Exploring the Situated Vulnerabilities of Robots for Interpersonal Trust in Human-Robot Interaction

Glenda Hannibal, Astrid Weiss

Abstract

The practical value of studying trust in human-robot interaction (HRI) rests on the assumption that people will, in the long-term, accept, interact, and collaborate more with robots that they trust or consider trustworthy. We propose in this book chapter to take our *event approach* to interpersonal trust in HRI and we argue why focusing on robot vulnerabilities will benefit current discussions on trust in robots and their perceived trustworthiness. On a theoretical level, we first argue that it is important to challenge the often negative view of the conceptual relationship between interpersonal trust and vulnerability in HRI as it has mainly comes to represent overexposure. Moreover, identifying robot-specific vulnerabilities is essential when exploring interpersonal trust in interactions between humans and robots (or HRI) because it overlaps but is not identical to those important to a human-centered perspective. To empirically explore robot vulnerabilities, we present the results of eight semi-structured expert interviews with experienced leaders in robotics. Based on these interviews, we identify the various robot vulnerabilities mentioned by the experts to present a systematic overview. Furthermore, we discuss how the experts interpreted the notion of vulnerability in relation to robots specifically and dive more into how malicious human behavior can be problematic when aiming to ensure mutual interpersonal trust in HRI. Moreover, we aim in this book chapter to lay down our motivation and arguments for why taking into account robot vulnerabilities provide a crucial and broader perspective on mutual trust in HRI, which is fundamental to strengthening interaction, collaboration and engagement between humans and robots.

Keywords

human-robot interaction, interpersonal trust, event approach, vulnerability, expert interviews, ethics, engagement

1 Introduction

The most common trust relation people have with artifacts and technologies is best described in terms of reliance and understood as a certain form of dependency. This dependency assumes that reliance on an inanimate object is necessary for the successful realization of some kind of plan given specific goals. Viewed as plan execution, trust as reliance mainly gets its value because of its ability to guide thoughts and actions from the perspective that seems reasonable given the means adopted to meet the concrete ends [Smith 2010; Alonso 2014]. Consequently, trust as reliance cannot be understood solely as something internal to the person trusting since it also depends on external conditions, which are laws of nature and the constraints of the specific design. Therefore, the main focus of trust as reliance is placed on making the interactions as smooth, efficient, and comfortable as possible in which artifacts or technologies are only to be considered instruments or tools to help people achieve their goals.

This instrumental view is the most traditional and widespread understanding of inanimate objects and also guides current understandings of robots [Coeckelbergh 2010a]. In robotics, *trust as reliance* is taken to mean that a person holds a predictive belief or assumption related to the performance of the robot given

https://doi.org/10.34727/2022/isbn.978-3-85448-052-5_2

the intended purpose and the predefined task. The performance of the robot then determines its trustworthiness and is considered important as it helps in establishing whether or not people are justified in trusting the robot. From this perspective, the robot's performance ensures that people can strike the right level of trust during interactions, collaboration, or engagement. Thus, an appropriate level of trust is treated as an indirect measure, which is later used to suggest specific design guidelines to prevent either under- or over-reliance [De Visser et al. 2020; Kok and Soh 2020].

The instrumental view has been significantly challenged with the recent aim in social robotics to make robots more socially capable and human-like (in regards to both physical appearance and style of behavior). Drawing on computational models of human cognition and social competence, "socially intelligent robots" [Breazeal 2001; Dautenhahn 1995] have built-in capacities to recognize and display cues for social interaction and communication. As such, they can behave and respond to people in a way they might interpret as intentional, influencing how people approach and treat robots. Similarly, endowing robots with anthropomorphic features only amplifies the tendency to perceive them as more human-like and is used as a deliberate design strategy to facilitate human-robot interaction (HRI) [Złotowski et al. 2016]. However, taking seriously human perception of robots as more socially capable and human-like also means that the current conceptualization of trust as reliance for HRI is no longer sufficient because it does not capture the additional social dimension of such interactions, which also extend to more ethical issues [Malle and Ullman 2021; Nyholm 2020]. Recent work on trust in HRI has attempted to adopt the notion of *interpersonal trust* to better study trust between humans and robots and uses this conceptualization as an explicit framework for the development of trustworthy robots [Lee et al. 2013; Ogawa et al. 2019].

In HRI research, speaking about interpersonal trust is taken to be unproblematic, and its meaning is connected to the observation that people seem to trust in robots and consider them trustworthy because of the assumed motives or intentions underlying their performance or actions due to their apparent agency. In such instances, speaking about interpersonal trust in HRI describes how people perceive robots as being concerned with their welfare, taking their views and personal interests into account, and working toward fair and unbiased outcomes. With these added social concerns, recent studies on trust in HRI have investigated how people attribute responsibility and blame to robots given an unfavorable outcome [Kaniarasu and Steinfeld 2014; Lei and Rau 2020]. These discussions bring forward the very ethical dimensions of human trust in robots and their perceived trustworthiness. The proposal to take into account the social and ethical dimensions of trust in HRI, through the application of the interpersonal trust no-

tion, is valuable as a first step to deepening our understanding of what happens in the interactions between humans and social robots. This work aids in recognizing that there is an added layer of complexity because it is no longer only a matter of performance but also about what follows from leveraging social rules and schemes to enhance the interaction. However, given the philosophical account of interpersonal trust compared to the technological advancement level in social robotics, Atkinson et al. asked the important question about the "appropriateness of using interpersonal trust as an analog for human-robot trust" [2012, p. 306]. They explained that making such an analogy has been argued as reasonable on the ground that some aspects of interpersonal trust also seem to be present in studies on HRI. However, not all fellow researchers are willing to draw such an analogy because of the lack of reciprocity in the interaction.

1.1. From Properties to the Event of Trust

What is interesting about this objection is that such concerns about reciprocity are a symptom of a more fundamental issue about the ontological status of the two kinds of agents. From a philosophical analysis, the issue of reciprocity touches upon the more basic ontological question of whether robots (as belonging to the class of inanimate objects) are of the *right kind* to be in the *category of objects that are appropriate targets of interpersonal trust* because their status as ontological equal to humans cannot be justified. Focusing on the ontological status of robots with a view to their properties is an intuitive and common way to reject robots as suitable objects of interpersonal trust. The needed argumentative step is to compare the relevant properties of robots with the criteria governing the category of objects that are appropriate targets of interpersonal trust established by "the 'official' philosophical inventory of things that are" [Loux and Crisp 2017, p. 13], which is also known as an ontology. The argumentative steps taken are of the general form:

- **Premise 1:** Having a certain property (P) is a necessary and sufficient criterion for belonging to the category of objects (C).
- **Premise 2:** All entities belonging to the category of objects (C) are appropriate targets of interpersonal trust.
- **Premise 3:** All entities that are part of the class inanimate objects (O) do not have the property (P).
- **Premise 4:** A robot (R) is a member of the class inanimate objects (O) .
- **Therefore:** A robot (R) does not belong to the category of objects (C) that are appropriate targets of interpersonal trust.

