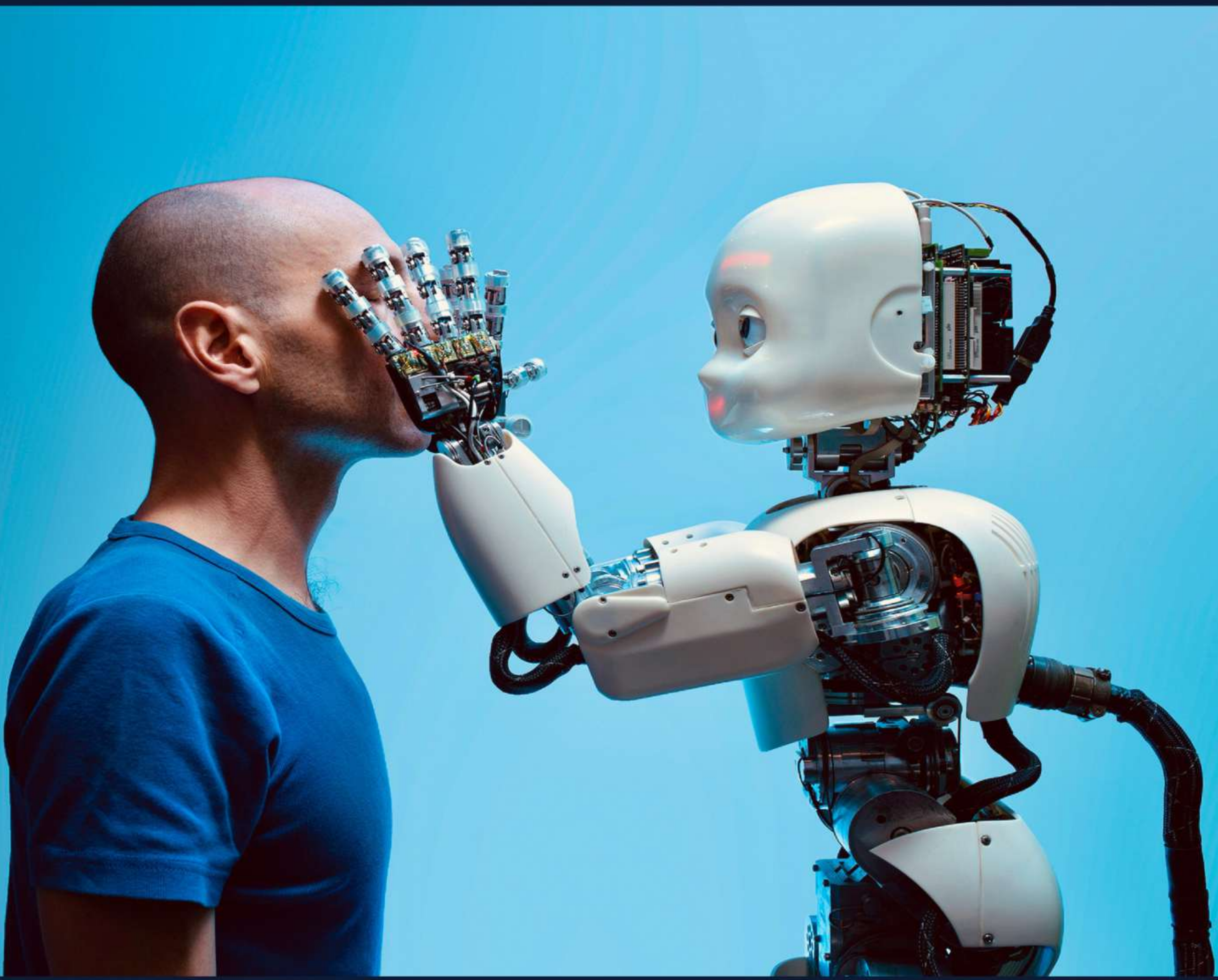
DISSERTATION



THE TRUST-VULNERABILITY RELATION

A Theory-driven and Multidisciplinary Approach to the Study of Interpersonal Trust in Human-Robot Interaction

Glenda HANNIBAL

Cover photo by ©2013 Émile Loreaux. Permission to use this photo was obtained via personal correspondence.

TU WIEN Informatics

The Trust-Vulnerability Relation

A Theory-driven and Multidisciplinary Approach to the Study of Interpersonal Trust in Human-Robot Interaction

DISSERTATION

submitted in partial fulfillment of the requirements for the degree of

Doktorin der Technischen Wissenschaften

by

Glenda Hannibal, BA MA
Registration Number 01549372

to the Faculty of Informatics
at TU Wien

Supervisor: Asst. Prof. Dr. Astrid Weiss
Co-supervisors: Assoc.Prof. Dr. Peter Purgathofer and Assoc.Prof. Dr. Markus Vincze

The dissertation has been reviewed by:

Prof. Dr. Bertram F. Malle

Prof. Dr. James E. Young

Vienna, 7th October, 2022

Glenda Hannibal

Declaration of Authorship

Glenda Hannibal, BA MA

I hereby declare that I have written this dissertation independently, that I have completely specified the utilized sources and resources and that I have definitely marked all parts of the work - including tables, maps and figures - which belong to other works or to the internet, literally or extracted, by referencing the source as borrowed.

Vienna, 7th October, 2022

Glenda Hannibal

Til Yago Bundgaard, for at have introduceret mig til filosofiens væsentlighed.

Abstract

Trust in robots and their trustworthiness has been studied and promoted in the literature on human-robot interaction (HRI) as an essential factor for how willing people are to interact, collaborate, and engage with them. Robots, deliberately designed to have apparent agency (in terms of both appearance and behavior), have motivated ongoing research on trust in HRI to go beyond an understanding of trust as mere reliance to that of *interpersonal trust*. While much work has focused on either identifying the various factors influencing human trust in robots or proposing which principles should guide the development and design of trustworthy robots, understanding how the concept of interpersonal trust in HRI is meaningful and useful for analyzing and studying trust in HRI is very often left unquestioned. However, speaking about and providing empirical evidence for interpersonal trust in the context of HRI, similarly to the way it has been characterized in relationships between people, requires that the trust-vulnerability relation is both identified and tested. To this date, no thorough and systematic investigation has been undertaken to show how an emphasis on vulnerability as a precondition of trust can help advance current understanding, analysis, and studies on interpersonal trust in HRI.

Using the philosophical method of conceptual analysis, I show that vulnerability stands as a precondition to the concept of trust. I argue that the inclusion of such a conceptual relation for our analysis and studies of trust in HRI must adapt an event approach: trust emerges from the interactions between humans and robots that occur in specific spatio-temporal situations. With this theoretical outset, I developed and conducted three empirical studies to explore either human experience of vulnerability or the vulnerabilities of robots in relation to trust in HRI. The human-centered online HRI studies revolve around the everyday life context of getting assistance from a robot in shopping for clothes. In the first online HRI study, we compare through the use of an interactive online survey three different interaction scenarios, each of which contained a subtle trust violation instance (relating to the theme of economy, privacy, and transparency respectively) and found that people do feel vulnerable in such situations to the degree that it correlates with their trust ratings. We conclude that the trust-vulnerability relation is

present and measurable in the ordinary, mundane and familiar situation of clothes shopping. In the follow-up online HRI study, we zoomed in on the privacy-related scenario and gained additional and deeper insights into people's motivations and reasoning by adding follow-up interviews to the iterated interactive online survey. We found that the experience of vulnerability that people felt was only slightly correlated to whether they would consider interacting or engaging again with the imperfect robot that posed a mild privacy breach. Among other things, we discussed and concluded that focusing on the trust-vulnerability relation sheds light on how the privacy paradox is also present when studying interpersonal trust in HRI, and that the reasons people had for trusting the imperfect robot were related to expectation of either utility or entertainment values. Taking a robot-centered perspective, I conducted several expert interviews to explore in which way robots could be considered vulnerable. By discussing how the concept of vulnerability could foster a new way of thinking about the shortcomings of robots, and by focusing on the identification of malicious humans as a challenge for robots, I conclude that interpersonal trust in HRI also needs to consider to what extent it needs to be mutual for discussions about ethical concerns and design considerations.

With my dissertation I contribute with fundamental knowledge about how to conceptualize interpersonal trust in HRI, what kind of conceptual knowledge needs to be considered and applied to guide empirical HRI studies on the trust topic or the development of trustworthy robots, and how the trust-vulnerability relation can be empirically studied through interaction scenarios between humans and robots in the everyday life situation of clothes shopping. As such, the work I undertook and presented in my dissertation can help advance future research on interpersonal trust in HRI by: (i) drawing attention to the importance of the trust-vulnerability relation for studies on trust in HRI, (ii) providing first steps to explore empirically the trust-vulnerability relation from both a human- and robot-centered perspective, (iii) and discussing what value interpersonal trust in HRI might have in supporting successful HRI in more commercial domains of application.

Acknowledgements

Given not only the regular challenges of PhD-ing, but also the unusual circumstances following the global COVID-19 outbreak in March 2020, it is with great pleasure and relief that I am now able to submit my work for evaluation. This achievement has only been possible with daily encouragement from my colleagues, collaborators, friends, and family over the last four years. Therefore, I am incredibly grateful for all of the support I have received in the pursuit of my PhD and the writing of this dissertation.

First and foremost, to my supervisor Dr. Astrid Weiss. Astrid, you have truly encouraged me to explore, try new ideas, and make mistakes so that I could mature as an interdependent early-stage researcher. Your eminent understanding and advice was also important to me during times when I needed to overcome moments of doubt and hardship. Thank you for your indispensable guidance and mentorship. I really hope that we will continue our collaboration beyond the time I have enjoyed working under your supervision.

To my first co-supervisor Dr. Peter Purgathofer, thank you for giving me the opportunity to do my PhD at TU Wien, for letting me develop my own research direction, and for helping me navigate through the bureaucracy that comes with working in a cross-faculty doctoral college. To my second co-supervisor Dr. Markus Vincze, thank you for co-organizing the Trust Robots Doctoral College at TU Wien with Dr. Sabine Köszegi. Being part of this doctoral college has added to my experience as a PhD student a sense of community on both a professional and personal level. Markus, I have learned a lot about how engineers think and approach problems in robotics from our many discussions about my own work and that of others. From our exchanges, Sabine, I have gained new perspectives on what is important about having a sociological perspective on the developments of robots and their deployment into society. To the members of my evaluation committee, Dr. Bertram F. Malle, Dr. James E. Young, Dr. Margrit Gelautz, and Dr. Martina Mara: thank you for your thoughtful advice and constructive feedback on my PhD project. Especially to Bertram and James for taking their time to provide extensive comments on my dissertation, which have improved greatly from your feedback. I hope our paths will cross again so we can continue our exciting and productive discussions about how to study trust in human-robot interaction.

To my fellow PhD students in the Trust Robots Doctoral College at TU Wien, working alongside all of you has been a great experience, and I wish you all the best of luck in the future. Thank you Matthias for helping me out with the videos for the online survey, and thank you Guglielmo and Dominik for allowing me to slowly start exploring the fascinating topic of explainable AI through our research collaboration. Also a special thanks to Darja, Helena, and Christina, for the many great moments we have also shared as friends over the last four years.

To the other members of the human-computer interaction (HCI) group at TU Wien, thank you for your help with the everyday life as a researcher and for the warm welcome I received as the first philosopher joining. Especially to Monika for always being so patient and helpful with any administrative tasks, to Michi for helping me out with the formalities of the Proficiency Evaluation and for listening to my struggles during the PhD-ing. Kay, thank you for having been the best office buddy and an important support as my proof-reader for the dissertation – I wish you all the best with your PhD project too.

To my collaborators in the Machine Intelligence and Human Behavior (HUMAINT) project at the EU Commission: thank you Emilia for enabling our collaboration and Marina for helping program the PEPPER robot before we had to give up the in-person HRI study due to the COVID-19 outbreak. Especially, thank you Vicky for your collaborative effort. Your feedback and input on my research, as well as your advice and guidance, has been important to the development of my dissertation. I hope we can continue our collaboration and fruitful discussions in the future.

To my other collaborators in Vienna and abroad that I had had the great opportunity to work with. Anna, let us stay in contact as I always enjoy or discussions about our life as academic misfits and how to overcome the challenges of working across disciplines. Felix, your always critical feedback on my PhD project and our work together has been valuable in ensuring that I feel well prepared for defending my ideas and research. I have truly enjoyed getting to know you Patrícia, you are such a positive and driven colleague that is not only very knowledgeable about design thinking, but also curious to learn from others. I hope we can continue our work to finish the things we have already started and maybe find new exciting projects to carry out. To Theresa and Nick, thank you for joining in my vision for more theory-driven HRI, and for the moments we have shared feeling part of something bigger than our individual PhD projects. I am really looking forward to meeting you in-person eventually.

To all the experts who accepted my invitation and to everyone who participated in my online studies, thank you for your valuable contributions to realizing this work. To all the reviewers in the HRI and philosophy community who along the way have been both challenging and supporting my work.

To my friends. Thank you Søren for your tireless encouragement and support throughout the years we were both students at Aarhus University, and during the more troublesome moments in my life trying to stay on top on life and my career plans. Mirjam, your are not only my fellow PhD student, but you are also a very kind friend in the moments when life outside academia would also put its pressure on me.

To my family, thank you for your unconditional love and consistent support. Thank you Gunhild for always encouraging me to follow my own path in life, and thank you Marlene for reminding me about daily life outside academia. A special thanks to Carmen: you helped me stick through the hard times of my PhD, never give up, and see the beauty of the process. The countless hours we have spent together talking about our projects and sharing our ongoing PhD struggles have enabled me to stay focused and motivated.

Last but not least, thank you Florian. Your continuous reminder to always see the small steps of progress in my work and to practice more kindness to myself means the world to me. Without having your help and support during the last year of finalizing my PhD project and writing this dissertation, I doubt my PhD journey would have been such an enriching experience as it was on both a professional and personal level. I am looking forward to sharing more adventures with you.

List of Figures

List of Tables

Related Publications

Part of this dissertation has already been published, accepted for publication, or is currently under review. I will specify throughout my dissertation which publications support or are the result of each chapter. This short overview provides an explanation of the main focus of the publications and my specific role in the collaborative efforts.

Published

Hannibal, G., Dobrosovestnova, A. & Weiss, A. (2022) "Tolerating Untrustworthy Robots: Studying Human Vulnerability Experience within a Privacy Scenario for Trust in Robots". Proceedings of the 31st IEEE International Conference on Robot & Human Interactive Communication (pp. 821-828). Naples, Italy: IEEE. https://ieeexplore.ieee.org/document/9900830

In this conference paper, we present and discuss the results of our follow-up online HRI study aiming to explore human experience of vulnerability for trust in HRI with a special focus on a privacy scenario. As the first author, I took the lead and responsibility for writing the first full draft of the paper, as well as the parts on the quantitative data analysis and the data analysis of the semi-structured interviews, which I mainly revised for the finalized version. I received continuous supervision and practical feedback from my supervisor throughput the whole writing process, and input for discussion from my second co-author.

Hannibal, G., Rabb, N. Law, T., & Alves-Oliveira, P. (2022). Towards a Common Understanding and Vision for Theory-Grounded Human-Robot Interaction (THEORIA). Proceedings of the 17th ACM/IEEE International Conference on Human-Robot Interaction (pp. 1254–1257). Sapporo, Japan (online): IEEE. https://dl.acm.org/doi/abs/10.5555/3523760.3523996

In this co-authored workshop proposal, we descried the motivation, theme and aim of the planned workshop on theory and theorizing in HRI that was held in March, 2022. As first author, I led the writing of the proposal and was responsible for the first full draft. Though I initiated the collaborative effort, my co-authors and workshop organizers played a big role in developing the workshop through continuous feedback and input for discussion.

Hannibal, G., Weiss, A. & Charisi, V. (2021). "The robot may not notice my discomfort" – Examining the Experience of Vulnerability for Trust in Human-Robot Interaction. Proceedings of the 30th IEEE International Conference on Robot & Human Interactive Communication (pp. 704-711). Vancouver, BC (online): IEEE. `https://ieeexplore.ieee.org/document/9515513`

In this co-authored conference paper we present and discuss the results of our online HRI study aiming to explore human experience of vulnerability for trust in HRI. As the first author, I took the lead and responsibility for writing the first full draft of the paper besides the parts on the quantitative and qualitative data analysis which I mainly revised for the finalized version. I received continuous supervision and practical feedback from my co-authors throughput the whole writing process.

Dobrosovestnova, A., **Hannibal, G.** & Reinboth, T. (2021). Service Robots for Affective Labor: a Sociology of Labor Perspective. AI & Society, pp. 1-13. `https://link.springer.com/article/10.1007/s00146-021-01208-x`

In this co-authored article we present a discussion how tensions through the concepts of affective and emotional labor offer insights for the design and evaluation of professional service robots. As second author, I have been in charge mainly of the theoretical discussion about how such perspectives relate to the research field of HRI, and have provided some of the main arguments put forward. From the early drafts of the papers, I provided substantial and detailed feedback or suggestions for improvements. I have also written the section on the different approaches to emotions in HRI.

Hannibal, G. (2021). Focusing on the Vulnerabilities of Robots through Expert Interviews for Trust in Human-Robot Interaction. Proceedings of the 16th ACM/IEEE International Conference on Human-Robot Interaction (pp. 288-293). Boulder, CO (online): ACM. `https://dl.acm.org/doi/abs/10.1145/3434074.3447178`

In this late-breaking report, I present the results of eight expert interviews I conducted to identify the vulnerabilities of robots. As sole author, I led and was responsible for writing this paper in its entirety. However, I received feedback from my supervisors throughput the whole process and received some helpful suggestions for improvement on earlier drafts.

Dobrosovestnova, A. & **Hannibal, G.** (2020). Working Alongside Service Robots: Challenges to Workplace Identity Performance. Proceedings of Robophilosophy 2020 – Culturally Sustainable Social Robotics (pp. 148-157), Frontiers in Artificial Intelligence and Applications Series, Vol. 335. Aarhus, Denmark (online): IOS Press. `https://ebooks.iospress.nl/volumearticle/56358`

In this conference paper, we set out to critically reflect on the potential challenges of working alongside social service robots through the lens of workplace identity. As a second author, I was part of developing the main idea and arguments put forward and responsible for structuring the paper. Throughout the whole writing process, I provided substantial and detailed feedback and suggestions for improvements.

