust fluctuates as a result.

High-quality longitudinal research is undoubtedly difficult to conduct because of cost, time and management commitments, participant attrition, ethical concerns of privacy and unforeseen impacts on daily living, and the high rate of mechanical robot failures. A somewhat more modest goal might be to study short-term temporal dynamics that will advance knowledge but also provide a launching pad for genuine longitudinal research. For the question of recovery from expectation-reality discrepancies we can focus on a few feasible but informative paradigms.

A first paradigm is to measure people's responses to a robot with or without information about the true capacities of the robot. In comparison to spontaneous inferences about the robot's capacities, would people adjust their inferences when given credible information? One could compare the differential effectiveness of (a) inoculation (providing the ground-truth information before the encounter with the robot) and (b) correction (providing it after the encounter). In human persuasion research, inoculation is successful when the persuasive attempt operates at an explicit, rational level [81]. By analogy, the comparison of inoculation and post-hoc correction in the human-robot perception case may help clarify which human responses to robots lie at the more explicit and which at the more implicit level.

A second paradigm is to present the robot twice during a single experimental session, separated by some time delay or unrelated other activities. What happens to people's representations formed in the first encounter that are either confirmed or disconfirmed in the second encounter? If the initial reactions are mere novelty effects, they would subside independent of the new information; if they are deeply entrenched, they would remain even after disconfirmation; and if they are systematically responsive to evidence, they would stay the same under confirmation and change under disconfirmation [82]. In addition, different response dimensions may behave differently. Beliefs about the robot's reliability and competence may change more rapidly whereas beliefs about its benevolence may be more stable.

In a third paradigm, repeated-encounter but short-term experiments could bring participants back to the laboratory more than once. Such studies could distinguish people’s adjustments to specific robots (if they encounter the same robot again) from adjustments of their general beliefs about robots (if they encounter a different, but comparable robot again). From stereotype research, we have learned that people often maintain general beliefs about a social category even when acquiring stereotype-disconfirming information about specific individuals [83]. Likewise, people may update their beliefs about a specific robot they encounter repeatedly without changing their beliefs about robots in general [82].

Conclusion

Trust is one agent’s expectation about the other’s actions. Trust is broken when the other does not act as one expected—is not as reliable or competent as one expected, or is dishonest or unethical. In all these cases, a discrepancy emerges between what one agent expected and the other agent delivered. Human-robot interactions, we suggest, often exemplifies such cases: people expect more of their robots than the robots can deliver. Such discrepancies have many sources, from misleading and deceptive information to the seemingly innocuous but powerful presence of deep-seated social signals. This range of sources demands a range of remedies, and we explored several of them, from patience to legibility, from incremental design to longitudinal research. Because of people’s complex responses to artificial agents, there is no optimal recipe for minimizing discrepancies and maximizing trust. We can only advance our understanding of those complex human responses to robots, use this understanding to guide robot design, and monitor how improved design and human adaptation, over time, foster more calibrated and trust-building human-robot interactions.

References

[1] D. Ullman and B. F. Malle, “What does it mean to trust a robot? Steps toward a multidimensional measure of trust,” in *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, New York, NY, USA: ACM, 2018, pp. 263–264.

[2] D. Ullman and B. F. Malle, “Measuring gains and losses in human-robot trust: Evidence for differentiable components of trust.,” in *Companion to the 2019 ACM/IEEE International Conference on Human-Robot Interaction, HRI ’19.*, New York, NY: ACM, 2019, pp. 618–619.

[3] L. Lewis and S. Shrikanth, “Japan lays bare the limitations of robots in unpredictable work,” *Financial Times*, 25-Apr-2019. [Online]. Available: https://www.ft.com/content/beece6b8-4b1a-11e9-bde6-79eaea5acb64. [Accessed: 04-Jan-2020].

[4] R. K. Moore, “Spoken language processing: Where do we go from here,” in *Your Virtual Butler: The Making-of*, R. Trappl, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 119–133.

[5] M.-A. Williams, “Robot social intelligence,” in *Social Robotics*, S. S. Ge, O. Khatib, J.-J. Cabibihan, R. Simmons, and M.-A. Williams, Eds. Springer Berlin Heidelberg, 2012, pp. 45–55.