Although different suggestions can be made for the exact necessary and sufficient properties for members of the class of animate objects that belong to the category of objects, the notion of interpersonal trust cannot be directly applied without violating the basic requirements of both parties to be ontological equivalent as they share the same properties. However, using only the conceptualization of trust as mere reliance for the analysis of trust in HRI is undesirable because this conceptualization tends to significantly downplay the social and ethical dimensions that have already empirically proven to be relevant for human trust in robots and their perceived trustworthiness. Left unaddressed, speaking about interpersonal trust in the context of HRI forces complex metaphysical discussions about whether the relevant facts of ordinary language use in light of the truth of the relevant prephilosophical claims requires us to reevaluate whether the application of the interpersonal trust concept must be granted to robots or not. Therefore, speaking about interpersonal trust for HRI poses a challenge to the metaphysical theory of trust proposed by philosophers. A discussion that is not going to be settled easily or anytime soon. For those eager to empirically explore trust in HRI, a more pragmatic solution is required for this conceptual challenge. HRI researchers need to know the implications of such intricate philosophical discussion upon their work on trust in HRI that is motivated and held to the standard of empirical investigations. From this perspective, studies on trust in HRI must account for what happens despite better knowledge, especially in those instances where the apparent agency of robots is reflected in their use of language and their actions and behaviors toward robots that they trust or consider trustworthy.

We propose to shift the focus on trust in HRI away from only speaking about the properties of the parties involved in the interaction, but instead consider the *event* of interpersonal trust itself. This new outlook simply extends the unit of analysis beyond the identification of properties ascribed to either humans or robots to the circumstances where interpersonal trust happens. Such an approach considers not only *who* or *what* can be included in the category of objects that are appropriate targets of interpersonal trust, but also takes into account the *conditions* under which interpersonal trust occurs. Taking the study of trust in HRI to be an event, poses a new central question that is open also to empirical investigation: *Are the kind of interactions that occur between humans and robots some that could be labeled as interpersonal trust?* So even though humans and robots are still ontological of different kinds, this broader perspective permits the study of trust in HRI to consider the properties of the parties involved in the trust event without making these properties the dividing line of how we speak or consider the analysis of trust in interactions between humans and robots. From a methodological perspective, the important difference between the property and event approaches is that they operate with different criteria for the inclusion or exclusion of robots from the category of objects that are appropriate targets of interpersonal

trust. The property approach focuses on class membership of the right kind as the criterion. In contrast, the event approach considers the criterion of identity, which is to be understood as a principle stating the necessary and sufficient conditions for an event *E* and an event *E** to be identical [Bennett 1988]. We argue that our event approach for studying trust in HRI would serve the practical aim of bypassing the issues of ontological asymmetry between humans and robots while still being able to speak appropriately about interpersonal trust as the focus is now placed on the occurrence. We argue that the occurrence of interpersonal trust is bounded by the preconditions of trust.

Figure 1 Abramović and Ulay performing *Rest Energy* (1980). Courtesy of Marina Abramović and Sean Kelly Gallery, New York [Abramović 2016]. DACS 2016.

To get a quick idea about these preconditions, consider the famous and stunning art performance *Rest Energy* (1980) by Marina Abramović and Ulay that was first shown at ROSC'80 (see Figure 1). In this art performance, the two artists draw a bow and arrow to hold each other in suspension while small microphones placed under their shirts capture their accelerating heartbeats during the performance. A strong atmosphere of tension is created for around four minutes, as any wrong movement or a lapse of attention could be fatal for Abramović because the

arrow is pointing directly at her heart. While no longer in control of the situation, she is left exposed and Abramović later explained that the piece was “the ultimate portrait of trust.”[Abramović 2016, p. 255].

What this art performance can teach us is that trust is required under very specific circumstances:

1. When there is a possibility of harm (i.e., risk).
2. When there is a future-oriented likelihood of harm (i.e., uncertainty).
3. When this exposure leaves people vulnerable (i.e., vulnerability).

This art performance also illustrates that the relationship between trust and vulnerability is fundamental for understanding trusting relationships and that the occurrence of trust is a careful balance between the two parties involved as they try to prevent harm from happening. As we can see, Ulay tries not to harm (or even murder) Abramović while she does not want to be harmed even though the risk and uncertainty are evident to both of them.

1.2. Avoiding Overexposure

As Cipolla [2018] points out, there is often some reluctance to highlight this precondition when studying trust in relation to technology because “vulnerability is not usually interpreted positively, particularly when related to design or engineering” [Cipolla 2018, p. 113]. Mainly associated with overexposure to danger (i.e. risk) and unfamiliarity (i.e. uncertainty), discussions about vulnerability in regards to technology usage tends to be something that needs to be avoided, solved or explained away. Dagan et al. [2019] elaborate on this tendency in their motivation for the designing of the social wearable technology “True Colors”. They state that an explicit focus on vulnerability as a design value is rarely considered in the human-computer interaction (HCI) community, because technology is mainly seen as a tool empowering people to live a better, more pleasant, and safer life. If there are any vulnerabilities in sight, Dagan et al. [2019] continues, the developers often call for technological fixes or new innovations to solve these issues or reestablish a sense of security or protection. By characterizing this instrumental view on technology as a project of modernity, Coeckelbergh [2017] explain how the underlying assumption for the development and use of information and communication technology (ICT) reflects the agenda of vulnerability reduction. Coeckelbergh writes:

> “By means of using electronic devices, the Internet, and all kinds of ICT infrastructures we hope to become less vulnerable, to control risk. We hope to be less dependent on ’nature’, on ’the earth’, on our vulnerable bodies. We might

> even hope to liberate ourselves from a kind of Platonic dark cave where vulnerability and mortality reigns, and instead walk into the bright light of a new, invulnerable future" [2017, p. 344].

Therefore, it can be deduced from his account that the perception of technology as a form of remedy to all the possible harm of the world is a coping mechanism that does not recognize or leave any space for vulnerability. As such, it might not be too surprising that vulnerability, as an important theme for technology development and design, is rarely considered as something positive or worthwhile, unless it is merely to optimize our technological instruments and systems.

In HRI, focusing on vulnerability may also be considered problematic, but for a different reason. Through many years of ethnographic research into the way children and older adults respond and relate to robots developed to offer them companionship, Turkle [2011] warns us against how such new forms of technology can leave people very vulnerable. With her critical view on the promise of eliminating vulnerability through the reduction and simplicity of relationships by using robots to meet people's basic needs, the bad association of vulnerability with technology is now related to the danger of deception and its consequences on how people form emotional attachments. Turkle writes:

> "Technology is seductive when what it offers meets our human vulnerabilities. And as it turns out, we are very vulnerable indeed. We are lonely but fearful of intimacy. Digital connections and the sociable robot may offer the illusion of companionship without the demands of friendship" [2011, p. 1].