Dobrosovestnova, A. & **Hannibal, G.** (2020). Teachers' Disappointment: Theoretical Perspective on the Inclusion of Ambivalent Emotions in Human-Robot Interactions in Education. Proceedings of the 15th ACM/IEEE International Conference on Human-Robot Interaction (pp. 471–480). Cambridge, UK (online): ACM. `https://ieeexplore.ieee.org/document/9484256`

In this conference paper, we make a case for the consideration of ambivalent emotions for the design of social robots for tutoring. As the second author, I provided substantial and detailed feedback and suggestions for improvements for the early draft of the paper. I also contributed by writing the first draft for the sections on how the aspects of disappointment relate to the concepts of trust and vulnerability, and the discussion on the ethical considerations.

Forthcoming

Hannibal, G. & Weiss, A. (2022). Exploring the Situated Vulnerabilities of Robots For Interpersonal Trust in Human-Robot Interaction (pp. 1-19). Vienna, Austria: TU Wien Academic Press.

In this book chapter, we argue that a focus on the vulnerabilities of robots will benefit current discussions on trust in robots and their perceived trustworthiness through the proposal of the Event Approach to interpersonal trust in HRI. As a first author, I took the lead and responsibility for writing the first full draft after a joint discussion with my co-authors about the focus and structure of the chapter. I received from my second co-author substantial and detailed feedback or suggestions for improvements throughout the whole writing process.

Contents

CHAPTER 1

Introduction

> "Visions of technology, whether overly optimistic or anxiously dystopian, consistently award new technologies the capacity to transform."
>
> Sturken et al. (2004, p. 3)

Section 1.2 and 1.3 of this chapter are based on discussions with Anna Dobrosovestnova (TU Wien). Some of the perspectives on the potential challenges related to the deployment of robots into society have already been published as Dobrosovestnova and Hannibal (2020a), Dobrosovestnova and Hannibal (2020b), and Dobrosovestnova et al. (2022).

In this introductory chapter, I aim to show that the development of industrial and service robots is no exception from previous sociological diagnoses of late modernity of questioning trust because of the growing sense of risk and uncertainty resulting from scientific and technological progress. I show that although the introduction of robots more widely into society and everyday life has brought us closer to the long-standing vision of a more robot-supported society, the potential of robot coworkers and companions requires us to reflect more critically on issues of trust in human-robot interaction. In the second half, I will then shift the focus of the chapter towards a short overview of recent studies in human-robot interaction that have been focusing on two different strands to ensure human trust in robots and to develop trustworthy robots. From this general contextualization, I will then present my research objectives, methodology, and main contributions, and provide a overview of my dissertation.

1.1 Trust in Late Modernity

As Misztal (1996) very well observed, trust serves as an important backdrop for a sense of harmony and social order that only gets questioned in unruly times or when big changes in society are happening. She writes:

> "[...] it can be said that in 'settled' or 'normal' periods the issue of trust, even in the context of a culture with a low level of coherence and unity of values, is not perceived as a social problem. In transitional periods, in contrast, the role of trust as the basis for negotiations and dialogues becomes more important and more visible" (pp. 62-63).

What this quote makes very clear is that trust is mainly of public concern only when people believe that it is under pressure or even threatened, which happens when society is undergoing major change or transformations. The unsettled or transitional moments can be caused by various social, economic, technological, political, scientific, and environmental crises, and there are plenty of examples throughout the history of mankind. In times when big change and transformations of society take place, people are often left vulnerable because they have to find new ways of organizing individual and collective life that are often arranged and established by reformed institutions and new practices. In such moments of social crises, either an erosion or explosion of trust can result from the increasing sense of risk and uncertainty because it forces people to reevaluate their idea of a good, safe, and meaningful life in light of the past, present, and future (Möllering, 2006).

A basic analysis of late modernity[1] through the work by contemporary sociologists Luhmann (1979), Giddens (1990) and Beck (1992) is very useful to our

[1] Whether life in a highly globalized, fast-moving, and uncertain society have to be considered a completely new area of modernity or merely as a specific phase of the modern era, is something that sociologists disagree about. Against those who have argued that people currently live in a post-modern age (or in post-modernity) as society is not a continuation of modernity but a

understanding of how the topicality of trust in society today is closely related to more critical views on how scientific progress and technological innovation have changed our lives and created new challenges. We live in a time, they argue, where old and new relations had to be negotiated and where trust came under a lot of pressure because it was being questioned.

1.1.1 Luhmann and System Trust

The work on trust by Luhmann (1979), is in many ways a primer for a contemporary understanding of trust from a sociological perspective. In his influential writing, Luhmann not only provides the first theoretical and systematic account of trust, but also points towards ways in which trust could be properly evaluated. Drawing on the work of various philosophical and sociological thinkers, Luhmann mainly elaborates and expands upon the functional approach to the integration between social subsystems and symbolic legitimation that was first taken up by the American sociologist Parsons (Misztal, 1996). His aim was to suggest that trust is important to any system[2] as it could serve as an antidote to problems of late modernity, in which the experience of complexity and contingency were increasing and intensifying due to new forms of mass communication. Central to his treatment of trust is the idea that its main function is to reduce uncertainty in social life by encouraging people to follow a risk-taking rationality as a replacement for dangerous behavior. Luhmann (1979) states that trust in the form of interpersonal relationships, as characteristic for life in modernity, was slowly being substituted by trust in systems that promised to perform and maintain fundamental conditions for life despite the increase in choices and opportunities. This kind of trust is not motivated by an emotional response to the request of others, but rather arises from an exercise of self-presentation that serves to

fundamental break from it, there are sociologists who have rather suggested that we live in late modernity (and sometimes also called high or liquid modernity) despite the major social change that has occurred since the 1960's in the developed part of the world. It was mainly in the work of Giddens (1990) that this concept was first introduced and is by now acknowledge as one of the dominant sociological theories of modernity.

[2] To develop his system theory, Luhmann adapted to notion of autopoiesis from biology to identify the elements of a system that are reproduced by the elements of the system itself. While biological systems, he argued, reproduce themselves on the basis of life, social systems reproduce themselves through communication and psychic systems. Social systems are in this sense non-physical substances made up by elements of meaning, which appear in the different sub-types of societies, organizations and interactions (Seidl, 2004).

reflect a high level of control and ability to set boundaries. The idea of "system trust" that Luhman presents rests upon the strong belief people have in each other to trust in systems that are based on symbols of unity, which are used and considered meaningful to people through reflexivity. The consideration of "trust in trust" by Luhmann (1979) then expresses how trust is closely related to the notion of confidence, as it supports the functioning of the society because people will strongly assume that others act and behave according to the logic of the system.

1.1.2 Giddens and Unstable Trust

In contrast, in the work of Giddens (1990) we are being alerted to how trust in authority cannot be automatically presumed in late modernity, but increasingly has to be actively earned and invested. Based on earlier criticisms of scientific rationality and technological determinism (see e.g., Ellul (1980); Horkheimer and Adorno (1972)), Giddens argued that modern life had moved into a second, reflexive age of modernity characterized by an increasing sense of unpredictability, unfamiliarity and unprecedented risks caused by modern science and technology. Moreover, he recognized that modernity involves not only the destruction of traditional orders, but also reconstitutes them at a new global level. With the widespread use of telecommunication and the development of the internet together with personal computers, the new developments in late modernity have set in motion the global transformation of modern life where people are no longer limited by the spatial and temporal boundaries of traditional society. This accelerating processes of "time-space distanciation" puts trust under great pressure, Giddens (1990) explains, because the mediation of social relations through technology dissolves local contexts of interaction and knowledge exchange. People are no longer basing their trust on face-to-face encounters with other people, but rather blindly trust in abstract systems and faceless commitment emerging from the process of globalization. Trust in these anonymous and often distant systems then comes down to their credentials and legitimacy, which are demonstrated by professional codes of practice, qualifications, accreditation, licensing, performance and reputation. Especially, expert systems used in late modernity to manage and reduce natural and technological risks are causing much anxiety and doubt because the views of experts as the only legitimate pathway to knowledge and truth has been shaken.

1.1.3 Beck and False Trust

According to Beck (1992), what is central to understanding life in late modernity is the shift from concerns of natural to manufactured risks resulting from scientific knowledge and technocratic decisions, as well as a heightened sensitivity to the social and political responsibility needed to deal with such risk events. In what he refers to as the "risk society", risks are an unanticipated consequence of accumulated knowledge and unmatched success by humans to unveil, colonize, manipulate, harness and transform the powers of nature for their own benefit. Gradually blurring the boundary between natural and technological risks, Beck (1992) also sees this change as a new attitude towards risk as partly socially constructed. In late modernity, human activity is the main source of the changes we see in our environment and the risks that follow are usually not restricted to a single location and point in time. In a risk society, people have to live with the all-pervasive awareness of low probability but high consequence risks that threaten their life and well-being despite efforts to manage and regulate these risks so they do not get out of control. As such, trust is very much needed in late modernity, Beck states, because it is getting harder to predict the consequences of manufactured risk and who is to be held accountable in cases in which a preventable catastrophe is unleashed. However, the perception and evaluation of risk is no longer decided solely by experts but also in perspective to growing reflexivity of individuals who are slowly losing trust in expert systems to reduce and mitigate human-made and global risks and sense of uncertainty. According to Beck (1992), trust in late modernity between science (through the institutions and practices put in place to provide its authority or legitimacy) and the general public is fragile and weakened because people no longer consider the knowledge claims made by experts about risk assessments and predictions assuring, when trying to cope with their fears of being harmed.

1.2 A Robot-Supported Society

What Luhmann, Giddens, and Beck have in common in their sociological diagnoses of society is that they all identify scientific and technological progress as a source of great change in late modernity, by transforming both social relations and structures. While they were not specifically focusing on the new advancements in AI and robotics, they did live to witness some of the changes that such developments would have on the organization and arrangement of society. Since developments in AI and robotics are posing unique challenges to society and forcing their own questioning of trust, a sociological diagnosis of late modernity with a focus on the trust topic can help us contextualize why the search for stability has been a concern after people started to envision such technology as an integrated part of society and everyday life more generally.

Today, we live in a time where the idea of a "robotic society" (Weiss et al., 2011) is being entertained regularly. Not only from the plethora of catchy media headlines reporting or commenting on the most recent developments in AI and robotics, but also from the steady stream of movies and series depicting our possible co-existence with robots. As Richardson (2015) observed when she conducted an ethnographic study among roboticists in the US and Japan in the 1980's, the distinction between facts and fiction is illusory when it comes to robots. With her research, she was able to show that "the fictional aspects were weaved into the factual aspects of the robotic science and practices" (p. 113). Richardson explained that this is understandable, because "prior to their work as robotic scientists, they had related to robots as cultural objects before technological ones" (p. 115). This distinction between robots as either cultural or technological object is one I will be tracing for the short historical account of how in late modernity, we have been able to welcome robots at work and in our homes.

1.2.1 Robots in Narratives

The idea of artificial, intelligent, and autonomous workers who serve or take over tedious work dates back to ancient times. As Nocks (2008) writes in her detailed account of robot developments through history:

> "The first incarnation of the robot was not a machine in our sense of the term, but it was a clear expression of the idea of manufacturing machines to replace humans in laborious tasks" (pp. 3-4).

What we mean by robots today first appeared as cultural objects with symbolic meaning in grand narratives about Gods and human fortune in both Greek mythology and Jewish folklore. For instance, Homer writes about the Greek god Hephaestus creating several statues that are able to come to life. Much later, the legend of the Golem tells the story of a artificial human-like clay statue intended to serve as a defender of the Jewish community from anti-Semitic attacks, accusations, and pogroms (Nocks, 2008). In the 19th and 20th century, we also find fictional narratives in which robots as artificial beings become a method of social critic. The story about the self-learning and speaking Frankenstein monster by Shelley (2012) teaches us about the consequences of social exclusion. The Czech play *Rossum's Universal Robots (R.U.R.)* by Čapek (2004) is a comment to the possible danger of industrialization that fosters a view on human workers as a mere means of production and exploitation, forced to do difficult, monotonous and dangerous work. This play also brought the word 'robot' to our language for the first time, and still symbolizes a classical dystopian narrative about violent conflicts between rebellious humanoid workers and their creator. In several short stories, Asimov (1995) presents a more nuanced perspective on what future co-existence between humans and robots might look like. Asimov developed robot characters with sophisticated personalities, and came up with the Three Laws of Robotics as a way to explore some of the moral dimensions of possible robot team members or companions.

Myths and stories like these represent how humans have always been imagining the possibility of creating artificial and human-like figures that would purposefully be developed to lift the burden of work in the role of servants, or even as a new forms of co-workers helping in various tasks. However, it is only in the modern age that it is has been possible to realize robots as actual workable machines good enough to take over human labor, given refined innovation and progress in science and technology.

1.2.2 Robots for Production

Resulting from efforts to create integrated circuitry and electronic computers, as well as advancing automation throughout the 1930 and 1940's, the first industrial robot was developed by George Devol in 1954[3]. Together with Joseph Engelberger, he founded the first commercial robotics company in 1961. Their first Unimate #001 robot prototype was sold to General Motors (GM), and operated as a die casting machine in their factory. Various car companies in Europe also recognized the great potential of industrial robots, and by the 1970's, they could be seen

[3] Automation played a key role during the second World War, where it was applied in fighter airplanes, landing crafts, warships, and tanks (Westerlund, 2000).

at BMW, Volvo, Mercedes Benz, British Leyland, and Fiat carrying out simple manufacturing tasks (Westerlund, 2000).

The generation of industrial robots developed during the 1970's performed very well at a number of industrial tasks, as long as the parts being manipulated were positioned in exactly the right place, and the parameters were under automatic control. While able to deal with monotonous and repetitive tasks in production, they were rather primitive, with no external capacity to sense and react to their environment in any significant way. Their control systems consisted of Programmable Logic Controllers (PLC) or were programmed by an operator by means of a teach box. They were typically programmed by recording each task as a series of points in space, and simply replayed whenever the task was to be performed. Because each robot had its own running program that was dedicated to a specific task, they were lacking any self-adaptive behavior. Furthermore, these robots were bolted to the floor or to a tabletop, making them completely immobile. As such, these industrial robots were in fact too slow and crude to effectively compete with human labor in assembly (Gasparetto and Scalera, 2019).

Determined to make these robots able to perform more complex tasks and manage more unstructured, constantly changing environments, billions of dollars were invested in the 1980's into robotics companies in the US and Japan to such an extent that robotics was celebrated as the "next industrial revolution" (Vincent, 1999, p. xv). During this time, much progress in robotics was achieved as robot engineers and manufacturers shifted their focus towards electric actuators, incorporating advanced sensors, improving robot controls and synchronization, as well as implementing rudimentary machine vision systems to detect and follow moving objects. Moreover, robots were also made far more versatile as the development of programming languages and better control interfaces enabled higher levels of adaptive and self-programming capabilities (Gasparetto and Scalera, 2019). This new generation of industrial robots was much more suitable for complex tasks and was considered capable of some low-level "intelligence" as they used "the data coming from vision or perception systems to locate the objects and guide the joint movements according to the task to be performed, taking into account the possibility of small changes in the position of the objects" (Gasparetto and Scalera, 2019, p. 31).

Despite the hype and willingness to buy industrial robots, many robotics companies had difficulties meeting the demand because their products were still too expensive for most potential buyers and there was also a lack of knowledge about how to implement such robots effectively. Consequently, a lot of the robotics companies vanished or left the field in the middle of the 1980's and only by the beginning of the 1990's new hope in robotics was regained with the prospect of improving their robots through major advancements in the software development (Vincent, 1999). Since the turn of the 21st century, efforts in developing industrial robots have been directed to refine the design of innovative kinematic structures and enable them to acquire "high-level 'intelligent' capabilities (such as performing advanced computations, logical reasoning, deep learning, complex strategies, collaborative behavior)" (Gasparetto and Scalera, 2019, p. 33).

1.2.3 Robots for Communication

While much of the research on developing industrial robots was dedicated to improving their ability to solve more complex tasks required for the assembly line, other researchers were focusing on ensuring that robots could navigate in human environments and interact intelligently with people. With the reduced cost of computer hardware, interest in making robots commercially available motivated the development of robots to serve the general public in the role of assistants, peers, and companions[4].

The goal of developing robots that could potentially do more than "dirty, dangerous, and dull" (Takayama et al., 2008) work required a more "intuitive" interaction between robots and humans, as many people outside factories would have very little knowledge about how they functioned. This led to a strong interest and focus on building human-like communication channels into robots. This was mainly done through the implementation of basic social principles adapted from work on human-human interaction in developmental psychology (Dautenhahn, 1995). The central idea was that robots should now imitate and learn through social cues and be able to adjust their behavior also through a form of emotional intelligence. Drawing on concepts of embodied and situated awareness in robotics by Brooks (1999), Breazeal and Scassellati (2002) developed the socially intelligent robot KISMET that was capable of responding to its human interac-

[4] I find it helpful to provide some specific examples of robots that have been developed for communication by considering the most widely known robots in the current HRI literature, though my own personal selection is not an exhaustive list.

tion partner with different types of facial expressions and simple vocal cues to indicate an emotion (e.g., sadness, happiness, disgust, tiredness, calmness, interest, surprise, fear, and anger). Equipped with a motivation and behavioral system, KISMET was able to learn from interacting and communicating with humans. Similarly, the socially intelligent robot LEONARDO was the first robot to demonstrate the implementation of a computational model that utilized the basic principle of "Theory of Mind" to recognize and account for the different perceptions and intentions of people in collaborative tasks and social game-play (Brooks et al., 2004). Resulting from their ambitious research project starting back in 1986 with their series of bipedal humanoid robots (P1, P2, and P3), the world famous ASIMO robot developed by Honda was presented to the public in 2000. It was the first bipedal walking robot capable of recognizing faces and speech, and able to help people with service tasks such as handing over a tray, pushing a cart, and pouring a drink (Hirose and Ogawa, 2007). Socially intelligent robots were also considered as tools to provide therapy for children on the Autism Spectrum (Dautenhahn et al., 2009) or to help post-stroke patients with rehabilitation exercises (Matarić et al., 2007).