[6] K. Fischer, H. M. Weigelin, and L. Bodenhagen, “Increasing trust in human–robot medical interactions: effects of transparency and adaptability,” *Paladyn, Journal of Behavioral Robotics*, vol. 9, no. 1, pp. 95–109, 2018, doi: 10.1515/pjbr-2018-0007.

[7] T. L. Sanders, T. Wixon, K. E. Schafer, J. Y. Chen, and P. Hancock, “The influence of modality and transparency on trust in human-robot interaction,” presented at the Cognitive Methods in Situation Awareness and Decision Support (CogSIMA), 2014 IEEE International Inter-Disciplinary Conference on, 2014, pp. 156–159.

[8] K. Fischer, “Tracking anthromorphizing behavior in human-robot interaction,” *Manuscript submitted for publication*, 2020.

[9] N. Eilan, C. Hoerl, T. McCormack, and J. Roessler, Eds., *Joint attention: Communication and other minds*. New York, NY: Oxford University Press, 2005.

[10] B. Mutlu, T. Kanda, J. Forlizzi, J. Hodgins, and H. Ishiguro, “Conversational gaze mechanisms for humanlike robots,” *ACM Trans. Interact. Intell. Syst.*, vol. 1, no. 2, pp. 12:1–12:33, Jan. 2012, doi: 10.1145/2070719.2070725.

[11] K. Fischer, K. Lohan, J. Saunders, C. Nehaniv, B. Wrede, and K. Rohlfing, “The impact of the contingency of robot feedback on HRI,” in *2013 International Conference on Collaboration Technologies and Systems (CTS)*, IEEE, 2013, pp. 210–217.

[12] K. Fischer, K. Foth, K. Rohlfing, and B. Wrede, “Mindful tutors – linguistic choice and action demonstration in speech to infants and to a simulated robot,” *Interaction Studies*, vol. 12, no. 1, pp. 134–161, 2011.

[13] M. Kwon, M. F. Jung, and R. A. Knepper, "Human expectations of social robots," in *Proceedings of the Eleventh ACM/IEEE International Conference on Human Robot Interaction, HRI'16*, Piscataway, NJ, 2016, pp. 463–464.

[14] R. Mermelshtine, "Parent–child learning interactions: A review of the literature on scaffolding," *British Journal of Educational Psychology*, vol. 87, no. 2, pp. 241–254, Jun. 2017, doi: 10.1111/bjep.12147.

[15] L. Jussim, S. L. Robustelli, and T. R. Cain, "Teacher expectations and self-fulfilling prophecies," in *Handbook of motivation at school.*, K. R. Wenzel and A. Wigfield, Eds. New York, NY: Routledge/Taylor & Francis Group, 2009, pp. 349–380.

[16] R. E. Kraut, "Effects of social labeling on giving to charity," *Journal of Experimental Social Psychology*, vol. 9, no. 6, pp. 551–562, Nov. 1973, doi: 10.1016/0022-1031(73)90037-1.

[17] M. de Graaf, S. B. Allouch, and J. van Dijk, "Why do they refuse to use my robot?: Reasons for non-use derived from a long-term home study," in *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, 2017, pp. 224–233.

[18] M. A. Bobes, M. Valdessosa, and E. Olivares, "An ERP study of expectancy violation in face perception," *Brain and Cognition*, vol. 26, no. 1, pp. 1–22, Sep. 1994, doi: 10.1006/brcg.1994.1039.

[19] J. K. Burgoon, D. A. Newton, J. B. Walther, and E. J. Baesler, "Non-verbal expectancy violations," *Journal of Nonverbal Behavior*, vol. 55, no. 1, pp. 58–79, 1989.

[20] U. Bruckenberger, A. Weiss, N. Mirnig, E. Strasser, S. Stadler, and M. Tscheligi, "The good, the bad, the weird: Audience evaluation of a 'real' robot in relation to science fiction and mass media," *ICSR 2013*, vol. 8239 LNAI, pp. 301–310, 2013, doi: 10.1007/978-3-319-02675-6_30.