The strong message provided in this quote is that serious psychological harm can result from a false sense of intimacy when engaging with robots who seek to establish an emotional connection and that there is a level of enhancement involved in such kinds of interaction. The work by Turkle [2011] revolves to a large extent on presenting that the fascination with robots capable of imitating signs of care and love will eventually lead to unhealthy and unauthentic emotional attachments. This is because the possibility that such technologies offer is to spare people from the hardship and disappointment integral to developing deeper relationships with other people. By focusing on the vulnerability of people during HRI as a form of exploitation of both children and older adults who are in need of special care and love, several attempts have been made to better understand and discuss what can be done to avoid that people are potentially being deceived by robots [Sharkey and Sharkey 2020; Grodzinsky et al. 2015; Danaher 2020].

This rather gloomy outlook on the role vulnerability plays in our relation to robots is unfortunate when discussing trust in HRI. Because vulnerability is one of the preconditions of interpersonal trust, aiming to avoid vulnerability or trying to explain it away will paradoxically also undermine the demand for trust "in the ab-

sence of vulnerability trust is not required" [Misztal 2011, p. 117]. As she explain, if vulnerability is not of any concern in the first place there would be no need for anyone to trust in others because they would be able to meet their goals, needs, or gain prosperity free from the support or help of people. To live an invulnerable life would mean to be completely and utterly self-sufficient, a state that some might strive for and work hard to achieve; however, it is also still to be seen. This point was also well explained by Möllering when he wrote:

> "[...] in order to describe the typical experience of trust we often refer to the fact that actors trust *despite* their vulnerability and uncertainty, *although* they cannot be absolutely sure what will happen. They act as *if* the situation they face was unproblematic and, although they recognize their own limitations, they trust *nevertheless*" [2006, p. 6].

Central to our understanding of trust, as he shows, is that we are aware of our vulnerability but can interact and engage with the world anyway. We will argue that this is similar when we aim to understand and study interpersonal trust in HRI. Therefore, it is essential to challenge the rather negative view of the relationship between trust and vulnerability. Considering more recent studies on trust in HRI, it seems that there is already some empirical support for the consideration of vulnerability as something that is not only problematic, but could also support the interaction and engagement with robots.

1.3. Vulnerability and Trust in HRI[1]

The notion of vulnerability is similar to that of trust; it is very abstract, and its exact meaning can be hard to grasp. One way to understand what people have come to understand with vulnerability in the HRI community is to show that it have been operationalized. Several studies on trust in HRI currently take vulnerability to be some form of self-disclosure by a robot through verbal expressions and communication. Using such an understanding of vulnerability is very helpful when designing empirical studies because it is made less abstract (i.e. specific linguistic statements), which eventually render it more easily manipulated and measured. Consequently, all existing studies so far are designed to explore how expressions or utterances of vulnerability by a robot can influence human behavior or communication during HRI.

For example, Siino et al. [2008] found that a robot using a style of affective disclosure during a collaborative task in a repair scenario would result in people feeling less in control of their data but increased its like-ability. Even though, this study is not directly about trust in HRI, it is still interesting as the findings could

1 Subsection 1.3 has already been published as Hannibal [2021].

be understood as an expression of either human experience of vulnerability or perception of the robot being more vulnerable when reporting its affective state. In another example, Kaniarasu and Steinfeld [2014] were able to show that an utterance of self-blame by a robot during a collaborative task in a navigation scenario leads people to find it less trustworthy. As discussed by the authors, the tendency by people to view others negatively, who constantly make an apology for themselves despite their intention of being honest, is an effect seen in HRI that shed light on issues of distrust. However, some studies have suggested that robot self-disclosure can improve trust in HRI. Martelaro et al. [2016] found in their more recent study that, a simple robot expressing statements of vulnerability during a learning task in a tutorial scenario would result in a higher level of trust and sense of companionship. More interested in group dynamics, Sebo et al. [2018] found that when a robot during a collaborative task in a game scenario made vulnerable statements, the members would display a much higher level of engagement with it. Traeger et al. [2020] extended their work and found that the communication between the team members would improve, and their experience as part of the group would be seen positively when the robot provided statements of vulnerability. Reducing vulnerability in HRI to a form of self-disclosure is problematic in two ways.

First, operationalizing vulnerability only as the robot's behavior fails to recognize that vulnerability as a precondition of trust must always be interpreted and linked to the situatedness and temporality of the interaction. Thus, vulnerability is something that arises from the given circumstance in relation to a real and perceived vulnerability, depending on how the interaction plays out. Second, designing the vulnerability behavior of robots in the form of linguistic statements is a very narrow understanding of how robots could be vulnerable because it is a form of mimicking human vulnerability. Considering the literature so far on robot failures (e.g., [Salem et al. 2015; Ragni et al. 2016; Honig and Oron-Gilad 2018]) and cybersecurity in robotics (e.g., [Clark et al. 2017; Miller et al. 2018]), the way in which robots can be vulnerable only partially overlaps with human vulnerabilities. In other words, given that robots are ontologically of a different kind, they have their own specific types of vulnerabilities. Hence, systematically identifying these robot-specific vulnerabilities is in fact equally important to the identification of human vulnerabilities when exploring trust in HRI. As such, a gap in the current research landscape has been identified, which serves as the motivation for the expert interviews presented in the next section. Moreover, reducing vulnerability only to a property of the robot's behavior fails to recognize that vulnerability, as a precondition of trust, must always be interpreted and linked to the specific situation or moment in time. As we wish to highlight also in the later discussion about the expert interview results, it is important to include the insight that vulnerability is relational in the research on trust in HRI, because it is highly sensitive to the

ongoing and ever-changing relationship between humans and robots during interaction.

2 Expert Interviews[2]

Given these theoretical perspective, we set out to explore the aspect of vulnerability as a precondition of trust in HRI by gathering knowledge about the possible robot vulnerabilities. Guiding this research with the question of *in which way robots could be considered vulnerable?*, we decided to conduct semi-structured expert interviews with experienced and leading roboticists.

2.1. Methodology

The method for conducting expert interviews is suitable for getting a more systematic overview of knowledge within certain domains, which experts have spent many years achieving through their professional training or experience [Meuser and Nagel 2009]. For this research, expert interviews are helpful in the initial stage of identifying the possible vulnerabilities of robots. Not only do robotics experts have an extensive knowledge about the technical challenges of developing robots, they can also provide insights into what types of vulnerabilities are common across various domains of application.

On a methodological level, using expert interviews is important because of the ontological status of robots. First, given that robots do not have an inner life that connects feelings of vulnerability to higher mental states or experiences, their particular vulnerabilities can only be studied from a third-person perspective. To paraphrase Bruno Latour, whose words about scientific facts are equally relevant to this discussions, expert interviews are required because robots cannot “speak for themselves” [Latour 1993, p. 29]. Thus, we take the specialized knowledge of roboticists as a vehicle for giving expression to the specific vulnerabilities of robots.