For more commercial use, robots were first presented as products for the entertainment market and they were not equipped with any advanced AI. As Stone (2005) explain, simple toy robots (e.g., FURBY by Tiger Electronics, MINDSTORMS by Lego, and AIBO by Sony) became very popular among younger and older children, who were fast in adopting them for play and educational purposes. With their interest in using robots for healthcare, Wada et al. (2005) developed and promoted in 2004 the PARO robot as a form of companionship robot specifically to the elderly in Japan, although it was later also tested and deployed in various retirement homes in Europe (e.g., Germany, Italy and Denmark). Aiming to develop more advanced commercial robots that were able to utilize AI advancements, the NAO Robot developed by Aldebaran Robotics in 2008 (acquired by Softbank Robotics in 2015) was developed for educational purpose as a method to get school children more engaged in STEM subjects (Bertel and Hannibal, 2015). In 2014 the very same commercial robot company issued the PEPPER robot, which was mainly intended for consumer service or support in smaller retail businesses, as well as for domestic use.

1.3 Trust Demand

In light of recent developments in AI and robotics that make robots increasingly capable, it is important to ask if these advancements are also contributing to increasing questioning of trust in late modernity, as they add to the growing sense of risk and uncertainty that eventually renders people more vulnerable. There is good reason to believe this is the case, and that this relates to two different but interrelated consequences of introducing robots into society and everyday life more broadly. First, the application of robots for both manual and emotional labor leading to increasing replacement of human workers. Secondly, the prospect of more regular encounters with robots in care and service has given rise to worries about whether they could eventually substitute human contact.

1.3.1 Replacing Human Workers

In the early days, the automation of production with industrial robots was met with optimism as the economy became more efficient and expanded, while people benefited correspondingly from better pay. Moreover, automation of work also led to new industries where far more people were hired than lost their jobs. However, this changed when new scientific and technological advancements improved the capacities of industrial robots, enabling larger-scale production with a smaller work force – a trend we see continuing today. From the mid-1970's, industrial robots become competitive enough to start outweighing the costs of human workers performing the same task. Consequently, the gap between pay and productivity began to increase: hourly compensation fell and stagnated while productivity kept rising. For the average factory worker, this was alarming, and soon the attitude towards automation became a source of worry and pessimism. As human workers lagged further and further behind, the social impact of industrial robots upon the labor market started to show in the 1990's as the fear of unemployment came to overshadow the benefit of automatization. Predictions of mass unemployment were also based on rapid changes in the organization of work given new forms of information- and telecommunication in office jobs and service industries. Both blue-collar and white-collar workers were at high risk of losing their jobs because employers simply considered their replacement the best strategy to reduce costs and improve profit performance through lean production and streamlined management. These drastic changes created a polarization and social disintegration of society: many of the less advantaged were left to their own devices in adapting to the new order established by a labor market that now demanded increasingly skilled workers to look after or work alongside robots (Rifkin, 1995).

Although many technological issues still stand in the way of developing fully functional robot assistants, there are now also many discussion about whether they will become advanced and versatile enough to take over work that was previously reserved for humans because of their mental and social capabilities. That is, with the ability to socially interact with people, robots could also take over emotional labor and consequently may have the potential to one day replace care givers, teachers, therapists and sales assistants. Not yet capable of providing adequate physical support, and still struggling with basic social skills for truly long-term care, support and service, such robots are currently marketed as mere supplement to various existing care and service work. They are mainly used to entertain and keep people company when their carers are too busy or tired to meet their emotional needs. Outsourcing emotional labor to robots has been envisioned as a way to ensure the comfort of sick or elderly people given a shortage of work force and the continuous demand of higher productivity with less time and resources. Because the attitude towards using robots for such work often differs between management and employees, many people have started to feel worried and unsure about what impact robots might have on their employment or work conditions (Tuisku et al., 2019): managers tend to have a very positive view on the introduction of socially intelligent robots, which they see as a new and innovative development that could modernize and renew their centers, facilities or businesses. As managers are mainly in charge of the organization of the workplace, their interest in bringing robots into their workforce has increased in connection with attempts to address increasing labor costs and a shortage of workers willing to do care and service work. Especially in the health and eldercare sector, it is sometimes even argued that technological solutions are required to address “demographic pressures in countries like Japan, Germany, and the USA to ensure the ability to care for a rapidly expanding population of aged or otherwise vulnerable persons, economic pressures upon individuals, private or public institutions to reduce the costs of care, social pressures to reduce growing institutional failures to provide quality human care, and recognition of the need to reduce the often-overwhelming physical and psychological burdens placed upon individual caregivers" (Vallor, 2011, p. 252). In the various service sectors, robots have mainly been introduced to increase profit by reducing the cost of labor and to increase customer satisfaction. In addition, the agenda to have fully autonomous robot-driven services receives much public attention through media coverage due to their promise of being the future of hospitality..

When people feel at risk and uncertain about whether their labor will continue to be needed as automation and robots continue to advance, they may feel like they can no longer take for granted a social order that ensures a basic income for living. This also means that in order not to lose "the race against the machine" (Brynjolfsson and McAfee, 2011, p.21), people are forced to not only reconsider their understanding of work, but also to consider how they can keep up with a demanding labor market that is quickly adopting faster and cheaper technological solutions. While the workers deemed redundant in the future high-tech workforce face unemployment, those still left will have to increasingly ask themselves if they can trust their new robot co-workers or so-called "co-bots" (Fast-Berglund and Romero, 2019). Since trust is crucial for successful collaboration, a robot-supported society where people are working alongside or together with robots has to ensure that people are willing to trust robots to carry out the required task on a daily basis, without fearing that their willingness to join forces could result in loosing their own job.

1.3.2 Substituting Human Contact

Beside worries that it may only be a matter of time before there will no longer be any work left for humans, there have also been concerns about whether advancements in robotics could have a serious impact on interpersonal relationships and access to sufficient human contact. This can be problematic, especially when the deployment of robots for emotional labor means that both children and elderly people are left with robot companions instead of the care or attention of their parents, friends or care givers. The possibility of robot companions that can offer friendship and companionship in a robot-supported society has raised the both social and ethical concern that "we will view robotic contact as a substitute for human contact and we will lose out on important human and social goods" (Danaher, 2019, p. 8).

Since the 1990's, Turkle (2011) has been studying how children and elderly people respond and relate to robots developed to keep them company. She refers to “the robotic moment” as a form of mind-shift where people express being at ease with the idea of substituting human connection with the comfort of robots. After observing and talking to children interacting with both simple robot toys and more sophisticated robots as if they were real creatures that could be befriended, she realized how quickly they get attached to robots and call for their attention. Turkle argues that these robot companions are problematic because they offer children the comfort of friendship without the need for reciprocity that regular contact with other people requires. She worries that allowing children to substitute human contact with the company of robots will lead to social deskilling or unauthentic friendship that will eventually make them addicted to “shallow, utilitarian and pleasure-seeking interactions” (Danaher, 2019, 20).

Interested in how the substitution of human contact affects the elderly, Turkle (2011) reports on how people in eldercare homes grow an attachment to robots that can provide emotional comfort. Again, she paints a gloomy picture in which the elderly left with a companion robot become isolated rather than cared for. They turn to robots, she explains, because they seem to listen and never get tired or impatient with them. Over time, robots that are only intended as tools to fill in the gaps when care givers are busy become their escape from emotional distress when feeling lonely. Others have also expressed great concerns that the replacement of care and service workers risk that people may no longer receive the human contact they need in order to develop properly as human beings and maintain their well-being. The work of Sharkey and Sharkey (2010) has been leading more philosophical discussions about the ethical issues of developing robot nannies. They argue that the lack of human contact could have significant psychological consequences for children and affect their well-being because robots are not capable of meeting their emotional needs. Sparrow and Sparrow (2006) even argue that the replacement of care workers is ethically misguided because human touch is essential to caring for the elderly. Their worries focus on how the elderly deprived from such contact are at risk of being objectified and becoming even more isolated from other people than they may already be.

Expecting this tendency to continue, there are growing concerns about the risk and uncertainties of what the consequences might be of replacing “the human touch” of care and service workers with the cold hands of machines. From the examples in the ethical debates, children and elderly might end up over-trusting so-called robot companions on the basis of forming strong emotional attachments regardless of whether the employment of these robots is by their own choice or as a consequence of human staff having too little time and resources. In this sense, one of the problematic aspect of substituting human contact with robots for the purpose of caring or providing service to children and elderly is that the kind of trust needed to ensure reciprocal and caring social relationships goes beyond what they are considered capable off. How willing people are in opening up their homes and hearts to robots from this point of view might suggest that the affective dimension of trust in meeting ones own emotional needs is the primary motivation when interaction with other humans is unattainable. Thus, anxieties and worries about using robots for care and as service workers relate to whether people would and even should trust these robots enough to share their personal life with them despite the fact that they are incapable of doing their intended job properly and to an acceptable standard, because robots may never be good enough for truly long-term care, support and service..

1.3.3 Are People Left Vulnerable?

Substituting human contact with that of robot co-workers and companions resulting from the replacement of workers for both manual and emotional labor have led to warnings about the potential ethical consequences of developing robot co-workers and companions.

The ethical concerns related to having robot co-workers in the workplace tend to focus on the basic question of whether people will be able to thrive among their new kinds of colleagues, and what kind of expectations people can have in terms of their conduct as part of a team. From their analysis of whether the introduction of robots into the workplace is to be considered a threat or an opportunity, Smids et al. (2020) discussed some problems that cause ethical concerns and if they could potentially have a negative impact on the human experience of meaningful work. Having robot co-workers, they argue, is problematic if they take over too many or the most challenging tasks: for instance, people may no longer find any purpose in their work, people may face reduced chances for social interaction at work or experience substantial change to their team dynamic, robots could make the acquired skills and competences that people use for their work obsolete, robots might keep people from developing their self-esteem or receiving social recognition from doing their work well, robots may deprive people from exercising their own judgment or autonomy when carrying out their work. Smids et al. (2020)

also discuss how the workplace itself will change along with the increasing use of robot co-workers, so that people might no longer feel like their work environment is optimal for their own needs and comfort – the infrastructure required for robots to navigate and operate well in a human workplace environment is considerable and comes with many constraints (e.g., people might like to personalize their office with ornaments that would be considered obstacles for robots or they might be required to have a certain lighting that that is straining to human eyes, but ensures that the robot is able to properly carry out its task). Beside these more general ethical concerns, the question asked by Nyholm and Smids (2020) about whether robot co-workers have the potential to be good colleagues is also important to this discussion. As they explain, for most people, satisfying and meaningful work consists not only of getting the job done successfully, but also in working together with good colleagues in a team. If robot co-workers developed to work alongside people cannot live up to the ideal of a good colleague, Nyholm and Smids (2020) argue, it is very likely that people will struggle on both social and psychological levels. Interestingly, reliability and trustworthiness were listed among the 10 different common sense criteria that they suggested are essential for a basic understanding of a good colleague. While they find it very plausible that robot co-workers could live up to the criterion of being a reliable colleague in terms of their task performance, the aspect of trustworthiness is much more challenging. In line with more pessimistic arguments for why robot co-workers cannot be trusted similarly to how people trust each others in teams, Nyholm and Smids (2020) agree that trustworthiness is a much more demanding (though overlapping) criterion. However, they are more optimistic about this possibility since the growing interest in empirically studying the acceptance of robot co-workers' work practices has helped shed new light on this question (see e.g., the work by Weiss et al., 2021).

Ethical concerns about the development of robot companions have gained much more attention in the philosophical literature because the challenge to be addressed no longer consists of simply ensuring mechanical and bodily safety, but ensuring that no harm is being done to people looking for entertainment, comfort or even friendship in their interaction with robots. Especially the extensive and deliberate use of the anthropomorphic design strategy to make companion robots more socially capable has lead to major worries about the possible danger of deception. Sharkey and Sharkey (2021) lay out two arguments for why the development of robot companions involves deception regardless of developer intentions, and why it is problematic to both the individual and society. First of all, as Sharkey and Sharkey (2021) argue, robot companions can deceive people by making them believe that they have emotions. The negative impact of this emotional deception on people stems from their attachment to robot companions that they believe care for them, or that they believe are capable of caring for others. Children and the elderly are particularly in danger of emotional deception by robot companions, because they are often in much need of care and not always able to understand what is behind the appearance, making it harder for them to protect themselves from neglect or exploitation. Secondly, Sharkey and Sharkey (2021) argue that the development of robot companions could have a harmful impact on society when they are delegated tasks or decisions which they are not qualified for because people overestimate their capabilities. From this point, Sharkey and Sharkey (2021) assert that it is much more likely that people will misplace trust in robot companions when they are left with a false impression about their abilities or competences. Accordingly, they call for a deliberation about who is ultimately responsible for wrongful deception in the development and deployment of robot companions by also holding the people behind the production of robots more accountable. In contrast, from his categorization of three different high-level forms of robot deception, Danaher (2020) states that there is something morally wrong in developing robot companions that use various superficial social or communicative cues to hide or cover up capacities they in fact posses. He even considers this form of robot deception a type of betrayal, because robots misleading or concealing their capabilities can severely undermine the relationships people hope to establish when they place their trust in them as companions. The ways in which robot deception as betrayal can be dealt with on a practical and legal level is an ethical concern that Danaher (2020) thinks is important to address in the pursuit of integrating robot companions in the everyday life of people[5].

From this brief presentation of some of the most pressing ethical concerns related to the development and more widespread use of robot co-workers and companions, it is clear that their introduction will instigate many new questions about how to evaluate or asses on a societal level whether they will eventually be

[5] For more comprehensive and detailed discussion about the ethical and legal aspects of robot deception see e.g., Matthias (2015); Sætra (2021).

an advantage or a disadvantage. While the benefits of robot co-workers and companions motivate further research in robotics, the strongest opponents in ethical debates claim that people are becoming increasingly vulnerable in the hands of machines. People finding themselves working alongside more and more robot co-workers are asked to place their trust in them in order to be successful at work or find their work meaningful, while people looking towards robot companions for support and care have to trust in them being aware of the potential of harmful deception. Will technological and scientific advancements lead to people being able to trust in their future robot co-workers and companions or consider them trustworthy? Our vision of a fully robot-supported society might come at too high a risk, with too much uncertainty of the long-term consequences. Given these concerns, efforts to strengthen trust in robots and their trustworthiness as a strategy for better diffusion of robots in society and everyday life are not only indicated, but also urgent.

1.4 Trust in Human-Robot Interaction

Research on trust in human-robot interaction (HRI) has started to emerge as a reply to these specific concerns of whether we can or should trust our new robot co-workers or companions. The practical value of studying trust in HRI rests on the assumption that people will in the long-term perspective be more likely to accept, interact, collaborate, and engage with robots that they trust or consider trustworthy. Yet, trust is not easy to achieve and maintain. It can be very fragile and, when violated, recovery can take a long time. Sometimes, the loss of trust will never be regained (Baier, 1986).

Previous work on trust in automated systems (Hoffman et al., 2013; Lee and See, 2004; Parasuraman and Riley, 1997) has been the primary to guide HRI studies on human trust in robots and the design of trustworthy robots, whether in dyadic interactions or in teams. However, the physical embodiment of robots and their usage for a wider range of tasks in more unstructured environments means that the factors relevant to studying trust in HRI pose additional and new challenges (Desai et al., 2009; Kessler et al., 2017).

As with regular human-human interaction, trust in HRI can be very complicated, and the avoidance of inappropriate levels of trust is argued as essential for ensuring that people do not have negative experiences in their collaboration, interaction and engagement with robots. Well-calibrated trust is important for co-existence with robot co-workers and companions (Lewis et al., 2018).

1.4.1 Appropriate Level of Trust

Kok and Soh (2020) rightfully point out that it is very important to pay careful attention to the distinction between the notion of trust and that of trustworthiness. They characterize the distinction as lying between trust as either a property of the human or the robot[6]. In their view, trust is a human property because it is dependent on human attitude towards robots, while trustworthiness is a property of robot that is independent of such attitude because it is determined by more general standards. Consequently, it is possible to end up *trusting an untrustworthy robot* or *not trusting a trustworthy robot* because these two notions are conceptually self-sufficient. To give a concrete example, people might be in a situation where they have to trust a service robot to help them find their way through a busy airport in time for take-off. If people trust the robot, which then turns out to be untrustworthy, they might head in the wrong direction. If people do not trust the robot, which turns out to be trustworthy, they might get lost. In both cases, they could miss their flight.

Taking this potential discrepancy into account is of great importance because trusting an untrustworthy robot can lead to over-trust (and eventually misuse), and not trusting a trustworthy robot can lead to under-trust (and eventually disuse). Both directions of inappropriate level of trust are problematic because they result in unhealthy trust relations between humans and robots (Kok and Soh, 2020). From this perspective, any concerns about whether people are left vulnerable when interacting, collaborating, and engaging with their future robot co-workers and companions depends to a large extent on whether or not they only trust in trustworthy robots. As such, trust is not merely to be maximized, but first of all be appropriate or well-calibrated. Current research on trust in HRI ultimately serves this goal of striking the right balance by either determining what factors influence human trust in robots, or by improving the technical aspects of what goes into developing trustworthy robots.

[6] I rather consider it a distinction between trust as an attitude and trustworthiness as a property, no matter the kind of agent in question. This is a point I will return to and elaborate on in chapter 2.

From their meta-analysis of various empirical studies on trust in HRI, Hancock et al. (2011) and later Hancock et al. (2020) categorize the factors that influence or regulate trust between humans and robots into those that are mainly human-related, robot-related, and environmental or contextual. While it can be discussed if this specific division is the most accurate one or not, their meta-analysis does help showing that there are many different antecedents and factors to consider when studying or evaluating trust in HRI.

1.4.2 Trust in Robots

It would take an extensive review to make an exhaustive list of all the factors so far studied in trust in HRI. For now, it is relevant to gain a sense of the broad scope of factors that have been studied empirically using different methodologies and measurements.

The main focus has been on identifying the various factors influencing the willingness of people to trust robots given their performance. Studies include questions about *reliability* (Desai et al., 2012), *capability* (Xie et al., 2019), *timing and magnitude of errors* (Rossi et al., 2017), *error types* (Flook et al., 2019), *failure and feedback* (Desai et al., 2013), *behavior style* (van den Brule et al., 2014), *approach to risk* (Bridgwater et al., 2020), and *vulnerability expression* (Sebo et al., 2018).

Another major strand in current HRI research is concerned with how the physical appearance of robots plays a role in the way people would consider trusting robots. Studies have been looking into aspects of *anthropomorphism* (Natarajan and Gombolay, 2020), *forms* (Schaefer et al., 2012), *embodiment* (Reig et al., 2019; van Maris et al., 2017), *lifelike impression* (Haring et al., 2013), *robot gaze* (Stanton and Stevens, 2017), *gender* (Bryant et al., 2020), and *expressive communication* (Hamacher et al., 2016).