[21] T. Komatsu, R. Kurosawa, and S. Yamada, "How does the difference between users' expectations and perceptions about a robotic agent affect their behavior?," *Int J of Soc Robotics*, vol. 4, no. 2, pp. 109–116, Apr. 2012, doi: 10.1007/s12369-011-0122-y.

[22] Vitsoe, "The power of good design," 2018. [Online]. Available: https://www.vitsoe.com/us/about/good-design. [Accessed: 22-Oct-2018].

[23] G. Donelli, "Good design is honest," 13-Mar-2015. [Online]. Available: https://blog.astropad.com/good-design-is-honest/. [Accessed: 22-Oct-2018].

[24] C. de Jong, Ed., *Ten principles for good design: Dieter Rams*. New York, NY: Prestel Publishing, 2017.

[25] D. J. Rea, D. Geiskkovitch, and J. E. Young, "Wizard of awwws: Exploring psychological impact on the researchers in social HRI experiments," in *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, New York, NY, USA: Association for Computing Machinery, 2017, pp. 21–29.

[26] W. C. Redding, "Ethics and the study of organizational communication: When will we wake up? In Jaksa, J.A., Pritchard, M.S. (eds.) (pp.). Hampton, Cresskill," in *Responsible Communication: Ethical Issues in*

Business, Industry, and the Professions, J. A. Jaksa and M. S. Pritchard, Eds. Cresskill, NJ: Hampton Press, 1996, pp. 17–40.

[27] E. C. Collins, “Vulnerable users: deceptive robotics,” *Connection Science*, vol. 29, no. 3, pp. 223–229, Jul. 2017, doi: 10.1080/09540091.2016.1274959.

[28] A. Matthias, “Robot lies in health care: When is deception morally permissible?,” *Kennedy Inst Ethics J*, vol. 25, no. 2, pp. 169–192, Jun. 2015, doi: 10.1353/ken.2015.0007.

[29] K. Wada, T. Shibata, T. Saito, and K. Tanie, “Effects of robot-assisted activity for elderly people and nurses at a day service center,” *Proceedings of the IEEE*, vol. 92, no. 11, pp. 1780–1788, Nov. 2004, doi: 10.1109/JPROC.2004.835378.

[30] E. Karakosta, K. Dautenhahn, D. S. Syrdal, L. J. Wood, and B. Robins, “Using the humanoid robot Kaspar in a Greek school environment to support children with Autism Spectrum Condition,” *Paladyn, Journal of Behavioral Robotics*, vol. 10, no. 1, pp. 298–317, Jan. 2019, doi: 10.1515/pjbr-2019-0021.

[31] H. H. Clark, “How do real people communicate with virtual partners?,” presented at the Proceedings of AAAI-99 Fall Symposium, Psychological Models of Communication in Collaborative Systems, November 5-7th, 1999, North Falmouth, MA., 1999.

[32] K. Fischer, *Designing speech for a recipient. partner modeling, alignment and feedback in so-called “simplified registers.”* Amsterdam: John Benjamins, 2016.

[33] J. Goetz, S. Kiesler, and A. Powers, “Matching robot appearance and behavior to tasks to improve human-robot cooperation,” in *The 12th IEEE International Workshop on Robot and Human Interactive Communication*, vol. 19, New York, NY: Association for Computing Machinery, 2003, pp. 55–60.

[34] B. F. Malle, P. Bello, and M. Scheutz, “Requirements for an artificial agent with norm competence,” in *Proceedings of 2nd ACM conference on AI and Ethics (AIES’19)*, New York, NY: ACM, 2019.

[35] E. Sanoubari, S. H. Seo, D. Garcha, J. E. Young, and V. Loureiro-Rodríguez, “Good robot design or Machiavellian? An in-the-wild robot leveraging minimal knowledge of passersby’s culture,” in *Proceedings of the 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, New York, NY, USA: Association for Computing Machinery, 2019, pp. 382–391.

[36] B. F. Malle and D. Ullman, “A multi-dimensional conception and measure of human-robot trust,” in *Trust in human-robot interaction: Research and applications*, C. S. Nam and J. B. Lyons, Eds. Elsevier, 2020.