2.2. Procedure

Over the period of nine months, we conducted in total eight semi-structured expert interviews. The purposeful sampling method [Patton 2015] was used to recruit the experts with the following selection criteria (see e.g., Table 1 for a quick overview of how the different expertise was divided among the different experts):

2 Section 2 of this book chapter has already been published as Hannibal [2021].

1. Having a disciplinary background in robotics.

2. Work experience in HRI or social robotics.

3. Research interest on the topic of trust.

To address the research question, it was enough if an expert would only fulfill one of the three criteria while ideally they would cover all of them.

Experts	**ID**	**Country**	**Expertise**
Justus Piater	Exp_JP	AT	computer vision, ML, robotics
Allan Wagner	Exp_AW	USA	AI, robotics, HRI, robot ethics, trust
Marc Hanheide	Exp_MH	UK	AI, robotics, HRI, social robotics
–	Exp_XX	–	social robotics, HRI, AI, trust
Birgit Graf	Exp_BG	DE	HRI, service robotics, applications
Kristin Schaefer-Lay	Exp_KS	USA	robotics, HRI, teams, trust
Michael Zillich	Exp_MZ	AT	computer vision, robotics, HRI
Paul Robinette	Exp_PR	USA	robotics, HRI, trust

Table 1 Overview of the experts and the used selection criteria for their inclusion.

All experts were contacted via email with an invitation to participate, which also contained more background information and the purpose of the interview. After indicating their willingness to participate in the interview, all experts were asked to sign a consent form that was sent to them in advance. The consent form clearly stated what their participation involved, their rights, and the data protection requirements set by the university. Each expert interview was conducted in English, audio recorded, and took about 30-40 minutes.

In the first part of the interview, all experts were given an opportunity to introduce themselves (i.e., “Could you tell me about your recent projects and main research interest?”). This information was needed to contextualize their disciplinary background and role as experts (see e.g., Section 2.3). Then five additional questions were asked to guide the semi-structured interviews:

- What do you consider as future application scenarios for agent-like robotic systems?
- Given your research background, how and when can an agent-like robotic systems be said to be vulnerable?

- Given your considerations of system-centered vulnerabilities, could you please rank or order them according to their importance?
- From your point of view, who would be disadvantaged if these vulnerabilities are left unaddressed?
- Considering cutting-edge technical knowledge used to develop agent-like robotic systems today, what has to be done to make agent-like robotic systems less vulnerable in your opinion?

After finishing the interview, all experts had the opportunity to give feedback and were again informed about their rights as participants.

2.3. Ethics

To ensure the protection and integrity of the experts participating, we generally followed the four-fold strategy suggested by Flick [2009]: (1) ensure voluntary consent by the participants in advance, based on sufficient and adequate information about the research project and its aim, (2) avoid causing any unnecessary harm to the participants in the process of collecting data, (3) do justice to the participants when analyzing and interpreting the collected data, and (4) guarantee the confidentiality and anonymity of all the participants when writing down and presenting the results and findings. However, given the nature of expert interviews, we excluded principle 4 for the informed consent of the experts, as it stated in the consent that we could use the name, professional title and affiliation for the purpose of direct quotations. Only one of the experts wished to remain anonymous, who we have given the expert code, Exp_XX.

On a practical level, it is important to mention that there was no official ethics board at TU Wien that was in charge of providing a standardized procedure for ethical approval of the expert interviews at the time when they were conducted. Only since 2020 has TU Wien been testing a concept of a Research Ethics Committee (Pilot REC) based on peer review to ensure a future procedure for basic standards of research ethics. However, we did our best to compensate for the lack of ethical approval because we were in contact with Dr. Marjo Rauhala about the development of the expert interviews. Since Dr. Rauhala supports all researchers at TU Wien daily with the identification of questions regarding research ethics in the role as the leader of the service unit of Responsible Research Practices[3], we received some feedback on the project description and consent form provided to the experts, so they would live up to basic standards for good research practice. For guidance about how to follow the EU regulations of GDPR, the third author

3 For more information about the service unit of Responsible Research Practices at TU Wien, we suggest visiting their website: https://www.tuwien.at/en/research/rti-support/ responsible-research-practices

ensured a check since he is in the role of Data Protection Coordinator at the Faculty of Informatics, TU Wien. This information was also provided on the consent forms that the experts were asked to sign to prepare for their interviews.

2.4. Analysis

After collecting all the expert interviews, the audio recordings were transcribed verbatim with the spoken word as the only focus [McLellan et al. 2003]. The first author coded the interviews solely using in vivo coding to summarize the exact wording, terminology, and formulations used by the expert. After a few coding cycles, related codes were then lumped into overall 13 different categories based on content and meaning similarity [Miles et al. 2020]. The decision on which category labels to use was also guided by prior classification of potential system-centered vulnerabilities as reported in previous literature on robot failures [Ragni et al. 2016; Honig and Oron-Gilad 2018] and cybersecurity in robotics [Clark et al. 2017; Miller et al. 2018]. We used a thematic analysis [Braun et al. 2019] to identify the common themes across the expert interviews. All coding, categorization, and thematic analysis of the expert interviews were done electronically using MAXQDA[4].

Theme	Category
(T1) Embodiment	(C1) Mechanical
	(C2) Sensory
	(C3) Functional
	(C4) Security
(T2) Processing	(C5) Understanding
	(C6) Learning
	(C7) Decision-making
(T3) People	(C8) Obstacle
	(C9) Perspective-taking
	(C10) Malicious
(T4) Setting	(C11) Infrastructure
	(C12) Environment
	(C13) Time

Table 2 List of the different categories and themes identified during the coding and analysis of the expert interviews.

4 Due to the COVID-19 outbreak in March 2020, all but the first expert interview were conducted online using the Skype platform.

From the analysis of the expert interviews, we were able to identify in total 13 categories of vulnerability that were then grouped into four different themes (see e.g. Table 2). Next, we provide a short description of each theme and offer few examples of how they were supported by different experts by drawing on their own wording, terminology, and formulations to summarize their main points.

2.4.1. Embodiment (T1)

Since robots are navigating and interacting with people in the real world, they have on the most basic level what experts Exp_JP and Exp_KS referred to as "physical vulnerability". Under this theme, we collected all the various vulnerabilities related to robots in the sense that they could be "fragile" (Exp_JP), "damaged" (Exp_KS), "worn down" (Exp_BG), "hacked" (Exp_XX), or "break down" (Exp_AW). As such, the aim of this theme was to highlight that regarding their various mechanical (C1), sensory (C2), functional (C3), and security (C4) aspects, robots can be exposed because their required embodiment creates tangible vulnerabilities.