Other factors that have been considered when exploring human trust in robots concern *framing* (Cameron et al., 2015; Washburn et al., 2019), *time* (Kaplan et al., 2021), *transparency* (Sanders et al., 2014), *sense of control* (Ullman and Malle, 2017), *prior experience* Sanders et al. (2017), *personality* (Salem et al., 2015b), *interaction scenarios* (Hannibal et al., 2021), *adaptability* (Fischer et al., 2018), *conformity* (Salomons et al., 2018), *social engineering* (Aroyo et al., 2018), *fault justification* (Correia et al., 2018), *intergroup bias* (Deligianis et al., 2017), *explainability* (Zhu and Williams, 2020), and *use choice* (Sanders et al., 2019).

1.4.3 Trustworthy Robots

When aiming to develop trustworthy robots, the challenge is to figure out how estimation of trust during HRI can be used to help guiding actions or decisions of robots for safe interaction and collaboration. This work requires developing a model of trust for robots that can also be implemented.

By using human trust as a latent variable to develop trustworthy robots, different strategies have been proposed which take on a *probabilistic graphical* approach (Xu and Dudek, 2015), a *Theory of Mind* approach (Vinanzi et al., 2019), a *relational-trust* approach (Ono et al., 2015), *contextual* approach (Abate et al., 2020), a *mutual trust* approach (Wang et al., 2014), a *Bayesian inference* approach (Fooladi Mahani et al., 2020; Guo et al., 2020), a *case-based reasoning* approach (Floyd et al., 2015), and a *trust-seeking* approach (Xu and Dudek, 2016).

These different models of trust for robots are intended to solve different challenges. Here, the focus has so far been on *motion planning* (Chen et al., 2018; Spencer et al., 2016), *collaboration* (Sadrfaridpour et al., 2016), *explanation generation* (Edmonds et al., 2019; Wagner and Robinette, 2021), *handover* (Walker et al., 2015), *long-term interaction* (De Visser et al., 2020), and *multi-task settings* (Soh et al., 2020).

1.5 PhD Project

After this broader contextualization of why research into trust in HRI can be considered more urgent than ever and how it has been picked up in recent work, I now shift the focus towards an account of how my PhD project has been developed and organized to engage and contribute new insights to such discussions.

1.5.1 Research Context

My PhD project was mainly located in the *Trust Robots Doctoral College* at TU Wien, which was co-lead by Prof. Markus Vincze and Prof. Sabine Köszegi. The aim of the TRDC was to provide a comprehensive analysis of trust in the application context of AI and autonomous robots. Unique to this doctoral college, the organizers invited early-stage researchers with various disciplinary backgrounds to join the college to share and learn from an interdisciplinary research collaboration on “trusting robots – trust in robots”[7]. After submitting my application, I was one of the selected 10 PhD candidates, and I was encouraged from the start to draw on my disciplinary background in philosophy and sociology when developing my PhD project. With some experience already working on issues in HRI from these two perspectives, I did not hesitate to take up this special opportunity because it was also important to my own development as an independent early-stage researcher. I wanted to to be part of this doctoral college because it gave me a research context where I could engage with research on trust in HRI without neglecting my knowledge and skills gained from my training as a philosopher and experience with sociological theory and methods. With the support and guidance by my main supervisor (Dr. Astrid Weiss), I found my own and unique way of bringing the insights from philosophy and sociology on the topic of trust into dialogue with discussions in the HRI community about how it is possible to understand, analyze, and study trust in HRI.

My PhD project was also closely connected to the *Human Behaviour and Machine Intelligence* (HUMAINT) project within the Joint Research Centre (JRC) at the EU Commission, which is still running. Overall, the HUMAINT project aims to explore the potential impact of machine intelligence on human behavior through an multidisciplinary understanding of human cognitive and socio-emotional capabilities and decision making. The HUMAINT project has a core team of experienced researchers and also establish a community of experts to support the project with their expertise in cognitive science, machine learning, human-computer interaction, and economics[8]. My own role as an expert in the HUMAINT project related specifically to the collaborative work I undertook with Dr. Astrid Weiss (TU Wien) and Dr. Vicky Charisi (JRC) where we through a series of online studies studied trust in HRI with a focus on the aspects of vulnerability and benevolence. This was a great opportunity for me to use again my disciplinary background in philosophy and my experience with more human-centered HRI research to, which was the valuable contribution I brought into our research collaboration.

Mentioning the research context is important to me, because it helps clarifying why my PhD project should be considered a hybrid between research in HRI and Philosophy. By hybrid, I simply mean that it might seem very theory-heavy

[7] More information about the TRDC can be found on the website (accessed on May 16th, 2022): `http://trustrobots.acin.tuwien.ac.at`

[8] More information about the HUMAINT project can be found on the website (accessed on May 16th, 2022): `https://ec.europa.eu/jrc/communities/en/community/humaint`

compared to a standard HRI dissertation while also atypical as a philosophy dissertation due to the added empirical work. I believe that this will shine through when reading the dissertation, as some chapters are purely theoretical (i.e., 2 and 3), while others are mainly reporting results from studies (i.e., 4, 5, and 6). While working on my PhD project, as well as while writing up this dissertation, I was always reflecting how to present such work that is in-between, which does not come without serious challenges. However, after presenting the aim and approach of my PhD project in the following sections, I hope that it will be apparent why such hybrid format is necessary for my dissertation. Better still, I might manage to show that it is exactly because of its hybrid format that my PhD project adds new perspectives and valuable insights into current discussions on trust in HRI. As such, I hope not only to contribute to the research field of HRI by suggesting how to demarcate the phenomenon of trust, but also to spark interest into the topic of trust in HRI among those coming from the field of philosophy by challenging their application of the concept of (interpersonal) trust to the case of robots that appear to have agency.

1.5.2 Research Aim

Trust in HRI is not easy to understand or achieve without diving into heavy theoretical and methodological debates. It requires not only a sensitivity to the nuances in the concept of trust, but also an awareness about the many ways it can be studied empirically as it unfolds in the interactions between humans and robots. Deliberately developing and designing robots with apparent agency adds an extra layer of complexity to our understanding, analysis, and studies on trust in HRI. The general aim of my PhD project was to bring the many important insights from the fields of philosophy and sociology into current research on trust in HRI in order to improve the way interpersonal trust between humans and robots is understood, analyzed, and studied. More specifically, I aimed to show *how a focus on the specific relation between trust and vulnerability is paramount for the advancement of research on trust in HRI, because it enables us to better grasp and evaluate to what extent the phenomenon of trust in interactions between humans is similar to, or differs from, the way trust plays out between people and robots*. By suggesting vulnerability as a necessary and active precondition of trust in HRI, I strive to lay the groundwork for the inclusion of uncomfortable experiences, emotions and encounters that will undoubtedly follow a stronger diffusion of robots into human everyday life. While I put much effort into presenting the theoretical perspectives required in order to bring serious attention to the trust-vulnerability relation, it was also crucial for my aim that my investigation into this relation for the context of HRI be supported by empirical work to provide some tangible evidence. My aim is to show that the work I present in this dissertation required a recursive exchange between a theoretical perspective and empirical work, which pushed my PhD project forward and places it firmly as a novel and highly relevant contribution to audience hailing from both HRI and Philosophy.

After a lengthy and careful identification of potential research gaps in the current literature on trust in HRI, my particular focus on the relation between trust and vulnerability has turned out to be rather unexplored terrain. However, my aim was not only to fill in this research gap, but also to show why my investigation into this trust-vulnerability relation in fact facilitates deeper understanding, better analysis, and more careful studies of trust in HRI. With a few exceptions (Martelaro et al., 2016; Sebo et al., 2019; Traeger et al., 2020), very little knowledge has so far been provided to the HRI community about how the interplay between trust and vulnerability as one of its preconditions actually influences our understanding, analysis, and studies of trust in HRI. With my PhD project and this dissertation, I hope to show how I have gained some of this important knowledge.

1.5.3 Research Questions

With my PhD project, I set out to address the following research question (RQ), which also serves as an anchor to my investigation into the topic of trust for HRI:

> **RQ:** *How can an investigation into the relation between trust and vulnerability advance current understanding and analysis of trust in human-robot interaction?*

In order to provide an adequate answer, it is helpful to break down the general research question into several sub-questions, which either serve to unpack the central aspects, or point to some of the underlying challenges at hand:

> **(1)** *How has trust been conceptualized in current HRI research?*

To ensure that my PhD project is situated in cutting-edge discussions on trust in HRI, it is important to first gain an overview of how the concept of trust has been considered and studied so far. This requires not only an understanding of the various conceptual components and relationships that exist within this concept, but also insights connected to the usage of trust in the context of HRI specifically. Drawing mainly on the vast philosophical literature on trust, I will dive into how the usage of the concept of trust in the context of HRI is both supported by our current conceptual knowledge, and simultaneously challenges it. Hence, with this first sub-question, I aim to understand how and to what extent the trust concept has been understood and explained both outside and inside the HRI context to evaluate how the discussion is currently framed, and whether it needs some re-framing. In chapter 2, I present the work I undertook to address this sub-question.

(2) *What is the conceptual relationship between trust and vulnerability?*

With my second sub-question, I shift the attention to the conceptual structure of trust and present how exactly it relates to the notion of vulnerability. This work will consist of a basic introduction to the nature and epistemology of trust, and proceed to discussing how it can be useful for application in the HRI context. My investigation into the trust-vulnerability conceptual relation is important because it is my intention to unfold and explain in more detail the remarkable position of vulnerability with regards to the concept of trust. I will mainly draw on literature from sociology and philosophy of technology to explain how the notion of vulnerability has been considered outside the field of HRI, and which conceptualization would be useful for guiding empirical studies on trust in HRI. In chapter 3, I present the work I undertook to address this sub-question.

(3) *What are the relevant differences between the vulnerability of humans and robots?*

For empirical studies on trust in HRI to start including the notion of vulnerability as a crucial conceptual aspect, it is necessary to be very clear about whether the meaning is consistently the same, or whether it changes when dealing with either the human or the robot side of the interaction. Although I explore the trust-vulnerability conceptual relation on a theoretical level, it has to be interpreted and used differently when conducting empirical work, because the vulnerabilities of humans and robots cannot be studied with the same methods. Consequently, the results and conclusions that can be drawn about the vulnerabilities of humans and robots will vary substantially and are in principle going in two different directions, even though the notion of vulnerability is a joint starting point for the empirical work. I present the work I undertook to address this sub-question in the chapters 4,5, and 6.

> **(4)** *How can vulnerability be studied empirically in relation to trust in HRI?*

Given that humans and robots are on an ontological level of very different kinds, an account of how the application of vulnerability is to be considered also needs further exploration for methodological discussions and the choice of methods. With my fourth sub-question, I point to the fact that a useful operationalization of vulnerability will need to be provided to support empirical studies on human vulnerability for trust in HRI. When talking about robots as being potentially vulnerable, the challenge is to ensure that the vulnerability notion even makes sense in such a context at all, which would also be part of what the empirical study itself aims to explore. I present the work I undertook to address this sub-question in the chapters 4,5, and 6.

> **(5)** *Assuming the design of trustworthy robots to be desirable, how can a focus on vulnerability for trust in HRI contribute to this aim?*

With theoretical perspectives to direct our attention to the trust-vulnerability conceptual relation, and empirical work to explore how it can be studied in the context of HRI, the last and fifth sub-question aims to guide the discussion towards a reflection on how this knowledge can be used for the design and development of robots that people will trust or consider trustworthy. Understanding the potential and limitations of the knowledge gained in assisting engineers in their development practice requires a deeper discussion about the transfer and integration of knowledge, which I consider by drawing on my interest in the philosophy of science. In chapter 7, I present the work I undertook to address this sub-question.

1.5.4 Research Approach

My PhD project has been strongly *theory-driven*, using a *multidisciplinary methodology* which provides an approach to the understanding, analysis, and study of trust in HRI that is rather unusual for a PhD in this field of research. This specific choice of research approach was motivated by my observation that there is currently a lack of deeper theoretical knowledge about the concept of trust within the HRI community, and that much work so far tends to neglect the inclusion of more philosophical and sociological perspectives.

While empirical studies on trust in HRI often begin in theoretical perspectives, the results are hardly ever fed back into the discussions from which they stem (e.g., philosophy, sociology, economics, psychology, and cognitive science). I have chosen a theory-driven rather than data-driven research approach because the very abstract nature of trust forces a clear and well-developed theoretical understanding to guide empirical work. Throughout my PhD project, I have gained a rich understanding of the theoretical assumptions and underlying commitments of the various concepts of trust used in current HRI research before conducting my own empirical work. At the same time, I made it a priority to ensure that any findings from the empirical work are always connected to the broader discussions about how to study or investigate the phenomenon of trust as it unfolds in the interactions between humans and robots.

My PhD project is multidisciplinary at its core, as I have been working in the intersection of philosophy, sociology, and computer science. My multi-disciplinary approach to the study of trust in HRI was beneficial because it allowed me to reduce the potential challenges with collaborative efforts across disciplines (that often surface in mixed-disciplinary HRI teams) by clarifying and addressing the various subtle differences in shared terminology, and supporting decisions about which methodological assumptions to accept (or reject). Throughout the process of developing my PhD project and writing up this dissertation, I continuously informed myself about various disciplinary perspectives on how to understand and study trust in HRI and complemented my research with them. Bringing together knowledge and theories from different disciplines provides not only a chance to understand, analyze, and study trust in HRI in a way that is sensitive to the multiple issues and dimensions of this matter, but also enables a more integrative stance when considering the development of trustworthy robots. Even though research in HRI is interdisciplinary by definition, much of the work tends to rely on more cognitive or psychological perspectives when it comes to the topic of trust in HRI. I will show that perspectives and methods from philosophy and sociology enrich current discussions on both a theoretical and methodological level.

1.6 Methodology and Methods

In my PhD project, I made use of a methodology that brings together methodology and methods from different disciplines. From the start, I planned to make use of conceptual analysis taken from philosophy (used for the chapters 2 and 3) together with expert interviews that originate from sociology (used for chapter 6), and experimental study design from the field of HRI (used for the chapters 4 and 5). In this sense, my project is also multidisciplinary on the methodological level, as I went beyond disciplinary borders for the choice of my methods (see e.g., Fig 1.1).

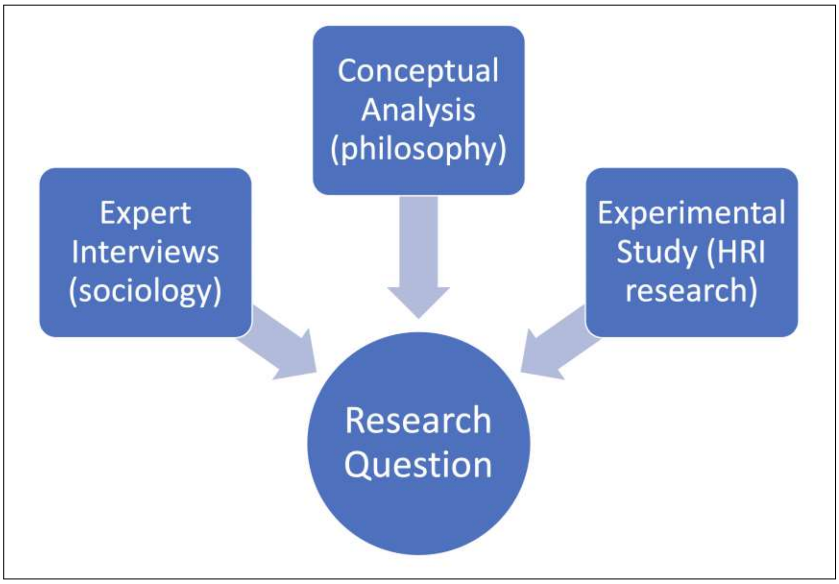

Conceptual Analysis (philosophy)
Expert Interviews (sociology)
Experimental Study (HRI research)
Research Question

Figure 1.1: The multidisciplinary methodology that I planned for my PhD project to address my overall research question.

However, I had to make significant changes to my methodology and choice of methods due to the global COVID-19 outbreak in March, 2020. As part of the HUMAINT project, together with the rest of the team, I had already planned and prepared the experimental study, and we also managed to run an in-person technical pilot that we conducted in our laboratory located in Seville. Though we were aware of the risk that the COVID-19 outbreak could turn into a global pandemic at that time, we hoped to be able to carry out our studies unless it was no longer permitted or caused any risk of harm to our participants. However, exactly in the week we were supposed to conduct our in-person experimental study, countries all around Europe went into full lockdown, and we had to stop our plans with immediate effect. We first thought, as did many, that we could simply postpone our experimental study by two weeks. Without any improvement in sight with the situation of the COVID-19 outbreak, and equipped only with home-office resources, we were forced to discuss over several meetings what alternative plan we might be able to use. In the end, it required a complete redesign as we decided to transfer the planned experimental design into an online study by using instead the format of an interactive survey (see e.g., Fig 1.2).

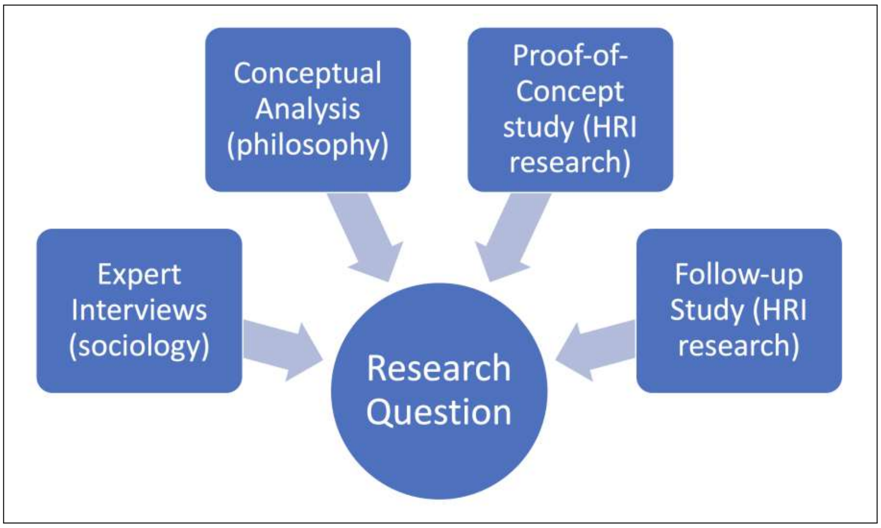

Conceptual Analysis (philosophy)
Proof-of-Concept study (HRI research)
Expert Interviews (sociology)
Follow-up Study (HRI research)
Research Question

Figure 1.2: The multidisciplinary methodology that I actually used for my PhD project to address my overall research question.