[37] K. Fischer and R. Moratz, “From communicative strategies to cognitive modelling,” presented at the First International Workshop on `Epigenetic Robotics’, September 17-18, 2001, Lund, Sweden, 2001.

[38] S. Payr, “Towards human-robot interaction ethics.,” in *A Construction Manual for Robots’ Ethical Systems: Requirements, Methods,*

Implementations, R. Trappl, Ed. Cham, Switzerland: Springer International, 2015, pp. 31–62.

[39] K. Fischer, *What computer talk is and isn't: Human-computer conversation as intercultural communication*. Saarbrücken: AQ-Verlag, 2006.

[40] K. R. Fleischmann and W. A. Wallace, "A covenant with transparency: Opening the black box of models," *Communications of the ACM - Adaptive complex enterprises*, vol. 48, no. 5, pp. 93–97, 2005, doi: 10.1145/1060710.1060715.

[41] M. Hind *et al.*, "Increasing trust in AI services through supplier's declarations of conformity," *ArXiv e-prints*, Aug. 2018.

[42] S. R. Fussell, S. Kiesler, L. D. Setlock, and V. Yew, "How people anthropomorphize robots," in *Proceedings of the 3rd ACM/IEEE International Conference on Human Robot Interaction, HRI '08*, New York, NY, USA: Association for Computing Machinery, 2008, pp. 145–152.

[43] J. Złotowski, H. Sumioka, S. Nishio, D. F. Glas, C. Bartneck, and H. Ishiguro, "Appearance of a robot affects the impact of its behaviour on perceived trustworthiness and empathy," *Paladyn, Journal of Behavioral Robotics*, vol. 7, no. 1, 2016, doi: 10.1515/pjbr-2016-0005.

[44] C. Nass and Y. Moon, "Machines and mindlessness: Social responses to computers," *Journal of Social Issues*, vol. 56, no. 1, pp. 81–103, Jan. 2000, doi: 10.1111/0022-4537.00153.

[45] H. Admoni and B. Scassellati, "Social eye gaze in human-robot interaction," *Journal of Human-Robot Interaction*, vol. 6, no. 1, pp. 25–63, May 2017, doi: 10.5898/JHRI.6.1.Admoni.

[46] X. Zhao, C. Cusimano, and B. F. Malle, "Do people spontaneously take a robot's visual perspective?," in *Proceedings of the Eleventh ACM/IEEE International Conference on Human Robot Interaction, HRI'16*, Piscataway, NJ: IEEE Press, 2016, pp. 335–342.

[47] C. Bartneck, T. Kanda, O. Mubin, and A. Al Mahmud, "Does the design of a robot influence its animacy and perceived intelligence?," *International Journal of Social Robotics*, vol. 1, no. 2, pp. 195–204, Feb. 2009, doi: 10.1007/s12369-009-0013-7.

[48] E. Broadbent *et al.*, "Robots with display screens: A robot with a more humanlike face display is perceived to have more mind and a better personality," *PLoS ONE*, vol. 8, no. 8, p. e72589, Aug. 2013, doi: 10.1371/journal.pone.0072589.

[49] F. Eyssel, D. Kuchenbrandt, S. Bobinger, L. de Ruiter, and F. Hegel, "'If you sound like me, you must be more human': On the interplay of robot and user features on human-robot acceptance and anthropomorphism," in *Proceedings of the 7th ACM/IEEE International Conference on Human-Robot Interaction, HRI'12*, New York, NY: Association for Computing Machinery, 2012, pp. 125–126.

[50] B. F. Malle, "How many dimensions of mind perception really are there?," in *Proceedings of the 41st Annual Meeting of the Cognitive Science Society*,

E. K. Goel, C. M. Seifert, and C. Freksa, Eds. Montreal, Canada: Cognitive Science Society, 2019, pp. 2268–2274.

[51] E. Phillips, D. Ullman, M. de Graaf, and B. F. Malle, “What does a robot look like?: A multi-site examination of user expectations about robot appearance.,” in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting.*, 2017.