2.4.2. Processing (T2)

On a more abstract level, but still related to the functioning of robots, the next theme is related to their ability to handle and use the information they get from the surroundings for understanding (C5), learning (C6), and decision-making (C7) as Exp_JP mentioned that "softwares are also vulnerable". Central to this theme are the different robot vulnerabilities that arise because they "lack a conceptual framework that allows them to understand what is going on in the world" (Exp_JP), could be "learning the wrong thing" (Exp_MH), or could "make decisions when they do not have all of the information" (Exp_XX). Thus, these kinds of robot vulnerabilities are mainly to be understood as a form of exposure in the sense of inadequate, misinformed, and hasty reasoning that eventually guides their behavior.

2.4.3. People (T3)

Moving on to those aspects that are more external to the robot, the next theme relates specifically to the action or behavior of the people interacting with them and that would have a direct effect on their level of exposure. For some of the experts, the robot vulnerabilities were not simply a matter of people sometimes hindering task completion by the robot because "the human does not move so the robot has to turn" (Exp_PR) or that the limited "understanding in humans how the

robots see the world" (Exp_MH) would result in robots getting into various accidents. In some cases people would in fact be downright "malicious" (Exp_PR) as they would intentionally engage in "abusive, aggressive behavior towards robots in the public" (Exp_MH). Thus, this theme intends to show that vulnerabilities are closely linked to both the unintentional and intentional conduct of the human counterpart because people expect that robots can easily deal with constantly moving obstacles (C8), fail to understand or take into account the perspectives of robots (C9) that leads to hazardous situations, and assume that mistreating robots by participating in malicious (C10) activities is unproblematic.

2.4.4. Setting (T4)

In the last theme, we collected the robot vulnerabilities mentioned by the experts, which relate to the framing or backdrop against which the interaction between humans and robots unfolds. In this theme, the often hidden technological and bureaucratic infrastructure (C11) was stressed because getting robots to properly function in real world scenarios often requires "ten engineers standing around" (Exp_BG) or "getting safety certificates" (Exp_MZ) to ensure robots could leave the laboratory and enter the market. Even when being tested for application, robots regularly get challenged when having to navigate in an environment (C12) designed for humans, which Exp_BG identified when she explained that "sometimes the corridor was simply too narrow" (Exp_BG) or that people would constantly be "moving stuff around". Time (C13) was also considered important given that according to Exp_KS there is a difference between those robot vulnerabilities that only show in the long-term compared to those "that happen right away". In the view of Exp_MH, the aspect of time might also be crucial in understanding why people "like to mess around" - because new and short encounters instigate a "novelty effect".

2.5. Discussion

There are several points to consider for discussing the results, which we will present in this section while relating them to existing literature in HRI and other relevant discussions.

2.5.1. Interpretation of Vulnerability

As expected, some of the experts would comment on how to interpret the notion of vulnerability in relation to robots. For example, Exp_PR considered how to understand robot vulnerability in light of how they are often portrayed in the media

and pop-culture. He noted that while people always see in movies that "robots are super strong and super fast and everything" this is far from the case because in "the real world they cannot get over a single step or they think that a bush is an obstacle that cannot be driven or something". Thus, Exp_PR concludes, that robots are "already pretty vulnerable in the real world" compared to the impression that the general public might have. This point is closely related to debates in HRI about managing public expectations regarding the robot's capabilities. Known by now as the "expectation gap" [de Graaf et al. 2016; Kwon et al. 2016], it is also highly relevant and recently linked to discussions regarding trust in HRI, as this gap could result in unwanted disappointment and even instigate fear [Malle et al. 2020].

More concerned with some conceptual challenges, Exp_AW expressed difficulties with speaking about robot vulnerabilities when saying that "vulnerability is just not a topic that's really very well suited for robots" because in his view using this notion would suggest that robots have some kind of volition or intentionality. Exp_AW further explained how this issue made him hesitate in using the common definition of trust by Mayer et al. [1995] and instead turned toward a "definition that involved risk", which is more practical and widespread in robotics since it is easier to operationalize. Another similar reflection was made by Exp_BG who said that "it's really hard to think about vulnerable in the sense of the robot because for me it's an attribute that's so human". Based on her more technical perspective, she then suggested reformulating the relevant aspect of considering robot vulnerability in terms of "situations where the robot could run into problems". This conceptual tension when studying trust in HRI has previously been identified by Malle and Ullman [2021] and it is still an open question whether human-robot trust necessarily comes with a feeling of vulnerability, which is a characteristic of human trust.

According to Exp_KS, such discussion must consider that speaking about robot vulnerabilities also contains a normative dimension because people in different contexts might need to ask themselves critically, "how vulnerable do we need to be to the system, how vulnerable does the system need to be to me?". She elaborates on this point by saying that robot vulnerabilities in a military context must always be avoided, whereas it might be useful in healthcare for building trust between people and robots. Questions about when and for what reasons robot vulnerabilities might be desirable or not are important to discussions about trust in HRI because the mere presence of a robot perceived as vulnerable can in fact influence human group dynamics for the better [Traeger et al. 2020].

2.5.2. Ethical Dimensions

From the expert interviews, it turned out that the theme of people (T3) ranked as the second most mentioned robot vulnerability despite different domains of application (coded 57 times). Especially the challenge of malicious humans was mentioned by several experts, who noted that people would intentionally be “kicking”,“pushing”, “hitting”, and “attacking” robots, which adds to previous HRI literature reporting how both adults and children would not shy away from such behavior [Scheeff et al. 2002; Brscić et al. 2015; Nomura et al. 2016]. This abusive behavior toward robots will only grow with their increasing application in public spaces, which according to Exp_KS is problematic for trust in HRI because “it will become an issue for their operation”. Given that the success or failure of a given task in fact depends on some level of mutual trust in HRI, it is relevant to study not only whether people can trust robots, but also whether robots can trust people [Vinanzi et al. 2019].

The necessity of mutual trust in HRI for task completion and collaboration requires a broader discussion about how to deal with human abusive behavior toward robots, and this challenge has already been recognized as an ethical dimension of HRI [Whitby 2008].

From a critical analysis of previous attempts in philosophy to account for trust that mainly originated from a liberal tradition, Baier [1986] argued that the significance of trust for thriving must be examined from a moral point of view. From her perspective, it is a bad starting point for any understanding of trust pertinent to human social life to consider it as some form of contract established between two equal parties in terms of power and capabilities. From her careful observation of interpersonal relationships of all kinds where cooperation and care are cardinal, she recognizes that some of them are fundamentally unequal and sometimes not even voluntary, which severely challenges the liberal ideal of the conditions of trust. Based on this insight, Baier [1986] propose instead to take trust as a form of reliance on others to act out of goodwill toward oneself. This so-called goodwill account of trust is essential in stressing the close connection between interpersonal trust with moral obligations and is one of the first views on trust that goes beyond reliance.