Without any experience making online surveys, I had to figure out in record time how to ensure that our online study could be designed in a way that made it similar enough to our planned experimental study design. Using the basic methodology and methods of experimental study design, I developed in collaboration with Dr. Astrid Weiss and Dr. Vicky Charisi from the HUMAINT project our *interactive online survey* that we ended up using for the online study, and the follow-up version (see more detailed explanation in the chapters 4 and 5). Without knowing that the initial lockdown would be followed by many more, in the summer of 2020, with the support of my supervisor, I started planning for an in-person experimental HRI study in collaboration with the HRI team lead by Prof. Young at the University of Manitoba, Canada. Again, after having put much effort into the preparations , we were forced to cancel all our plans because traveling to Canada was not possible within the time frame needed to include this work as part of my PhD project. Furthermore, all the expert interviews (with the exception of the very first one) that I conducted for my PhD project took place online (see chapter 6 for further details). While in-person interviews are the most ideal way of conducting data collection, I believe that the online format was a reasonable alternative in a time where there was no other option. Luckily, the effort I put into the conceptual analysis was not effected by the global COVID-19 outbreak in any significant way.

To ensure that the methods I planned and used for my PhD project are familiar to those who belong mainly to only one of the disciplines I bring together, in the following section, I will provide a brief explanation of the basic idea behind each of them.

1.6.1 Conceptual Analysis

Careful conceptual analysis is rare within the field of HRI because lexical or borrowed definitions are often considered sufficient. The lack of careful conceptual analysis is problematic because it hinders adequate understanding of the specific phenomenon in question as well as possible explorations of how best to study it given our current scientific theories and methods. Despite recent discussions in philosophy about whether philosophical questions are merely conceptual ones (Williamson, 2007), conceptual analysis is by many considered one of the most predominant and defining methods in philosophy today. This method gained influence mainly in the analytic tradition of philosophy that developed in the 20th and 21th century. Today, it is characterized as the "linguistic turn" because it was heavily promoted by those members of the Vienna Circle that adhered to logical positivism, later also known as logical empiricism (Glock, 2008). Roughly, the goal of conceptual analysis is to provide an answer to the question of the structure "what *is* X?" by stating the individually necessary and sufficient conditions for the given phenomenon denoted by X. By asking this kind of question, as explained by Knobe and Nichols (2008), the philosophical method of conceptual analysis "attempts to identify precisely the meaning of a concept used to capture the phenomenon in question by breaking the concept into its essential components, which themselves typically involve further concepts. In an attempt to determine the meaning of a philosophically important concept, one often considers whether the concept applies in various possible cases" (p. 4).

In my PhD project, I used this philosophical method of conceptual analysis to explore the notions of trust and vulnerability to guide the empirical work (as presented in 4, 5, and 6). Specifically, I made use of the variation of conceptual analysis that aims to be constructive and to provide "explicit relation among terms or concepts of a language within our conceptual theory of that language" (Kosterec, 2016, pp. 221-22). Using a constructive conceptual analysis broadens our knowledge about how various concepts relate, either by postulating new relations, or by affirming that some already known relations are present within new regions of a specific language or discourse. As a result, it is possible to introduce new terms or concepts into language usage that were previously lacking in the initial and explicit conceptual theory. Though many philosophers learn to use the method of constructive conceptual analysis indirectly by reading and writing philosophical texts on various topics, the guideline proposed by Kosterec (2016) consisting of six basic steps to take for a good analysis: (1) specify the initial conceptual background CB, (2) formulate the conceptual problem P, (3) state the new conceptual relation R, (4) formulate tests T of the conceptual relation R within CB, (5) elaborate the new relation R by tests T respecting CB, and (6) if the relation R succeeds in tests, declare it a part of CB.

As (Bennett, 2017) argues, it is important to keep in mind that serious problems can arise when using the philosophical method of conceptual analysis exclusively because it is essentially operating on the descriptive level of analysis. Thus, any normative proposals to change or improve current concepts needs to be guided by considerations external to the specific conceptual theory as it is expressed in any given language (Thomasson, 2015). In my PhD project, I have included theoretical perspectives on trust and vulnerability from both philosophy and sociology discussions to ensure that my constructive conceptual analysis would be able to support my empirical work on trust in HRI.

1.6.2 Experimental Study

For the empirical work I planned and did undertake (with some modification) to examine human experience of vulnerability as a precondition for trust in HRI, I used the methodology and methods of an experimental study. As Bartneck et al. (2020) write in their introductory textbook into HRI research, it has become a very established practice in the HRI community to conduct experimental studies. The primary unit of analysis in such studies is the interaction between humans and robots. The results of experimental HRI studies are often used to gain either knowledge about the required functionality and design of robots (robot-centered), or to understand the attitudes or behaviors of people towards robots (human-centered) for successful HRI. I used the style of a human-centered experimental study for my PhD project, which was also exploratory in nature, as research into the vulnerability of humans as a precondition for trust in HRI has not yet been carried out. As such, my experimental human-centered HRI study aimed to gain fundamental knowledge about whether the vulnerability of humans as a precondition for trust was also of relevance in the context of trust in HRI.

On a very practical level, I planned to use a between-subjects design for my experimental study, and I required at least 30 participants for each experimental condition[9]. Moreover, the target population of my experimental human-centered HRI study is that of the general public, since I connect my understanding of human vulnerability to the discussions of vulnerability as part of the human condition. Consequently, the sample of participants consisted of people that were diverse in terms of gender, nationality, age, educational level and type, as well as familiarity with robots. Ideally, the population sample for my experimental study would have to also be collected randomly in the sense that all members of the target population have an equal chance of being included in the participant sample (Hesse-Biber, 2010), though our actual sample poll tended to be more representative of the special groups of university students, researchers, and people from the more Western part of the world. It is also important to mention that I used only a single and short interaction between the human participant and the robot for my experimental HRI study design, even though this decision also lends itself to problems with the novelty effect[10] (Belpaeme, 2020). However, by acknowledging this methodological limitation, I can be transparent about how this physiological effect might influence the results and do my best to take this into account when analyzing and interpreting the data I collected .

For the data collection and analysis, I used the methodology of mixed-methods as established within the social sciences in the late 1980's to early 1990's (Creswell and Creswell, 2018; Hesse-Biber, 2010), which has also gained some popularity

[9] I decided on this specific number of participants because it is recommended in the HRI community to have a minimum of 25 participants per condition (Bartneck et al., 2020).

[10] According to Belpaeme (2020), the novelty effect in the current HRI literature is used to explain the possibility of finding either too positive or too negative effects in the results because participants are new or unfamiliar with the robot, the study setting, or interaction design.

for experimental studies in the HRI community (Bartneck et al., 2020). Mixed-methods is considered pragmatic in terms of its methodology because the underlying worldview argues that quantitative methods (rooted in a positivist worldview) and qualitative methods (from a constructivist or transformative worldview) are complementary and provide a more complete understanding of a research problem than either of them separately. More precisely, I used a convergent mixed-methods design methodology, "in which the researcher converges or merges quantitative and qualitative data in order to provide a comprehensive analysis of the research problem" (Creswell and Creswell, 2018, p. 52). By using questionnaires, open-ended questions and semi-structured interviews, I collected and analyzed both forms of data more or less in parallel, so that the interpretation of the overall result has integrated important findings from various datasets. With this choice of design, the rigor and credibility of my experimental human-centered HRI study was ensured by the method of triangulation, as the validity of the results and findings was always cross-checked against the different datasets (Hesse-Biber, 2010).

Given that I was not able to in fact carry out my planned in-person experimental HRI study, I will point out the important changes that we have made to the methodology and methods so far. First of all, I used the format of an interactive online survey instead of having people come into the laboratory to interact with the robot. I addressed this challenge with my new study design by having a module in the interactive online survey in which the participants had to click their way through an interaction scenario with the PEPPER robot (see a more detailed description of the structure and components in the chapters 4 and 5). Secondly, we decided for our follow-up online study to complement our interactive online survey with a collection of semi-structured interviews with some of the participants (see chapter 5 for a more detailed description). Thus, while the studies conducted for my PhD project are not officially regular in-person experimental HRI studies, many of the overall methodological assumptions and decision for the choice of methods are very similar to what I have described in this section. I also provide further explanations in the discussion section of the chapters 4 and 5 on how our development and use of the interactive online survey proves a very promising alternative for data collection along the lines of an experimental study when in-person meetings are either very limited or impossible.

1.6.3 Expert Interviews

The empirical work that focused on exploring the vulnerabilities of robots was informed and supported by the methodology and method of interviewing experts with state-of-the-art knowledge about the technical challenges of developing robots. An expert interview is a method used for qualitative research in sociology used to gather and utilize expert knowledge (Meuser and Nagel, 2009). Though expert interviews are widely and commonly used within sociology, it is only recently that a strong methodological foundation of this method has been discussed and established, as part of discussions within sociology of knowledge (Bogner et al., 2009). While expert interviews are mainly considered a very effective method for gathering more specific information about a particular subject matter or domain, careful reflections are required when using this method because "issues of what constitutes an expert, the differences between the various forms of expert interviews and their role in research design" (Bogner et al., 2009, p. 1) quickly arise.

Differently from early approaches within sociology of knowledge to the method of an expert interview, I take into account for my own PhD project that the expert knowledge that can be collected cannot be reduced or purely defined by the practice of specific professions with which it might be associated (i.e., expert knowledge on robots may not be limited to the knowledge an engineer in robotics has). Therefore, I interpret experts as people who have privileged access to information and play an active role in structuring and defining the relevant issues or problems. Thus, what makes someone an expert is that their practice and experience clearly address a demarcated range of problems within a specific domain, which others find meaningful and can guide their actions by (Meuser and Nagel, 2009). In more precise terms, I conducted interviews according to the methodology of "systematic expert interview" (Bogner and Menz, 2009; Flick, 2009) because the purpose was to obtain systematic and complete information on "objective" matters about trust in HRI. That is, providing an individual portrait of the expert and their knowledge is not the primary object of interest. The focus is rather on the function of experts as informants who provide information about the problems or questions being investigated, by presenting "facts" and elaborations drawn from their specialized knowledge. For this reason, data collected by means of systematic expert interviews always has to be comparable to the specific subject matter or domain, which in the case of my PhD project was the topic of trust.

The pool of experts selected for my PhD project was based on purposive sampling, where a potential expert for interviewing is chosen according to their domain expertise as it relates to the specific research question. As such, “expert” is to be conceptualized in my PhD project as a relational term, insofar as the selection of persons to be interviewed depends on the question at hand and the domain being investigated. Additionally, I asked the experts interviewed whether they could suggest other relevant experts for consideration, thus extending and supplementing my purposive sample strategy with that of snowball sampling (Biernacki and Waldorf, 1981; Handcock and Gile, 2011; Miller and Brewer, 2003). Snowball sampling is particularly beneficial for interviewing experts, as this strategy can help overcome the difficulty of gaining access to wider circles of experts more effectively, and has the potential of ensuring a saturation principle is reached. The data collection for my PhD project rested on the use of semi-structured interviews. Ensuring that my interviews with experts were semi-structured was important to ensure a systematic exploration, because the topic of robot vulnerabilities is still rather uncharted and scarce in the current HRI literature, therefore requiring a more exploratory approach as a start.

I analyzed the data collected from the expert interviews using the six-step procedure provided by Meuser and Nagel (2009): (1) transcription (audio record the interview and thereafter transcribe only the thematically relevant passages. I focused on the thematic units rather than prosodic and paralinguistic elements), (2) Paraphrasing (paraphrase the conversation as it unfolds and account for the opinion of the interviewee according to the intended meaning), (3) Coding (order the phrased passages thematically while keeping as close to the text as possible and adopting the terminology of the interviewee. The amount of codes for sorting the passages will depend on how many topics will be addressed by the interviewee), (4) Comparison (go beyond the individual interviews by comparing the thematic passages with each other. To ensure soundness, completeness, and validity, each category for the different passages might be revised during this process), (5) Conceptualization (condense and formulate the commonly shared knowledge among the interviewees as it has been categorized through the process of coding and comparison. The result will be statements referring to structures of expert knowledge as the common and different features of the interviews are elaborated and conceptualized according to the theoretical knowledge base), and (6) Generalization (contextualize and frame the empirical general findings through theoretical perspectives where the meaning structures emerging from the interviews will be connected and will form topologies and theories. This reconstructive process enables previous loose ends and unconnected findings to be brought into new light).

1.6.4 Research Ethics

I planned and did my best to ensure the protection and integrity of the participants taking part in my empirical work for my PhD project by following 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.

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 my empirical work at the time when I was conducting my two online studies and expert interviews. 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, I did my best to compensate for the lack of ethical approval of my empirical work. I was in contact with Dr. Marjo Rauhala about my expert interview. Because Marjo supports all researchers at TU Wien on a daily basis with the identification of questions regarding research ethics in the role as the leader of the service unit of Responsible Research Practices[11], I received some feedback on both my project description and consent form so they would live up to basic standards for good research practice. For guidance about how to follow the EU regulations of GDPR, I was in contact with Assoc. Prof. Peter Purgathofer, who has 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 as

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

preparation for their interviews (see e.g., Figure C.3 in appendix C). Additionally, I was able to get ethical approval for the online study and its follow-up version from the ethics board at the EU commission, as this empirical work was conducted as part of the the HUMAINT project. At the very beginning, we discussed with and asked the PI of the HUMAINT project whether a separate ethical approval was necessary, but our slight changes to the consent forms were considered acceptable, as the original approval also covered research with children that normally requires even stricter procedure.

Overall, we were able to impose all four principles of ethical research when conducting my online study and the follow-up version. This was partly ensured through the information provided on the consent forms we developed for our interactive online survey (see e.g., Figure A.1 in appendix A). We also provided more detailed information about the project aim and what they as participants were asked to do for their participation in our online study, though only after consent was provided. We made sure to provide this basic information earlier for our follow-up and improved online study (see e.g., Figure B.1 in appendix A). For conducting my expert interviews, I provided all expert with a PDF file with the project description that they were going to be interviewed for (see e.g., Figure C.2 in appendix C), which was attached to the email invitation (see e.g., Figure C.1 in appendix C). Given the nature of the expert interview, I left out principle 4 for the informed consent of the experts, as it stated in the consent that I could use the name, professional title and affiliation for the purpose of quotations (see e.g., Figure C.3 in appendix C).

I would also like to mention that I obtained an official certificate on research ethics as part of my preparation for the planned in-person experimental study with the HRI team at the University of Manitoba, Canada. It was a requirement from the side of the university that all members of the project team must prove that they completed the course "Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans" (TCPS 2)[12] as instructed by the Interagency Advisory Panel on Research Ethics in Canada[13]. Because the TCPS 2 course consisted of multiple tests on how to navigate research ethics involving human subjects, I

[12] The content and structure of the TCPS 2 course can be found on the following website: `https://tcps2core.ca/welcome`. Back in June 2020 I took the course "TCPS 2: CORE-2018", which has now been updated to the new version of "TCPS 2: CORE-2022".

[13] Visit the website of the Interagency Advisory Panel on Research Ethics for more information: `https://ethics.gc.ca/eng/home.html`

received some training that I now also consider a very important part of my path towards becoming a responsible HRI research even though this was not required for obtaining my PhD project at TU Wien.

1.7 Main Contributions

With my PhD project I contribute mainly to the research field of HRI by closing a gap in discussions and studies on trust in HRI that currently overlook the importance of vulnerability, though I also provide some input to current discussion on trust for research in philosophy and sociology. My contributions to these discussions relate to improvements on both a theoretical and methodological level, as well as considerations about how this knowledge gained can be of benefit to the engineering practice specifically. While I present in chapter 8 an account of how my work contributes to current discussions about trust in HRI by answering my overall research question and sub-questions, I will highlight already in the following sections some of the main contributions.

1.7.1 Theoretical Refinement

From the work on my PhD project mainly presented in the chapters 2 and 3, I achieve several theoretical contributions through my careful conceptual analysis:

- I account for and emphasize *vulnerability as a necessary and active precondition of trust*, where risk and uncertainty are the two others. This work requires a deep understanding of how the various concepts that make up the rich notion of (interpersonal) trust relate to each other.
- From this analysis, I provide a theoretical account of how to conceptualize the relation between trust and vulnerability beyond characterizations that take it to be exclusively a matter of properties. As an alternative, I present my *event approach for trust in HRI*, which focuses on how to identify and evaluate whether trust between humans and robots emerges from the interaction where both parties play an important role in establishing trust (be it in a positive or negative reinforcement).
- Zooming in on the very specific trust-vulnerability relation, I also provide a *definition of trust that includes the preconditions of vulnerability* as it has not

been considered in research on trust in HRI so far. Moreover, this working definition can support hypothesis generation when aiming to empirically study trust in HRI.

1.7.2 Empirical Anchors

My empirical work for this PhD project is going to be presented in the chapters 4, 5, and 6. Though each study contains several interesting contributions, I list for now only the main ones for each of them:

- In the online study where we investigate human experience of vulnerability for trust in HRI through different interaction scenarios , we were able to show that there is a *relationship between the very ordinary, mundane, and familiar situation of clothes shopping with the help of a robot and the experience by people of feeling vulnerable* given their expression of discomfort during the interaction. From a discussion of our results, we argue that such findings provides interesting input to current methodological discussion about how to study trust in HRI in situations that are closer to the everyday lives of people.
- By further investigating human experience of vulnerability for trust in HRI within the privacy scenario specifically as presented in our follow-up online study, we also found that although some people did report that they felt vulnerable in the time of the interaction they also did consider interacting or engaging with the same robot again in a real-life situation despite having their trust violated through a mild privacy breach. Their motivation for trusting the imperfect robot was based mainly on their *expectation of gaining utility or entertainment value*, which we concluded suggests that the trust people place in the robot within this specific privacy scenario is based on a lower opinion of the interaction benefit that is commonly higher for interpersonal trust between humans when considering the philosophical literature.
- Aiming to explore the vulnerabilities of robots through expert interviews with leading roboticists, I identified and discussed that successful (or positive) trust in HRI is not only about how well robots function, navigate, and behave in a dynamic world. The development of strategies for *how robots cope*

with malicious people turns out to be important for trust in HRI, as mutual recognition of the others vulnerability is essential for trust to play a role in the bettering of interaction, collaboration or engagement between humans and robot.