[52] E. Phillips, X. Zhao, D. Ullman, and B. F. Malle, “What is human-like? Decomposing robots’ human-like appearance using the Anthropomorphic roBOT (ABOT) database,” in *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, New York, NY, USA: ACM, 2018, pp. 105–113.

[53] X. Zhao, E. Phillips, and B. F. Malle, “How people infer a humanlike mind from a robot body,” PsyArXiv, preprint, Nov. 2019.

[54] K. Fischer, “Interpersonal variation in understanding robots as social actors,” in *Proceedings of \em HRI’11, March 6-9th, 2011. Lausanne, Switzerland*, 2011, pp. 53–60.

[55] S. Payr, “Virtual butlers and real people: Styles and practices in long-term use of a companion,” in *Your virtual butler: The making-of*, R. Trappl, Ed. Berlin, Heidelberg: Springer, 2013, pp. 134–178.

[56] K. M. Lee, Y. Jung, J. Kim, and S. R. Kim, “Are physically embodied social agents better than disembodied social agents?: The effects of physical embodiment, tactile interaction, and people’s loneliness in human–robot interaction,” *International Journal of Human-Computer Studies*, vol. 64, no. 10, pp. 962–973, Oct. 2006, doi: 10.1016/j.ijhcs.2006.05.002.

[57] K. S. Haring, K. Watanabe, M. Velonaki, C. C. Tossell, and V. Finomore, “FFAB—The form function attribution bias in human–robot interaction,” *IEEE Transactions on Cognitive and Developmental Systems*, vol. 10, no. 4, pp. 843–851, Dec. 2018, doi: 10.1109/TCDS.2018.2851569.

[58] G. M. Lucas, J. Gratch, A. King, and L.-P. Morency, “It’s only a computer: Virtual humans increase willingness to disclose,” *Computers in Human Behavior*, vol. 37, pp. 94–100, Aug. 2014, doi: 10.1016/j.chb.2014.04.043.

[59] T. Uchida, H. Takahashi, M. Ban, J. Shimaya, Y. Yoshikawa, and H. Ishiguro, “A robot counseling system — What kinds of topics do we prefer to disclose to robots?,” in *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2017, pp. 207–212.

[60] R. Pak, N. Fink, M. Price, B. Bass, and L. Sturre, “Decision support aids with anthropomorphic characteristics influence trust and performance in younger and older adults,” *Ergonomics*, vol. 55, no. 9, pp. 1059–1072, Sep. 2012, doi: 10.1080/00140139.2012.691554.

[61] L. D. Riek, T. Rabinowitch, B. Chakrabarti, and P. Robinson, “Empathizing with robots: Fellow feeling along the anthropomorphic spectrum,” in *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, 2009, pp. 1–6.

[62] S. H. Seo, D. Geiskkovitch, M. Nakane, C. King, and J. E. Young, “Poor thing! Would you feel sorry for a simulated robot? A comparison of

empathy toward a physical and a simulated robot," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, New York, NY, USA: Association for Computing Machinery, 2015, pp. 125–132.

[63] D. Y. Geiskkovitch, D. Cormier, S. H. Seo, and J. E. Young, "Please continue, we need more data: An exploration of obedience to robots," *Journal of Human-Robot Interaction*, vol. 5, no. 1, pp. 82–99, 2016, doi: 10.5898/10.5898/JHRI.5.1.Geiskkovitch.

[64] M. Salem, G. Lakatos, F. Amirabdollahian, and K. Dautenhahn, "Would you trust a (faulty) robot? Effects of error, task type and personality on human-robot cooperation and trust," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction, HRI '15*, New York: ACM, 2015, pp. 141–148.

[65] A. Powers and S. Kiesler, "The advisor robot: Tracing people's mental model from a robot's physical attributes," in *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-robot Interaction*, New York, NY, USA: ACM, 2006, pp. 218–225.

[66] V. Chidambaram, Y.-H. Chiang, and B. Mutlu, "Designing persuasive robots: How robots might persuade people using vocal and nonverbal cues," in *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction (HRI '12)*, New York, NY, USA: Association for Computing Machinery, 2012, pp. 293–300.