However, debates about mutual trust rooted in a liberal tradition have become challenging for HRI because they presume that the two parties stand to each other in an equal moral and power relation [Faulkner and Simpson 2017]. The acknowledgment of robot vulnerability in relation to their human counterpart is ethically problematic as they can at most be considered “moral patients” [Coeckelbergh 2018], and they do not have a choice whether or not to engage in the interaction [Baier 1986].

Considering both the limited moral standing of robots and the inequality of power in HRI, we agree with Tolmeijer et al. [2020] that future work needs to focus more on developing concrete trust-repair strategies for what they refer to as “user failure” to mitigate robot vulnerabilities resulting from abusive behavior. From their main focus on interaction design strategies for mutual trust in HRI, they have suggested that robots could use methods of apology, showing emotions, and involving authority figures. More concerned with ethical and legal strategies, debates in philosophical circles have been revolving around granting some form of “robot rights” [Coeckelbergh 2010b; Gunkel 2018], which is a rather controversial suggestion [Tavani 2018].

3 Relational Dimension of Vulnerability

Throughout his work on developing a normative anthropology of vulnerability, Coeckelbergh [2013] draws on the traditions of phenomenology and pragmatism for analyzing vulnerability in relation to technology, as an alternative to the more classical scientific approach. As he writes, the understanding that the classical sciences brings to the foreground of the discussion is one where “vulnerability appears as an objective, essential feature of human nature, and the vulnerability of people is studied in an objectivist way” [Coeckelbergh 2013, p. 38-39]. From this perspective, he continues, vulnerability is something external to people, which can be evaluated from a third-person point of view. Vulnerably is thereby characterized in objective terms; is vulnerability *real* compared to the possible risk and uncertainty posed by a threat to the livelihood or well-being of people. In this sense, the individual experience of being vulnerable is not considered or at least something that can be managed when understood properly. As Coeckelbergh [2013] explains, those who do in fact speak about vulnerability as tied to the subjective feelings or emotions of people still presuppose that the perception of being vulnerable is seen in the light of an objective standard. Taking the complete opposite view, is to consider vulnerability only as subjective where the first-person perspective is in focus, how the “I” (or individual) comes to experience the vulnerability. However, he argues that this view is also problematic because it does not acknowledge that the subjective experience of vulnerability is influenced by the surroundings and conditions people find themselves in. Vulnerability is connected to the way people interact and engage with the world, which contains both risk and uncertainty as part of daily life. Thus, Coeckelbergh [2013] aims to challenge this overall idea of the object-subject dichotomy to our understanding of vulnerability ingrained in the Western thought. As a way out of this dualistic view on vulnerability, he proposes to shift the focus on how vulnerability emerges out of this tension so that it “[...] is neither a feature of the world (an objective,

external state of affairs) nor something that we create or perceive (a subjective construction by the mind, an internal matter), but is constituted in the subject-object relation" [Coeckelbergh 2013, p. 43].

From this critical discussion, Coeckelbergh [2013] elaborates on what he means when he takes vulnerability to be relational, that closely connected with the notion of engagement. He states that vulnerability arises from or comes into view only within the relation that manifests when people engage with the world. It is nothing that already belongs to people or the world in advance, but something that unfolds in that meeting. Following this understanding of vulnerability as something emergent during the interaction is also relevant to the way it is possible to think about vulnerability for studies on trust in HRI. Given that vulnerability fundamentally emerges from the interaction or engagement between humans and robots, it would be a mistake to reduce it to being a property of the robots nor of the perceptions people have, as reported in from previous work. Rather, it is something that must be located in the event of a meeting. As relational vulnerability in HRI, we can take the co-constitution of vulnerability as a result of both the human and robot who are coming into interaction or engagement. While Coeckelbergh puts a lot of effort into stressing the value of this analysis because it makes room for the existential dimensions of a "vulnerable being" [2013, p. 44], we argue that the more important point he makes, and the most relevant for the HRI community, is that it also enables us to see vulnerability as a process; vulnerability is continuously ongoing. Since vulnerability is relational in terms of interaction and engagement, it also means that it is always in the making. Coeckelbergh makes this point clear when he writes:

> "Vulnerability is not merely passive. To understand vulnerability as something entirely passive would be to turn the human being into an object once again or a *property* of that object. But *openness* does not mean passivity, and vulnerability is not merely a characteristic of our body or our mind. We are not vulnerable in the way a building or a bridge is vulnerable. Rather, we *make* ourselves vulnerable; we put ourselves at risk, by our mental and physical actions. We eat, we travel, we work, we love, we hope, and these actions make us vulnerable. Vulnerability, therefore, is not a property of the human person but a feature of the relation between us and the world. It is a feature of our way of being (in the world) and a way of existing" [2013, p. 44].

Translating this insight into the context of trust in HRI, we can say that it is possible to consider vulnerability as a result of the exchange between the human and robot that always occur. Although robots are of a completely different kind than humans, we believe that this does not hinder the recognition that they play their own important role in the creation of vulnerability. Just as anything else in the world, which confronts people as part of their everyday life, our meeting with

robots can potential shape the way we come to experience and understand our vulnerability through encounters. This is similar to how robots can be considered vulnerable in the meeting with people. They are also affected by the actions and behaviors of humans, even though the issues that robots face from such meetings might not have the same existential consequences. However, there are potential risks and uncertainties that robots face when navigating in human spaces, which render them vulnerable and thus bring the theme of trust as bidirectional into the discussion.

4 Conclusion

In this book chapter, we have considered some theoretical and empirical work in deepening an understanding of interpersonal trust in HRI. First, we considered how trust had been understood in the context of HRI on a conceptual level, leading to deeper philosophical questions about the metaphysics of taking trust to be an event rather than a property as a way to highlight vulnerability as one of the preconditions less explored. Then, we then presented the results of eight expert interviews that aimed to explore how robots could be said to be vulnerable in interactions requiring trust. Based on the systematic overview, we discussed how robot vulnerability is challenging our conceptual associations and how such a stance leads to broader social and ethical discussions on trust in HRI, where mutual trust is essential in strengthening the interaction or collaboration. Finally, we reflected on how the current shift toward vulnerability as an emergent aspect of mutual trust in HRI aligns with a general view on how interpersonal trust is always a result of the ongoing exchange between humans and robots, even though they are of ontological different kinds.

In summary, our book chapter presents an interdisciplinary perspective on the analysis of trust for current HRI research. Although there are still many open questions to be addressed and further empirical work to be carried out, we believe that the initial steps have been taken toward new directions of understanding and studying trust in HRI. Furthermore, our work is also helpful in fostering a stronger dialog about how to combine both theoretical and empirical perspectives on the complex way of recognizing robot vulnerabilities that can support trust in HRI.

Bibliography

Marina Abramović. 2016. *Walk Through Walls: A Memoir.* Crown Archetype, New York (NY), USA. 304 pages.