1.7.3 Potential Alliance

In chapter 7, I discuss how the accumulation of all the things I learned from my PhD project gave rise to a multitude of meta-reflections. Following the same structure, I will provide a short description here of how the different points for discussion add to the list of my contributions:

- To find a good way to infuse my theoretical perspectives on trust in HRI into the HRI community, I initiated and organized together with some colleagues the HRI'22 "Theory-Grounded Human-Robot Interaction" (THEORIA) workshop. With our attempt to bring together HRI researchers interested in theory and theorizing (with 76 people preregistering to participate in total), we were able to *identify and meet the increasing demand* within the community to discuss how to establish and promote more theory-driven HRI research. With my own theory-driven PhD project, I believe I have also provided some basic insight into how theoretical perspectives from philosophy can be beneficial in making the complex and very abstract concept of trust accessible to HRI research, and shown how this approach can help tease out conceptual relations worth exploring through empirical work.
- Together with a colleague, I set out to explore how to best infuse the knowledge gained about (interpersonal) trust in HRI into the community of robot engineers by facilitating a pilot workshop on the specific theme of "designing for trust in HRI". By combining our card-based design tool with a scenario-based design methodology, we were able to provide a foundation for the *knowledge transfer* of the various dimensions of the concept of trust for engineers developing robots mainly through constraint-based problem solving. From further reflection on how successful we were with our pilot workshop, I also anticipate the potential *knowledge transfer* of the concept of trust for the specific engineering practice on the stages of testing and evaluating whether robots function or behave in a way that would enable people to trust in them based on their perceived trustworthiness.

- Last, but not least, I have throughout the whole process of my PhD project come to know myself better in terms of how to position myself in the typical identification of philosophers as either logicians or ethicists: I have come to identify as an *experimental philosopher*. This also allows me to position myself clearly in interdisciplinary collaborations. What I hope others from the HRI community can learn from my own development as an (experimental) philosopher doing HRI research is that the old wisdom of *knowing yourself* (as famously inscribed over the temple in Delphi) can foster collaborations across disciplines borders because the realization of knowing what you do not know ensures that members of the team are ready to learn from each other.

1.8 Dissertation Overview

My dissertation presents a written documentation of my PhD project in a selective and organized manner. It consists of eight chapters in total that collectively serve to bring the various processes and components of my PhD project into a joint narrative addressing the research question and sub-questions. Overall, there is first an introductory chapter, then two chapters that present the theoretical perspectives, followed by three chapters laying out the empirical work, a longer discussion chapter about the knowledge transfer and integration of the main findings, and finally, a last chapter containing the conclusion. (see e.g., Fig. 1.3).

With **chapter 1**, I have started my dissertation with a very broad introduction. In this first part, I presented how current research on trust in HRI sits in the slipstream of a longer and ongoing discussion about how people in late modernity interact with and relate to technology. I also presented the current state-of-the-art in HRI regarding the topic of trust specifically. Against this backdrop, I used the second part of the introduction to presented the overall motivation and aim of my PhD project, and how I planned to address my research question using a theory-driven and multi-disciplinary approach.

In **chapter 2**, I will provide a basic account of the concept of trust and how it is possible to apply it in the context of HRI. This work leads to my proposal of an event approach to trust in HRI, which highlights how vulnerability stands in distinctive conceptual relation to trust as one of its preconditions. Based on this insight, I offer a working definition of trust for use in the context of HRI,

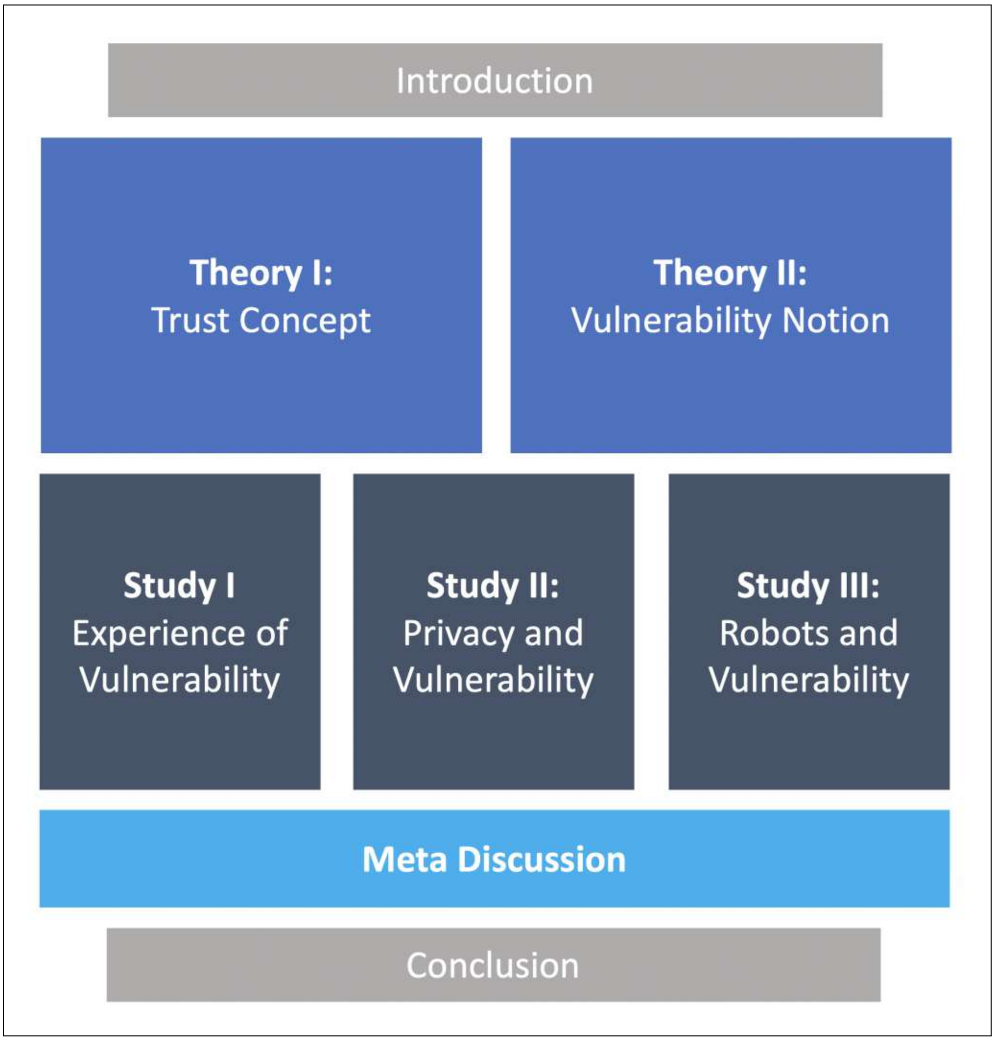


Figure 1.3: A visual overview of the dissertation structure and the connection between the chapters.

which takes into consideration the conceptual relationship between trust and vulnerability.

Zooming in on the notion of vulnerability, I will in **chapter 3** explain why addressing vulnerability is often avoided in current research on trust in HRI. From this, I account for how I understand and apply the notion of vulnerability in the context of HRI, and show how this allows me to focus on different themes in the analysis and study of trust in HRI. I also discuss how the notion of vulnerability cannot be interpreted and treated in the same manner when considering the perspective of humans and robots.

In **chapter 4**, I present the results of a theory-driven HRI study that aimed to explore human experience of vulnerability in the context of clothes shopping with a robot assistant through the themes of economy, privacy, and transparency. Intended as a proof-of-concept study, we designed and conducted an interactive online survey that demonstrates that it is possible to measure that there is a relationship between human experience of vulnerability and the ordinary, mundane, and familiar situation of clothes shopping when people have to trust in robots or find them trustworthy.

I then present a follow-up study in **chapter 5**, which dives into the theme of privacy for a more detailed view on how people experience their vulnerability when they have to trust in a robot for clothes shopping. Among several interesting findings and points for discussion, we conclude that the main motivation for people to interact with a robot again after the trust-violation instance of a mild privacy breach that also left them with an experience of vulnerability was driven by the simple added values of utility and entertainment.

The results of several experts interviews that I conducted to explore in which way robots can be considered vulnerable will be presented in **chapter 6**. From this work, I was able to identify in total 13 categories of vulnerability that can be grouped into the different themes of embodiment, processing, people, and environment. In the discussion, I specifically focus on how experts interpreted the notion of vulnerability, and how malicious human behavior can be problematic when considering mutual trust in HRI.

I take a step back in **chapter 7** to reflect and bring into discussion how my work contributes to the field of HRI as part of more contemporary computer science. My aim is to clarify how I transfer and integrate the knowledge I accumulated in a way that it is useful to the HRI community. I also consider the different roles philosophers in the HRI community can take, and position myself as an experimental philosopher given the work I have undertaken in my PhD project and present in this dissertation.

In **chapter 8**, I conclude this dissertation by providing short answers to my posed sub-questions and overall research question, pointing towards future work given the knowledge and insights I have gained, and adding a last remark to express how my work in carrying out this PhD project is part of a longer journey.

CHAPTER 2

Foundation of Trust

> [...] "neglect leads to fragmentation of meaning, which seems to justify further neglect and further fragmentation until eventually a concept can disappear entirely."
>
> Zagzebski (2009, pp. 141-142)

Part of section 2.3 and 2.4 in this chapter will appear as a book chapter organized and edited by the co-organizers of the Trust Robots Doctoral College (TU Wien) as: *Hannibal, G. & Weiss, A. (forthcoming, 2022). Exploring the Situated Vulnerabilities of Robots For Interpersonal Trust in Human-Robot Interaction (pp. 1-19). Vienna, Austria: TU Wien Academic Press*.

I will in this chapter take a specific philosophical perspective on the complex topic of trust and offer some basic understanding and conceptual clarification. This work is important not only because trust has been so widely studied across various disciplines that each highlights different aspects and methods for its exploration, but also because trust is a very abstract concept that easily slips out of hand. While this broad perspective on the topic of trust might seem very far away from the more narrow theme of this dissertation, I believe it is necessary to start out with this groundwork to limit any unnecessary confusion later on when moving on to how trust can be understood in the specific context of human-robot interaction.

Reading my way through some of the most influential work on the topic of trust, and considering also the continuously growing body of literature, this chapter is motivated by my observation that much discussion on the topic of trust is to a large extent easily blurred because of an inability to clearly separate from the start the *metaphysical* questions about trust from its *epistemological* questions. For this reason, I have chosen to use this distinction also for the structure of this chapter, to show why these different questions are treated best separately. The important difference I wish to highlight is that the metaphysical questions deals with *the nature of trust* while the epistemological questions are more concerned with identifying *the reasons* for trusting to be placed well.

2.1 The Metaphysics

Very simply put, metaphysics[1] is a branch of philosophy that aims to account systematically for *existence* by answering questions in the most general way about *what there is* and *how it is* (Loux and Crisp, 2017). In my dissertation, I will allude to metaphysical questions about trust by asking what trust is *in itself*. To provide any satisfactory answer, the aim must be to get as close to the true nature of the trust phenomenon as possible. Even though there are many metaphysical questions that should be addressed for a proper and extensive account of trust, I will only be dealing with its relation type and objects , which I deem most relevant to any further investigation of studying trust in HRI.

[1] To those not so familiar with philosophy and its core disciplines, metaphysics might sound like something above or beyond physics in a way that evoke thoughts about the mysterious or even occult. The origin of metaphysics, however, is actually rather accidental. It was coined by an (unimaginative) editor in ancient Greece who decided to name the books of Aristotle according to their order. Since “meta” actually meant “after” in ancient Greek, all the books referred to as the metaphysics were simply those that came after the books on physics. As a more serious reflection, there is today ongoing discussions in contemporary philosophy about what exactly is meant with “metaphysics” as a special philosophical discipline with its own subject matter and methodology.

2.1.1 Trust Relation

There is a fundamental distinction to be made in our understanding of trust concerning its scope, that revolves around the *kind of relation* it can be said to constitute in the sense of one-place, two-place or three-place. As the one-place relation of trust is recognized as central to discussions about what can be said about the trust-giver (i.e., trust as virtue, emotion, character or personality trait), I will in this section lay out the difference between the two-place and three-place relation of trust, with a view towards more recent discussions in the philosophy of trust.

According to Domenicucci and Holton (2016), trust in its most basic form can be considered a two-place relation with the predicate instance of 'A trusts B'. Drawing a parallel to the notions of love and friendship, they explain that conceiving of trust as two-place means that an agent A trust in another agent B *simpliciter* because there is no particular feature, behavior or accomplishment on which this trust is contingent. Trust as a two-place relation between agents *is* that which "we might have in a parent, or a partner, or a child who is old enough" (Domenicucci and Holton, 2016, p. 151). For this reason, they consider the two-place relation of trust as more fundamental than trust as three-place, which they support with a logical argument. Against the argument that two-place relation is just a more generalized version of the three-place relation of trust given the range of the quantification (i.e., "I trust you to F, for some class of F" (p. 152)), Domenicucci and Holton (2016) assert that trust in this view does not support the common intuition that often stands as the explanation or justification for F in the ordinary use of language (e.g., "Alright, I trust you. You can borrow my car"). To stress this interpretation, they also point out that trust in the two-place relation undergoes a significant change if considered in the present continuous construction (i.e., "I am trusting you") because it suddenly require further inquiry (i.e., "trusting me to do what?" (p.153))[2]. Understanding trust in terms of a two-place relation as presented and defended by Domenicucci and Holton (2016) means that there is not a third predicate instance included, which means that

[2] How their account of the two-place relation of trust is able to deal with trust as partial or being a matter of degree, is something Domenicucci and Holton (2016) also address because it is considered a strong objection. Although such a discussion is very interesting for a more in-depth philosophical account of the trust relation, it is not necessary to further my own argument .

they take trust to be “primarily an attitude to a person” (p. 155). Overall, trust from this perspective will always have to start from consideration of what is going on *internally* with the agent A (e.g., private thoughts, opinions, motives, beliefs, attitudes, emotions). Thus, trust as two-place relation is not always visible or observable from a third-person perspective.

Quite differently, to consider trust as a three-place relation is the claim that it involves an agent A, an agent B, and that which an agent B is being entrusted with (Baier, 1986). It takes the predicate instance of ’A trust B to do C’. To get a feeling of the logical structure more clearly in the way it is also reflected in more ordinary language use, consider the following examples:

> (a) Lisa trusts her boyfriend to look after her purse. (b) Anne trusts Clare to bring their cat to the vet. (c) Peter trusts the plumber to repair the sink. (d) The instructor trusts the student to be on time.

It is evident in all four examples that the trust mentioned is very closely connect to the *action* of agent A rather than how this agent is regarded – it is determined by the overt behavior of agent A to entrust agent B to do (or to not do) something. This added predicate instance of C most often brings into the picture a specific task conducive to the achievement of a certain goal (i.e., Lisa wanting to avoid loosing her possessions, Anna wanting to ensure the health of the cat, Peter wanting to get the sink working again, the instructor to start the lesson). As such, the three-place relation of trust is viewed and understood as a way of *interacting* with the world or other agents, which makes it determined by something that is *external* to agent A. This might be one reason for why the three-place relation of trust is to this day still the most examined and endorsed form of trust relation in current literature debating the nature of trust. The appeal, as I see it, stems from the eagerness among most philosophers (and indeed also among most social scientists) to show that trust is always tied to the social reality of humans, to which it is often said to be crucial. So while trust as a three-place relation cannot be considered as fundamental as that of a two-place relation considering the argument provided by Domenicucci and Holton (2016), it is nevertheless taken by many, in the discussion about trust, as the starting point for its understanding and analysis[3].

[3] Especially some philosophers from the modern analytic tradition tend to advocate for the three-place relation of trust as primary to such an extent that someone new to the discussion

However, because trust in the form of a three-place relation is conceived of as something that an agent *does* first of all, it is of worry that its more simplistic version is too similar to another very well known, but still rather different, three-place relation – that of reliance (i.e., agent A relies on agent B to prepare the dinner)[4]. While it is recognized that trust cannot be reduced or collapsed to *mere* reliance, one of the biggest reasons for disagreement within the philosophy of trust is in identifying what exactly is present (or absent) for this distinction to be upheld[5]. Thus, proponents of the three-place relation of trust will usually add some extra factor related to norms or morality to their account because consideration only of performance highlighted by this simplistic view would otherwise make it indistinguishable from the three-place relation of reliance (Thompson, 2017). This desire to allow reliance to be embedded into the three-place relation of trust only in an accompanying role, not as something that exhaust trust, results from the observation that people tend to muddle together trust and reliance in their ordinary way of speaking. Since philosophers do to a large extend concern themselves with how language reflects our intuition of trust, they do want to understand and analyze those instances in which the phenomenon is both logically and grammatically *about* something (as shown in the examples of the three-place relations of trust already provided).

While discussing whether the two-place or the three-place relation is to be preferred over the other for a proper understanding and analysis of trust, I will in this dissertation only be dealing with the latter. First of all, it is a more appropriate aim to demand trust in the form of a three-place relation for the context of HRI, because it brings into scope the behavior of a robot. The two-place relation of trust allows it to remain only a mental disposition detached from its interactive component, which is counterproductive since HRI is about interactions between humans and robots. Furthermore, because the main argument for the development of robots is related to their utility (i.e., the robots should support or assist workers in factories, the elderly at home, children with learning, or provide a good

might stay uncritical about exploring the other kinds of relations trust can take.

[4] Other examples of a three-place relation that is often confused with trust is testimony (i.e., agent A believes/accepts agent B's word about C) and cooperation (i.e., agent A and agent B cooperate to achieve some specific end C).

[5] I will later on in this chapter briefly engage with this discussion when providing my arguments for why the understanding and analysis of trust specifically in the context of HRI requires a rethinking on a metaphysical level, though for now, it is sufficient to simply know that this is a differentiation that is being made.

service), it is reasonable to look towards the three-place relation of trust, as it more easily accommodates such added value in the form of reliance (or cooperation for that matter). When trust is mentioned in this dissertation, henceforth, it is always in the form of a three-place relation (i.e., a human trusting a robot to do (or to not do) something).