[67] S. Turkle, C. Breazeal, O. Dasté, and B. Scassellati, "First encounters with Kismet and Cog: Children respond to relational artifacts," in *Digital media: Transformations in human communication*, P. Messaris and L. Humphreys, Eds. New York, NY: Peter Lang, 2006, pp. 313–330.

[68] C. Lichtenthäler and A. Kirsch, *Legibility of robot behavior: A literature review.* https://hal.archives-ouvertes.fr/hal-01306977, 2016.

[69] M. Kwon, S. H. Huang, and A. D. Dragan, "Expressing robot incapability," in *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction (HRI '18)*, New York, NY, USA: Association for Computing Machinery, 2018, pp. 87–95.

[70] G. Briggs and M. Scheutz, "'Sorry, I can't do that:' Developing mechanisms to appropriately reject directives in human-robot interactions," in *Proceedings of the 2015 AAAI Fall Symposium on AI and HRI*, 2015.

[71] M. de Graaf and B. F. Malle, "How people explain action (and autonomous intelligent systems should too)," in *2017 AAAI Fall Symposium Series Technical Reports*, Palo Alto, CA: AAAI Press, 2017, pp. 19–26.

[72] H. H. Clark, "Communal lexicons," in *Context in Language Learning and Language Understanding*, v, 198 vols., K. Malmkjær and J. Williams, Eds. Cambridge University Press, 1998, pp. 63–87.

[73] E. A. Schegloff, "Notes on a conversational practise: formulating place," in *Studies in Social Interaction*, D. Sudnow, Ed. New York: Free Press, 1972, pp. 75–119.

[74] S. Paepcke and L. Takayama, "Judging a bot by its cover: An experiment on expectation setting for personal robots," in *2010 5th ACM/IEEE*

International Conference on Human-Robot Interaction (HRI), New York, NY: Association for Computing Machinery, 2010, pp. 45–52.

[75] E. R. Teoh, "What's in a name? Drivers' perceptions of the use of five SAE Level 2 driving automation systems," *Journal of Safety Research*, 2020.

[76] J. Bhuiyan, "A federal agency says an overreliance on Tesla's Autopilot contributed to a fatal crash," *Vox*, 12-Sep-2017. [Online]. Available: https://www.vox.com/2017/9/12/16294510/fatal-tesla-crash-self-driving-elon-musk-autopilot. [Accessed: 05-Jan-2020].

[77] F. Lambert, "Tesla driver was eating and drinking during publicized Autopilot crash, NTSB reports," *Electrek*, 03-Sep-2019. [Online]. Available: https://electrek.co/2019/09/03/tesla-driver-autopilot-crash-eating-ntsb-report/. [Accessed: 05-Jan-2020].

[78] M. A. Regan, C. Hallett, and C. P. Gordon, "Driver distraction and driver inattention: Definition, relationship and taxonomy," *ACCIDENT ANALYSIS AND PREVENTION*, vol. 43, no. 5, pp. 1771–1781, Sep. 2011, doi: 10.1016/j.aap.2011.04.008.

[79] M. Rothbart and B. Park, "On the confirmability and disconfirmability of trait concepts," *Journal of Personality and Social Psychology*, vol. 50, pp. 131–142, 1986.

[80] C. V. Smedegaard, "Reframing the Role of Novelty within Social HRI: from Noise to Information," *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 411–420, 2019, doi: 10.1109/HRI.2019.8673219.

[81] W. J. McGuire, "Inducing resistance to persuasion: Some contemporary approaches," in *Advances in Experimental Social Psychology*, vol. 1, L. Berkowitz, Ed. Academic Press, 1964, pp. 191–229.

[82] M. J. Ferguson, M. Kwon, T. Mann, and R. A. Knepper, "The formation and updating of implicit impressions of robots.," presented at the The Annual Meeting of the Society for Experimental Social Psychology, Toronto, Canada, 2019.

[83] M. Rothbart and M. Taylor, "Category labels and social reality: Do we view social categories as natural kinds?," in *Language, interaction and social cognition*, Thousand Oaks, CA, US: Sage Publications, Inc, 1992, pp. 11–36.