Facundo M Alonso. 2014. What is reliance? *Canadian Journal of Philosophy* 44, 2 (4 2014), 163–183.

David Atkinson, Peter Hancock, Robert R Hoffman, John D Lee, Ericka Rovira, Charlene Stokes, and Alan R Wagner. 2012. Trust in Computers and Robots: The Uses and Boundaries of the Analogy to Interpersonal Trust. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 56, 1, 303–307.

Annette Baier. 1986. Trust and Antitrust. *Ethics* 96, 2 (1986), 231–260.

Jonathan Bennett. 1988. *Events and Their Names.* Vol. 28. Clarendon Press, Oxford, UK. 243 pages. https://doi.org/10.2307/2215925

Virginia Braun, Victoria Clarke, Nikki Hayfield, and Gareth Terry. 2019. Thematic Analysis. In *Handbook of Research Methods in Health Social Sciences*, P. Liamputtong (Ed.). Springer Nature Singapore Pte Ltd., Singapore, Chapter 48, 843–860.

Cynthia Breazeal. 2001. Socially Intelligent Robots: research, development, and applications. In *IEEE International Conference on Systems, Man and Cybernetics,* Vol. 4. IEEE, Tucson (AZ), USA, 2121–2126.

Drazen Brscić, Hiroyuki Kidokoro, Yoshitaka Suehiro, and Takayuki Kanda. 2015. Escaping from Children's Abuse of Social Robots. In *Proceedings of the 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI'15)*. ACM, Portland (OR), USA, 59–66.

Carla Cipolla. 2018. Designing for Vulnerability: Interpersonal Relations and Design. *She Ji: The Journal of Design, Economics, and Innovation* 4, 1 (2018), 111–122.

George W Clark, Michael V Doran, and Todd R Andel. 2017. Cybersecurity issues in robotics. In *Proceedings of the IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA)*. IEEE, Savannah (GA), USA, 1–5.

Mark Coeckelbergh. 2010a. Artificial Companions: Empathy and Vulnerability Mirroring in Human-Robot Relations. *Studies in Ethics, Law, and Technology* 4, 3 (2010), Article 2.

Mark Coeckelbergh. 2010b. Robot rights? Towards a social-relational justification of moral consideration. *Ethics and Information Technology* 12, 3 (2010), 209–221.

Mark Coeckelbergh. 2013. *Human being @ risk: Enhancement, technology, and the evaluation of vulnerability transformations*. Springer (Science & Business Media), Berlin. 372 pages.

Mark Coeckelbergh. 2017. The Art of Living with ICTs: The Ethics-Aesthetics of Vulnerability Coping and Its Implications for Understanding and Evaluating ICT Cultures. *Foundations of Science* 22, 2 (2017), 339–348.

Mark Coeckelbergh. 2018. Why care about robots? Empathy, moral standing, and the language of suffering. *Kairos. Journal of Philosophy & Science* 20, 1 (2018), 141–158.

Ella Dagan, Elena Márquez Segura, Ferran Altarriba Bertran, Miguel Flores, and Katherine Isbister. 2019. Designing 'True Colors': A Social Wearable that Affords Vulnerability. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.* ACM, Glasgow, Scotland, 1–14. https://doi.org/10.1145/3290605.3300263

John Danaher. 2020. Robot Betrayal: a guide to the ethics of robotic deception. *Ethics and Information Technology* 22 (2020), 117–128.

Kerstin Dautenhahn. 1995. Getting to Know Each Other - Artificial Social Intelligence for Autonomous Robots. *Robotics and Autonomous Systems* 16, 2-4 (1995), 333–356.

Maartje M A de Graaf, Somaya Ben Allouch, and Jan A G M van Dijk. 2016. Long-term evaluation of a social robot in real homes. *Interaction Studies* 17, 3 (12 2016), 461–490.

Ewart J De Visser, Marieke, M M Peeters, Malte, F Jung, Spencer Kohn, Tyler, H Shaw, Richard Pak, and Mark A Neerincx. 2020. Towards a Theory of Longitudinal Trust Calibration in Human-Robot Teams. *International Journal of Social Robotics* 12 (2020), 459–478.

Paul Faulkner and Thomas Simpson. 2017. Introduction. In *The Philosophy of Trust,* P Faulkner, T Simpson (Eds.). Oxford University Press, Oxford, UK. 3–14 pages.

Uwe Flick. 2009. *An Introduction to Qualitative Research* (4th ed.). SAGE Publications Ltd., London. 504 pages.

Frances S Grodzinsky, Keith W Miller, and Marty J Wolf. 2015. Developing Automated Deceptions and the Impact on Trust. *Philosophy & Technology* 28, 1 (3 2015), 91–105.

David J Gunkel. 2018. The Other Question: Can and Should Robots Have Rights? *Ethics and Information Technology* 20, 2 (2018), 87–99.

Glenda Hannibal. 2021. Focusing on the Vulnerabilities of Robots through Expert Interviews for Trust in Human-Robot Interaction. In *16th ACM/IEEE International Conference on Human-Robot Interaction.* ACM, Boulder, CO, 288–293.

Shanee Honig and Tal Oron-Gilad. 2018. Understanding and Resolving Failures in Human-Robot Interaction: Literature Review and Model Development. *Frontiers in Psychology* 9, 861 (2018). https://doi.org/10.3389/FPSYG.2018.00861

Poornima Kaniarasu and Aaron M. Steinfeld. 2014. Effects of blame on trust in human robot interaction. In *The 23rd IEEE International Symposium on Robot and Human Interactive Communication.* IEEE, Edinburgh, UK, 850–855.

Bing Cai Kok and Harold Soh. 2020. Trust in Robots: Challenges and Opportunities. *Current Robotics Reports* 1, 4 (12 2020), 297–309.

Minae Kwon, Malte F. Jung, and Ross A. Knepper. 2016. Human expectations of social robots. In *Proceedings of the 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI'16)*. IEEE, Christchurch, New Zealand, 463–464.

Bruno Latour. 1993. *We Have Never Been Modern.* Harvard University Press, Cambridge (MA), USA. 168 pages.

Jin Joo Lee, W. Bradley Knox, Jolie B. Wormwood, Cynthia Breazeal, and David DeSteno. 2013. Computationally modeling interpersonal trust. *Frontiers in Psychology* 4 (12 2013), 893. https://doi.org/10.3389/fpsyg.2013.00893

Xin Lei and Pei-Luen Patrick Rau. 2020. Should I Blame the Human or the Robot? Attribution Within a Human–Robot Group. *International Journal of Social Robotics* (4 2020), 1–15. https://doi.org/10.1007/s12369-020-00645-w

Michael J Loux and Thomas M Crisp. 2017. *Metaphysics: A contemporary introduction* (4 ed.).Routledge, New York (NY), USA. 1–356 pages.