2.1.2 Trust Objects

By looking at the predicate instance of trust as a three-place relation, it is also possible to identify its constitutive objects, which are as follow:

(1) trust object A: agent A
(2) trust object B: agent B
(3) trust object C: task

Being more sensitive to the different *kinds of trust objects* is not only useful in making explicit how to link other relevant concepts and distinctions into the analysis of trust, but also in furthering the discussion about *who* or *what* can be an appropriate target of trust. Because I will return to the latter point for discussion in this chapter, I will in this section only be focusing on presenting the trust objects and briefly account for the particular role they play. I will start from the bottom of the list because the standing of the trust object C is not going to cause much dispute.

As explained, for ’A trusts B to do C’, agent A must have a certain goal (e.g., in the form of concern, need, desire, fear, intention) in mind from which specific actions are required for its achievement or maintenance. As Castelfranchi and Falcone (2010) explain, the goal of agent A brings to view the motivational aspect of trust because “if I don’t potentially have goals, I cannot really decide, nor care about something (‘welfare’): *I cannot subjectively trust somebody*” (p. 13). It would simply be hard to see how related notions such as “expectation”, “outcome”, “safety”, “confidence”, “stake”, “dependence”, or “vulnerability” come into play for an understanding of trust if agent A did not have any goal. Moreover, the goal of agent A is also what makes the discussion about the *rationality of trust* pertinent, because it anchors the evaluation of agent B by agent A relative to a specific task to ensure that the set goal will be reached or preserved. Similarly, trust specifically used to enhance collaboration is characterized by the presence of a *common* goal held by agent A and B, not merely the personal goal of agent A. Given this goal, the actions to be taken when agent A entrusts agent B with something C can vary greatly in number and complexity, and also depend on how they support a specific task. As a trust object, the task (or set of tasks) is something that agent A has to consider essential to fulfilling or sustaining the goal because it would otherwise not serve the function of providing the three-place relation of trust with its *aboutness*. Whether the task that agent A entrusts agent B with doing is nothing more than executive actions (i.e., already specified sequences of actions) or actions that are related to problem-solving (i.e., actions that require understanding) is a matter closely tied to the expectations and evaluation of agent B by agent A (Castelfranchi and Falcone, 2010).

Since the next two trust objects (A and B) are of the same kind, I will present them together even though they play very different roles in the understanding and analysis of trust[6]. Consequently, I will not be changing my terminology to align with common trust discourse but rather continue to refer to agents A and B as the trust objects A and B. In the specific case of trust in HRI, I will simply fill in the linguistic place-holder with the definite object of concern (i.e., human and robot)[7]. No matter the choice of wording for the trust object A and B, the significant distinction between the two is that the role of the trust object A is to carry the weight of trust as discussed so far, while the trust object B raises questions about *trustworthiness*. It often happens that trust and trustworthiness are confused with one another, which sometimes causes problems for the measurement of trust if not detected in due time. Moreover, the interconnection between trust and trustworthiness also contains a normative dimension that Scheman (2020) puts into words very fittingly when she writes that "it is generally imprudent

[6] In specialized language (among e.g., philosophers, sociologist, economists, psychologists, lawyers) this difference is often marked by the choice to use more specific terminology to speak about the trust object A and B. Most common in the literature on trust is the reference to trust object A as the "truster" and trust object B as the "trustee". In the specific case of studying trust in HRI, the intention to speak about a truster and a trustee is to permit for a broad interpretation so that non-human entities can also be included. In my own view, it is not a good strategy to use this terminology because, while using these linguistic denotations as practical place-holders, speaking of a truster and trustee poses the risk of losing sight of how that makes them qualitatively different from the trust object C: they are to be classified by their *agency* (or at least apparent agency).

[7] Note that this does not imply that agent A is necessarily to be the human whereas agent B is the robot. It can be otherwise, which is part of the discussion in chapter 6 on the consideration of mutual trust in HRI.

to trust those who are not trustworthy, and it typically does an injustice to the trustworthy to fail to trust them" (p. 28). In this sense, trust and trustworthiness are self-standing concepts because one can be present without the other and they qualify equally as properties of either trust object A or B depending on the perspective taken for the analysis and study of trust. Discussions about well-calibrated trust in HRI as already mentioned in chapter 1 aim at finding the best strategies to ensure that agent A is continuously updating "the perception of an actor's trustworthiness with its actual trustworthiness so that the prediction error is minimized" (De Visser et al., 2020, p. 3). The role of predictability (or unpredictability) in placing trust well is an interesting discussion that I will also briefly touch upon in the following section. At this point, I only stress that the trust objects A and B have different functions in the understanding and analysis of trust, and that these functions also help us see how other related notions come into view such as e.g., "betrayal", "competence", "care", "commitment", "control", "benevolence", and "mistrust". How the trust objects A, B, and C in the three-place relation of trust provide support for each other is something that changes the discussion from the metaphysical to the epistemological questions of trust.

2.2 The Epistemology

Epistemology, as another core philosophical discipline, seeks to thoroughly lay bare all aspects of *knowledge* by asking questions about *what can be known* and *how it can be known* (Zagzebski, 2009). As specifically related to the topic of trust, I will consider epistemological questions in this dissertation by focusing on discussions that aim to understand *why* agent A trusts in agent B to do C. As such, the epistemological questions of trust that I want to focus on concern the interpretation of the strength of the *explanation for trust* and the exploration of the various *reasons for trust* as suggested in previous literature on trust.

2.2.1 Trust Explanation

As mentioned, trust as a three-place relation requires not only that agent A trust agent B, but also that there be a specific task that agent B is entrusted to do (i.e., A trust B to do C) and that the trust of agent A can be considered relative to the trustworthiness of agent B. To understand *why* agent A trusts agent B with a specific task, an explanation for trust must be added to the picture, which can be expressed with the predicate instance of 'A trusts for X reason'. How to interpret the link between trust and its explanation (leaving the various suggestions for the exact reason to be presented in the next two sections) is something to be mindful about, because it guides an understanding and analysis of trust that will also be helpful when considering the specific context of HRI. Overall, it is possible to have either a weak or strong interpretation of this link.

In the weak interpretation, the explanation only suggests that it is *co-occurring* with trust – the link is nothing more than a conjunctive. To see how this is the case more clearly, consider its logical structure:

(A trusts B to do C) *and* (A trusts for X reason)

In this sense, the link between trust and its explanation involves the combination of trust and its reason (no matter what it might be exactly) as two associated but still independent components. I believe that a practical example of how this interpretation works is helpful at this point. For this purpose, I will be drawing on the work by Faulkner (2007), who in his discussion about the epistemology of testimony also examined an account of trust as predictive more broadly. Here, trust is defined as "A trusts S to ϕ (in the predictive sense) if and only if (1) A knowingly depends on S ϕ-ing and (2) A expects S to ϕ (where A expects this in the sense that A predicts that S will ϕ)" (Faulkner, 2007, p. 880). In this example of trust, there is a weak interpretation of the link for its explanation, indicated by the conjunctive that enables the combination of trust as reliance with its two conditions of known and expected dependency. Continuing with the example of trust as reliance, the mere co-occurrence between trust and its explanation is a problem for an understanding and analysis of trust given that such a weak link does not offer much actual explanatory power. What is known is only that trust and its explanation are logically the same (i.e., they always coexist), but it could happen that there is an underlying factor that precedes to both of them (e.g., trust as reliance and the two dependency conditions result from the contextual circumstance of implementing a technological solution to a demographic problem[8]) or that they are brought about independently by two different underlying

[8] Think here about how elderly might trust in robots to assist them in getting out of bed, not owing to the fact that they rely on them by having in mind both the known and expected dependency (assuming also that the elderly do not suffer from dementia that impact their cog-

factors (e.g., trust as reliance originating from the wish of a favorable outcome while the two dependency conditions are rooted in the decision to cooperate, so that the coupling of the wish and decision lead to trust and its explanation to co-occur[9]). Thus, what is needed is a stronger interpretation of the link between trust and its explanation.

This possible stronger interpretation would be to say that the explanation provides a *cause* for trust – the link suggest that it is a matter of causation. To make the difference more intelligible, it is again useful to see the logical structure that is as follow:

(A trusts B to do C) *because* (A trusts for X reason)

Since trust and its reason are actually causally linked now, with this stronger interpretation, there is greater alignment with the common intuition that any explanation must support a better understanding and analysis of trust by effectively eliminating cases where trust turns out to be nothing more than accidental, or cases where it is mistaken for another underlying constrain. For simplicity, in offering an example here, I will again adhere to the example of trust in the broad predictive sense as used before (Faulkner, 2007). In some cases, then, people might trust others by relying on them not because the two conditions of known and expected dependency are the reasons, but instead because their reliance stems from necessity, persuasion, or obligation[10]. Being able to explain the link between trust and its reason in causal terms has the additional advantage that

nitive functions), but given that the eldercare facility or their family members wish to use such technological advancement as way to deal with the care-taking responsibility. It is exactly such possible scenarios that have caused much ethical debate around the development and use of robots for eldercare.

[9] Think here about how manufacturing workers might trust in robots to assist them at the assembling line as they wish to keep up a high level of productivity for keeping their job, which imply that these workers also decide to cooperate with the full awareness of both the known and expected dependency. Trust and its explanation in this case can co-occur because this specific scenario allows for a connection between wishing and decision-making that not all understandings of trust allow for.

[10] Whether trust is the kind of phenomenon that can be forced by oneself or other people is a large and lively discussion in the philosophy of trust. I will not be able to include these reflections in the work presented in this dissertation, although it is very interesting and could be relevant for consideration when using robots among more vulnerable groups within a population (e.g., children, adolescents, elderly). For my aim to study trust in HRI that concerns only people who are free to choose if they trust (or not) in robots, going into such discussion would be an unnecessary detour.

such interpretation can also shed some light on the particular way people tend to act when they trust in others to do something. When trust and its reason are causally linked, a prediction of whether agent A will actually trust in agent B to do something can be made, and this prediction can be examined partly through the observable display of behavior between the two agents when they engage in an interaction requiring trust. That is, the explanation for trust in this strong interpretation permits drawing an inference between what is taking place internally in agent A and the external patterns of behavior of agent A *as a response to the way agent B is perceived or behaves*. The impression that agent A has or gains of agent B not only plays a part in forming the reason for trust (as related to trustworthiness), but also extends to influence the actions taken by agent A towards agent B during interactions that require trust. I will later on in my dissertation (i.e., chapters 4 and 5) introduce and focus on the important dimension of *benevolence* for the identification of trustworthiness that is ascribed to agent B (i.e., the robot) by agent A (i.e., the human). In my view, the dimension of benevolence is important to take into account since it provides a specific reason for agent A to trust that is strongly reactive to what agent B *will do*: the intention and actions of agent B to avoid harming agent A is related directly to assuring that the vulnerability of agent A will not be exploited (even if the opportunity is available). Without considering the trustworthiness of agent B in terms of benevolence (at least to some degree), any demonstration of e.g., care, sincerity, fairness, and empathy towards agent A would be deemed irrelevant for understanding or analyzing the reasons for trust resulting from interactions between people. Yet, and as I will show, the intention of agent B towards agent A really matters for whether people trust in others to do something.

2.2.2 Trust Reasons

After this short deliberation on how the explanation for trust can vary in strength, and why it is best to adopt a strong interpretation for our understanding and analysis of trust in HRI given the required focus on the interaction between two agents, I will briefly present some of the various suggestions that have been proposed in the current philosophical literature about what counts as the X reason for agent A to trust agent B – what makes trust *justified*. While there are many different proposals in this discussion, there is a very strong divide between those who think that the justification for trust rests on agent A (wholly or partly) believing that agent B is trustworthy (i.e., doxastic accounts of trust), and those who argue that agent A can trust in agent B in the absence of belief (i.e., non-doxastic accounts of trust). As Keren (2020) points out, since trust is often formed on beliefs and those beliefs often find support in trust, it seems important to determine the exact relationship between trust and belief to understand why people would trust in other agents. In the following short discussion about what constitutes the justification for trust from the perspectives of a doxastic and a non-doxastic account, I present what can be considered the so-called "x-factor" of trust, and argue that determining the reason for trust also reveals what is essentially involved in trust as it emerges in the interaction between humans and robots.

According to Keren (2020), doxastic accounts argue that trust entails *beliefs* held by agent A about agent B as the reason for what agent B is being entrusted with. These trust-beliefs relate either to (i) the trustworthiness of agent B or (ii) whether the specific task that agent B is being trusted with will be done. As such, doxastic accounts of trust state that trust-beliefs provide necessary and sufficient criteria for trusting, so that trusting is simply a form of good believing[11]. The systematic relation between trust and holding certain beliefs that are not themselves trust-beliefs is at least one convincing argument for the acceptance of a doxastic account of trust, which can be shown in the specific case of trusting in testimony. Considering the logical relations between beliefs, much of the things that people come to believe stem from what others have told them, and the truth of what they say is what is being trusted in. It would be quite difficult to explain how trusting in the truth of what a speaker says supports our intuition of how testimonies gain their merit, without this trust also entailing the belief that this person is trustworthy. More generally, Hieronymi (2008) has been arguing for a strong doxastic notion of trust that entails a trusting belief as she writes that "one person trusts another to do something only to the extent that the one trustingly believes that the other will do that thing" (p. 214). Moreover, given that *trusting is believing* the normative evaluation of trust rests on its rationality – the assessment of good trusting (or well-calibrated trust) presumes that it follows the common epistemic norms (e.g., reliability, truth, coherence, knowledge). As Keren (2020) explains, those in favor of a doxastic account of trust conclude that

[11] There are also less strict (or "impure") doxastic accounts that only take trust-beliefs to be necessary, but not sufficient for trusting (Keren, 2014).

"if trust is a belief, we should be able to derive the conditions for the rationality of trust from the epistemological study of rational belief" (p. 111). Since rational beliefs are mental states that enjoy essential distinctive features, any doxastic account of trust must therefore be able to show to what extent beliefs have a primary place in the rationality of trust in relation to questions about evidential considerations, voluntary control, and logical standards (Keren, 2020).

Even though there are other noteworthy proponents of a doxastic account of trust in the philosophical literature (Adler, 1994; Hieronymi, 2008; Keren, 2014; McMyler, 2011), it is the influential work on trust by Hardin (1992; 2002) that provides the most comprehensive picture of such an account. From the following quote about his understanding of trust, Hardin (2002) argue that it is the belief in the trustworthiness of others that eventually determines if people decide to trust or not:

> "Trust is in the cognitive category with knowledge and belief. To say I trust you in some way is to say nothing more than that I know or believe certain things about you – generally things about your incentives or other reasons to live up to my trust, to be trustworthy to me. My assessment of your trustworthiness in a particular context is simply my trust of you. The declarations 'I believe you are trustworthy' and 'I trust you' are equivalent" (p. 10).

He also expresses that trust is rational, since it rests on the knowledge people have about others, which Hardin (2002) later develops into a theory of trust as "encapsulated interest". The rationality of trust, from his perspective, merely extends the principles of rational-choice theory because the belief of agent A about the trustworthiness of agent B to do C rests on the expectation of self-interest in establishing and maintaining long-term beneficial relationships. This *expectation* of self-interest, Hardin (2002) continues to explain, not only provides a strong incentive for a person to trust in others but also serves to help the *prediction* of trustworthiness in new or similar circumstances. As such, trust is being reinforced as people draw on the *accumulated knowledge* of both past experiences of interacting with trustworthy others and continuous updating of the assessment of their trustworthiness. In his argument for a more "street-level epistemology of trust", Hardin (1992) proposes that predictions of the trustworthiness of others are grounded in some common sense version of Bayesian generalization over prior experiences. Accumulated knowledge is very important to a doxastic account of trust because it provides the required evidence for believing that others are trustworthy or will do what they are entrusted with. Also known for his contribution to the rational-choice theory of trust, Coleman (1990) views trust simply as a form of decision-making under risk, where the risk that people take is contingent on the action(s) of others. To provide a solution to the problem of trust (i.e., whether or not to trust in others) as a basis for cooperative interaction, he suggests that people should make a cost-benefit analysis of the possible outcomes and act in such way that they are better off placing trust than not getting their needs and interests met. Coleman (1990) explains that trusting in others is, in this implication, a rational future-oriented belief about the *expected gain or loss*, guided by calculative reasoning similar to placing a bet against the odds. Consequently, in his view, the estimation of placing trust requires only that people be able to know exactly the potential gain, potential loss and the chance of others being trustworthy[12]. However, since there are cases where these variables are not always known or only known to a certain extend, it is according to Coleman (1990) rational for people to try to obtain or search for all the relevant and available information for their estimation before deciding to trust in others.

Even though doxastic accounts of trust are seemingly advantageous, as they can easily explain why people are right in their intuition that it would be rather irrational for agent A to trust agent B to do C if agent A has strong evidence for believing agent B to be untrustworthy, various alternative accounts of trust have recently been proposed in response to some of the objections raised against these accounts. Considered merely as the oppositional view, non-doxastic accounts of trust contest that trust itself involves or entails a belief (or has some belief-like properties). As Keren (2020) writes, there are at least two compelling arguments in support of non-doxastic accounts of trust.

First, from a closer examination of the relation between trust and evidence, a serious tension surfaces which is problematic if we are to accept that trusting is believing, and good believing usually requires complying with the standard of evi-

[12] Coleman (1990) presented the mathematical equation $p/(1 - p) > L/G$ to express when a rational person will place trust, where p represents the probability value of trustworthiness, L the potential loss, and G the potential gain.

dence norm. Whereas the securing or strengthening of evidence always improve the quality of beliefs, it turns out that this is not the case with trust. Differently from belief, “trust appears to exhibit a certain resistance to counter-evidence” (Keren, 2020, p. 114) because people often do trust in others or find them trustworthy while lacking favorable evidence, sometimes even in spite of evidence to the contrary. That every so often people trust without proper justification would hardly surprise anyone. Furthermore, trust is not like belief in terms of how much scrutiny it can sustain, which Baier (1986) formulated well when she wrote that “trust is a fragile plant, which may not endure inspection of its roots, even when they were, before the inspection, quite healthy” (p. 260). The PARANOID PARENT case by Wanderer and Townsend (2013) is helpful in understanding this intuition. While it might seem very rational for a paranoid parent to monitor the babysitter remotely via ’nanny-cam’ to obtain evidence of trustworthiness, it rather reflects a lack of trust in the babysitter. As they note, in the effort to gain the best available evidence for trust, the paranoid parent simultaneously undermines trust, because monitoring the performance of the babysitter ends up eliminating the possibility for the paranoid parent to be vulnerable to harm. To truly trust, according to non-doxastic accounts of trust, people might have to occasionally ignore counter-evidence for the trustworthiness of others or refrain from overthinking their trust or how trustworthy others are.