Bertram F Malle, Kerstin Fischer, James E Young, Ajung Moon, and Emily C Collins. 2020. Trust and the discrepancy between expectations and actual capabilities of social robots. In *Human-robot interaction: Control, analysis, and design,* D. Zhang and B. Wei (Eds.). Cambridge Scholars Publishing, New York, NY, USA, Chapter 1, 1–23.

Bertram F Malle and Daniel Ullman. 2021. 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.). Academic Press, London, UK, Chapter 1, 3–25.

Nikolas Martelaro, Victoria C Nneji, Wendy Ju, and Pamela Hinds. 2016. Tell me more: Designing HRI to encourage more trust, disclosure, and companionship. In *11th ACM/IEEE International Conference on Human-Robot Interaction (HRI).* IEEE, Christchurch, New Zealand, 181–188.

Roger C Mayer, James H Davis, and F David Schoorman. 1995. An Integrative Model of Organizational Trust. *The Academy of Management Review* 20, 3 (1995), 709–734.

Eleanor McLellan, Kathleen M. MacQueen, and Judith L. Neidig. 2003. Beyond the Qualitative Interview: Data Preparation and Transcription. *Field Methods* 15, 1 (2 2003), 63–84.

Michael Meuser and Ulrike Nagel. 2009. The Expert Interview and Changes in Knowledge Production. In *Interviewing Experts* (1st ed.), Alexander Bogner, Beate Littig, and Wolfgang Menz (Eds.). Palgrave Macmillan, Hampshire, UK, Chapter 1, 281.

Matthew B Miles, A Michael Huberman, and Johnny Saldaña. 2020. *Qualitative Data Analysis: A Methods Sourcebook* (4 ed.). SAGE Publications Ltd., Thousand Oaks, CA. 381 pages.

Justin Miller, Andrew B Williams, and Debbie Perouli. 2018. A Case Study on the Cybersecurity of Social Robots. In *Proceedings of 13th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. ACM, New York (NY), USA, 195–196.

Barbara A Misztal. 2011. *The Challenges of Vulnerability: In Search of Strategies for a Less Vulnerable Social Life.* Palgrave Macmillan, Hampshire, UK. 272 pages.

Guido Möllering. 2006. *Trust: Reason, Routine, Reflexivity.* Emerald Group Publiching Limited, Bingley, UK. 244 pages.

Tatsuya Nomura, Takayuki Kanda, Hiroyoshi Kidokoro, Yoshitaka Suehiro, and Sachie Yamada. 2016. Why do children abuse robots? *Interaction Studies* 17, 3 (12 2016), 347–369.

Sven Nyholm. 2020. *Humans and Robots: Ethics, Agency, and Anthropomorphism.* Rowman & Littlefield International, London, UK. 236 pages.

Rui Ogawa, Sung Park, and Hiroyuki Umemuro. 2019. How Humans Develop Trust in Communication Robots: A Phased Model Based on Interpersonal Trust. In *Proceedings of the 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI).* IEEE, Daegu, South Korea, 606–607.

Michael Quinn Patton. 2015. *Qualitative Research & Evaluation Methods: Integrating Theory and Practice* (4 ed.). SAGE Publications, Inc., Thousand Oaks, CA. 806 pages.

Marco Ragni, Andrey Rudenko, Barbara Kuhnert, and Kai O Arras. 2016. Errare humanum est: Erroneous robots in human-robot interaction. In *The 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, New York (NY), USA, 501–506.

Maha Salem, Gabriella Lakatos, Farshird Amirabdollahian, and Kerstin Dautenhahn. 2015. Would You Trust a (Faulty) Robot?: Effects of Error, Task Type and Personality on Human-Robot Cooperation and Trust. In *Proceedings of the 10th Annual ACM/IEEE International Conference on Human-Robot Interaction (CHII)*. ACM, Portland (OR), USA, 141–148.

Mark Scheeff, John Pinto, Kris Rahardja, Scott Snibbe, and Robert Tow. 2002. Experiences with Sparky, a Social Robot. In *Socially Intelligent Agents - Creating Relationships with Computers and Robots,* Kerstin Dautenhahn, A. Bond, L. Cañamero, and B. Edmonds (Eds.). Springer, Boston (MA), USA, Chapter 21, 173–180.

Sarah Sebo, Margaret Traeger, Malte Jung, and Brian Scassellati. 2018. The Ripple Effects of Vulnerability: The Effects of a Robot's Vulnerable Behavior on Trust in Human-Robot Teams. In *Proceedings of the 13th ACM/IEEE International Conference on Human-Robot Interaction*. ACM, Chicago (IL), USA, 178–186.

Amanda Sharkey and Noel Sharkey. 2020. We need to talk about deception in social robotics! *Ethics and Information Technology* (2020), 1–8. https://doi.org/10.1007/s10676-020-09573-9

Rosanne M Siino, Justin Chung, and Pamela J Hinds. 2008. Colleague vs. tool: Effects of disclosure in human-robot collaboration. In *The 17th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN).* IEEE, Munich, Germany, 558–562.

Matthew Noah Smith. 2010. Reliance. *Noûs* 44, 1 (2 2010), 135–157.

Herman Tavani. 2018. Can Social Robots Qualify for Moral Consideration? Reframing the Question about Robot Rights. *Information* 9, 4 (3 2018), 73.

Suzanne Tolmeijer, Astrid Weiss, Marc Hanheide, Felix Lindner, Thomas M Powers, Clare Dixon, and Myrthe L Tielman. 2020. Taxonomy of Trust-Relevant Failures and Mitigation Strategies. In *15th ACM/IEEE International Conference on Human-Robot Interaction (HRI).* ACM, Cambridge, United Kingdom, 3–12.

Margaret L Traeger, Sarah Strohkorb Sebo, Malte Jung, Brian Scassellati, and Nicholas A Christakis. 2020. Vulnerable robots positively shape human conversational dynamics in a human-robot team. *Proceedings of the National Academy of Sciences of the United States of America* 117, 12 (3 2020), 6370–6375.

Sherry Turkle. 2011. *Alone Together: Why We Expect More From Technology and Less From Each Other.* Basic Books, New York (NY), USA. 400 pages.

Samuele Vinanzi, Massimiliano Patacchiola, Antonio Chella, and Angelo Cangelosi. 2019. Would a robot trust you? Developmental robotics model of trust and theory of mind. *Philosophical Transactions of the Royal Society B* 374, 1771 (2019), 1–9.

Blay Whitby. 2008. Sometimes it's hard to be a robot: A call for action on the ethics of abusing artificial agents. *Interacting with Computers* 20, 3 (2008), 326–333.

Jakub Złotowski, Hidenobu Sumioka, Shuichi Nishio, Dylan F Glas, Christoph Bartneck, and Hiroshi Ishiguro. 2016. Appearance of a robot affects the impact of its behaviour on perceived trustworthiness and empathy. *Paladyn* 7, 1 (2016), 55–66.