Secondly, when considering instances of trust, a problem arises from the observation that people can trust *at will*. Unlike belief, which is not subject to voluntary control[13], people can decide to overcome their initial hesitations to trust in others, or decide to abolish any doubts they might have in viewing others as trustworthy. The DRAMA CLASS GAME example that Holton (1994) uses in his work provides a good illustration of this intuition: in the moment of trusting your class mates to catch you when letting yourself fall, “it feels as though you are *deciding* whether or not to trust” (p. 63). Most people will be able to recall a situation where they had to consciously make up their mind about whether or not to trust others or consider others trustworthy. People trusting others at will, according to Keren (2020), are also found in the case of so-called therapeutic trust.

[13] Discussions about the exact relationship between the nature of belief and the nature of the will have not reached a final consensus because e.g., those in favor of doxastic voluntarism would argue to the contrary. My interest in this discussion is merely to explicate that certain conceptualizations of trust require an account of trust broader than belief (see e.g., the work by Frost-Arnold, 2014).

The idea behind therapeutic trust is to encourage trustworthiness by entrusting others with something important in the absence of any prior belief that they will rise to the occasion – trusting by itself serves to establish their trustworthiness. With her frequently cited TRUSTING MOTHER example, Jones (2004) ask us to imagine a mother trusting her teenage daughter to take care of the house over the weekend, with the hope of evoking her trust-responsiveness even though she has previously failed to be trustworthy. As Jones (2004) explains, it is possible for the mother to trust therapeutically not because her daughter has proven trustworthy, but rather given her normative expectation that her daughter will take good care of the house in the long run – the mother decides at will to trust in her daughter although she is not very confident in her trustworthiness[14].

If the reason for trusting in others is not a belief about their trustworthiness, what is it then? With the more ambitious aim to substantiate an non-doxastic account of trust, several views have been presented on what characterizes our reasons to trust in others or find them trustworthy. Baier (1986) famously argued that people trust in others given the assumption that they will act out of *goodwill* – they recognize and show concern for the well-being of those who are trusting them. In this sense, the trustworthiness of other people introduces a moral dimension to trust, because it emphasizes that their wish to avoid harming those who trust in them is based on the right motivation (i.e., the will to do good). While this proposed non-doxastic account of trust by Baier (1986) has contributed immensely to explaining why the reactive attitude of feeling betrayed is appropriate for people to have when their trust is violated, the presumption of goodwill was immediately criticized for being neither necessary nor sufficient for trust[15]. Modifying this view on trust as goodwill, Jones (1996) adds that the one trusting also has to be *optimistic* about the goodwill of the other person. This focus on optimism reveals that there is an emotive component to trust, she argues, because being optimistic is altogether a higher-order affective attitude which corresponds to that of hopefulness. However, the effort by Jones (1996) to defend this approach to trust as goodwill plus some optimism failed, because

[14] For a much more detailed discussion about the challenges of therapeutic trust to doxastic accounts of trust, please see also e.g., Frost-Arnold (2014); McMyler (2017); Pace (2021).

[15] Holton (1994) provided his CONFIDENT TRICKSTER and ESTRANGED COUPLE cases to show why it is possible for someone to presume the goodwill of others without trusting them. As McLeod (2021) points out, there has also been much discussion about how to even interpret the meaning of goodwill without such a concept being either too narrow or broad in scope.

it was vulnerable to other forms of counterarguments[16]. Although Holton (1994) agrees that there is a moral dimension to trust, he offers a different view on what counts as a reason for trusting. Applying the basic idea of reactive attitudes by Strawson (1962) to the analysis of trust, he came to the conclusion that people take a *participant stance* towards those who they trust or find trustworthy. Taking a participant stance, according to Holton (1994), means that people who trust not only see other persons as responsible for their actions, but also that they make themselves ready for the various kinds of feelings that result from their decision to trust (e.g., feeling hurt, resentful or angry when trust has been violated). As with the other proposals, there was a sense of dissatisfaction with the requirement of a participant stance for the justification of trust[17]. Hawley (2014,1) recently presented the view that it is an expectation that trustworthy people will follow through with the normative binding *commitment* to do what they are entrusted to do, which provides a reason for others to trust in them. She explains that in trusting others, such commitment must be of the right sort, and that it has to be broad enough in scope since it can be "implicit or explicit, weighty or trivial, conferred by roles and external circumstances, default or acquired, welcome or unwelcome" (Hawley, 2014, p. 11).

Even though there are many different non-doxastic accounts of trust, they have in common that they apply social and ethical norms as the criteria for what makes trust justified. From this perspective, the reasons for trust are not simply a matter of calculative, self-interested, and utility-maximizing rationality. Rather, it must be seen as something that goes beyond such rationality, as the various and complex interactions between people contain highly normative and ethical dimensions that are more often seemingly arbitrary from an external perspective because they may to a large extent be guided by the feeling, desires, wishes, and anticipations of people..

[16] From her work on how an overemphasis on individual autonomy ends up undermining trust between patients and health professionals, O'Neill (2002) challenged the view that an attitude of optimism is necessary for trust because people can be considered trustworthy merely from obligation or sense of duty.

[17] With her MAKING DINNER example, Hawley (2014) presents a case in which people take a participant stance towards one another but without it being a matter of trust. Jones (2004) showed that another problem with this view on trust is that it does not rule out instances of mere accident in which readiness to feel hurt or angry would no longer be appropriate.

2.3 Beyond Mere Reliance

In the ordinary use of language, it is normal to speak about how people in everyday life trust inanimate objects for various activities or goals with the mere assumption that they can depend on them to function properly. For instance, consider what is involved in preparing yourself for getting to work. You trust that the alarm clock or app will go off on time to ensure you do not oversleep, you trust that the cup containing your hot coffee will not break and cause a serious burn, and you trust that the chair you are going to sit on while eating breakfast will not suddenly collapse underneath you. However, it is not only those very simple objects like alarm clocks, cups, and furniture that you trust in your everyday life. You also trust that the car you are using to get to work will function properly and your drive will to be safe, you trust that your office computer will operate well so you can complete your tasks, you trust that the plane you are boarding for your business trip is not going to crash and end your life.

While examples are countless, the main message remains the same: Every single day we trust in numerous and various inanimate object to support everyday life to such an extent that our dependency on them hardly ever comes to our attention. Most of the time, we simply rely on the working of these inanimate objects because they form the unremarkable backdrop of everyday life activities, goals and tasks. The way people trust in artifacts and technologies is thus, on the most fundamental level, best characterized as an instance of a background relation (Ihde, 1990, p. 108) because interactions, engagement, or usage of such inanimate objects is something we have grown accustomed to.

2.3.1 Trust as Reliance

Considering these examples, the particular kind of relation people have to artifacts and technologies is best described in terms of reliance and understood as a certain form of dependency[18]. This dependency assumes that reliance on inanimate objects is necessary for the successful realization of some kind of

[18] It is important to point that there are in philosophical discussions different understandings of reliance that do not understand it as dependency, but rather to a mental state or attitude (Alonso, 2014). Consider, for instance, the example where Mary relies on John to drive her to work but is not dependent on John's doing so because she could otherwise take the train. Nevertheless, I will consider reliance as dependency because it is important due to its link to trust, which is of my concern in this dissertation.

plan given specific aims or goals. Why practical reasoning is central to the idea of *trust as reliance* is well explained by Smith (2010), who states that "in the absence of plans, there would be no relyings, and when an adopted plan is abandoned or completed, the relyings associated with that plan fall away just as do the associated intentions" (p. 137). From this perspective of plan execution, trust as reliance gains value mainly from its ability to guide thoughts and actions from a point of view that seems reasonable given the means adopted to meet the concrete ends, whether they are simple (e.g., reaching out with a hand to pick up a cup and take a sip of tea) or complex (e.g., the various and innumerable steps involved in finishing a longer research project). Consequently, trust as reliance cannot be understood solely as something internal to the person trusting the artifact or technology in question, but also hinges on the external conditions because it "is a relation that exists between the agent and the world" (Smith, 2010, p. 136). These conditions exist both in relation to the laws of nature and the constraints of a specific design.

While trust as reliance is "typically formed and revised in response to pragmatic considerations" (Alonso, 2014, p. 163), it is also important to keep in mind that this form of trust is closely related to the degree or conditions in which agency can be exercised. That is, there are certain situations where people must rely upon forces beyond their control (Smith, 2010) when trusting that e.g., the office building will not collapse, or that the elevator will not malfunction. These are situations in which the dependency is no longer voluntary and can as a result be experienced or regarded as problematic. However, because reliance on artifacts and technologies is free from any intentions from the inanimate objects to be reliable themselves, this also means that cases of malfunction, breakdown, or error are issues of disappointment. The fault or blame resulting from any serious consequences caused by unreliable artifacts and technologies has to be placed somewhere other than in the inanimate objects themselves. Accordingly, the main focus of trust as reliance is on making interactions with artifacts and technologies as smooth, efficient, and comfortable as possible, so that they may be considered only instruments or tools to help people achieve their aims or goals. This instrumental view is the most traditional and widespread understanding of inanimate objects and is also common to current understandings of AI and robots (Coeckelbergh, 2010a).

Trust as reliance within the specific context of robotics is understood as a predictive belief or assumption related to the robot's performance, given its intended purpose, relative to its specific task or predefined goal (Lee and See, 2004; Lewis et al., 2018). Given the actual performance of the robot, it is then possible to asses its trustworthiness, which is very important as it helps establish whether or not people are justified in trusting the robot in the given situation. From this perspective, the performance of the robot ensures that people are able to assess the appropriate level of trust during interactions, collaboration or engagement (Cai and Lin, 2010). This level is then treated as an indirect measure of trust, 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; Lee and See, 2004). It is for this reason that issues of safety become central to many discussions about trust in robots and their trustworthiness. Because humans who misplace trust could be exposed to serious danger or even risk their lives, safety measures and calibration of trust are also used as basic standards for evaluation (Freedy et al., 2007; Lindblom and Wang, 2018; Maurtua et al., 2017).

2.3.2 Apparent Agency of Robots

However, the instrumental view on robots has been challenged with the aim of making them more socially capable and human-like in both physical appearance and style of behavior.

Drawing on computational models of human cognition and social competence, "socially intelligent robots" (Breazeal, 2001; Dautenhahn, 2007) have built-in capacities to understand and display cues for social interaction and communication, similarly to how people naturally engage with each other. As such, they are able to behave and respond to people in a way that people might interpret as intentional, which influences how people approach and treat socially intelligent robots (Breazeal, 2003; Dautenhahn, 1995). In addition, the deliberate use of anthropomorphism as at design strategy to facilitate HRI only amplifies the tendency to perceive robots as more human-like (in both appearance and behavior)[19].

[19] While some might subsume socially capable robots under the general idea of anthropomorphic design, I decided to keep them as two seperate but interrelated aspects. To give an example of how a robot can be highly social but not very anthropomorphic, consider the TARS robot from the sci-fi movie Interstellar (2014). This robot is able to indicate high social intelligence solely through language use, by engaging in Theory of Mind and sarcasm. It has the appearance and behavior of a solid modular cube. In contrast, the nameless robot from the movie Robot and Frank (2012) looks and behaves very human-like but does not manage to escape the manipula-

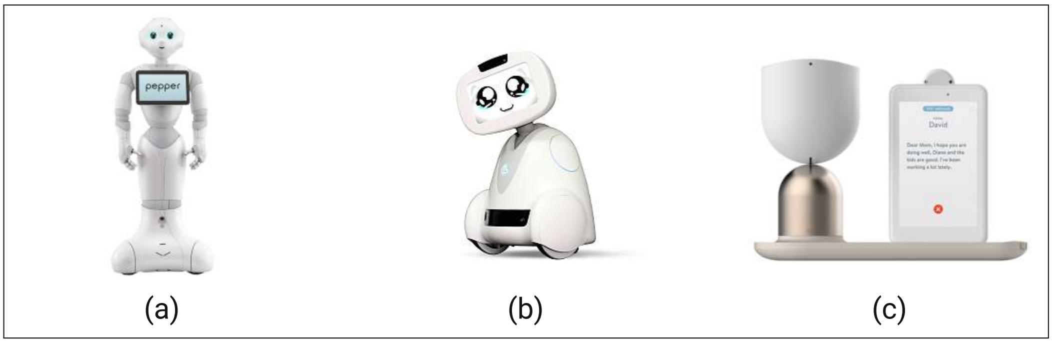


Figure 2.1: Three examples of robots using different degrees of anthropomorphic design: (a) the PEPPER robot by Softbank Robotics, (b) the BUDDY robot by Blue Frog Robotics, and (c) the ElliQ robot by Intuition Robotics.

The advantages of such a design strategy is that it allows people to use their highly developed social and cultural schemes to interpret their experience and perception of robots through these more familiar and intuitive channels or modalities (Złotowski et al., 2015). As such, socially capable and anthropomorphic robots are by design outliers in the discussion of anthropomorphism and the human disposition to project human capabilities onto inanimate objects. As with trust, the distinction between anthropomorphizing an object and experiencing an anthropomorphic design is important to keep in mind, because it is possible for people to anthropomorphize non-anthropomorphic robots or refuse to anthropomorphize anthropomorphic robots. Turkle (2011) illustrates this important difference with the example of a robot that speaks or cries out for attention, which is not animated solely based on anthropomorphic projections. These robots are experienced and appear to people *as if* they are sentient creatures through their display of social capabilities and human-like embodiment that comes in various degrees (see e.g., Figure 2.1).

Consequently, empirical studies have already suggested that people might interpret highly socially capable and anthropomorphic robots as some kind of "otherness" worthy of both social and ethical consideration (Coeckelbergh, 2010a; Kahn et al., 2012; Melson et al., 2006). Some have even argued that a new ontological category is now required to properly capture the experience children have with socially capable and human-like robots, as they perceive them as simultaneously animate and inanimate (de Graaf, 2016; Kahn et al., 2011). This

tion of Frank, which eventually results in the deletion of its memory (that marks the end of the relationship).

idea that robots are not just inanimate and passive machines is a further step towards the perception of robots as having some form of agency, which was first discussed with the “computers are social agents” (Lee and Nass, 2010; Reeves and Nass, 1996) design paradigm in the HCI literature. By now, there is an extensive body of literature in HRI focusing on studying the apparent agency of socially capable and anthropomorphic robots with an interest in how this influences their acceptance in everyday life and the willingness of people to interact with them (see e.g., Bishop et al. (2019); Goudey and Bonnin (2016); Zhang et al. (2021)).

It is important to already state here more explicitly why I have chosen to speak only of *apparent* agency in contrast to *actual* agency. From the very start of HRI developing into an independent research area, there has been a discussion about whether it makes an important difference if robots are in fact believed to be genuine agents or merely perceived as such, given their ontological status as inanimate objects (Takayama, 2012). While attempts to address such issues also depend largely on the definition of agency used, there has been a tendency in the HRI community to argue that only apparent agency is required for the aim of establishing and supporting interaction, collaboration, and engagement between humans and robots. Mainly concerned with the kinds of behaviors that are prompted and leveraged through social cues and signals displayed by robots, it does not really matter if robots are considered genuine agents by people, as long as they perceive and treat them as such for the sake of task completion or to meet their needs (Duffy, 2006). Given this pragmatic view, demanding that people in fact believe robots to have an inner life that manifests their agency is understood only as a sufficient criterion, but not as a necessary requirement for having successful human-robot interactions. As such, focusing on the human perception of robots as agents helps avoid any strong ontological commitments that follow from claiming genuine agency as seen in humans (or animals with high cognitive function). Such perspectives have also gained increasing support from more theoretical debates among philosophers. The work of Coeckelbergh (2011) has been very influential here, as he assumes a phenomenological approach: he argues the phenomenological perspective on human-robot interaction takes a non-traditional stand towards how we can understand and evaluate relations between humans and robots, as the focus does not lie on the ontological status of robots, but rather on how they appear to the human mind. This appearance of robots as agents is what matters, because the phenomenological perspective is focused on how the human experience is constructed through engagement with the world. As Coeckelbergh (2011) continues, the choices of developers and designers to endow robots with social capabilities and anthropomorphic design cannot be ignored in such discussions, because appearance strongly guides our inferences and actions in everyday life. While the debate about whether actual vs. apparent agency of robots should be a significant distinction for research on HRI is interesting, it does not undermine the point that apparent agency of robots must be a basic assumption underlying human-robot interaction. In my view, the apparent agency of robots enables us to grasp the uniqueness of human-robot interactions and precedes any additional proof of whether people in fact take robots to be genuine agents. If only for pragmatic reasons, we can stay agnostic about why people believe or do not believe that robots are genuine agents, as long as it is possible to observe that the consequences of more socially capable and anthropomorphic design of robots bring about human perception of them as having some form or degree of agency (a point I will elaborate on further in the following sections).

However, taking human perception of robots as more socially capable and human-like seriously also means that work on trust within HRI that rests on an understanding of trust as reliance is no longer sufficient for capturing the social dimension of such interactions, which also often extends to more ethical issues (Malle and Ullman, 2021). Human perception of robots as having agency has a direct consequence for how trust in HRI can be measured and evaluated, because the social competences and anthropomorphic design strategy affects what cues or features people include in their decision to trust or not to trust robots. By now, several studies have already suggested that a more human-like appearance also influences the extent to which people deem robots trustworthy or not (Stanton and Stevens, 2017; Złotowski et al., 2016). Moreover, it has been shown that the application of a trust scale developed for either a machine-looking automatic system or one developed for a more human-like robot yield different results during a social interaction scenario, thereby indicating that they cannot be used interchangeably (Kessler et al., 2017). Studying trust in HRI that is more tailored to interactions between humans and human-like robots, therefore, necessitates a stronger or more solid understanding of trust – one that is closer to the kind of trust people would place in each other.

2.3.3 Trust as Interpersonal

The work on trust (and antitrust) by Baier (1986) has without doubt set the tone for contemporary reflections on the philosophy of trust. She highlighted the specific social and ethical dimensions distinctive to interpersonal relationships, and from a critical analysis of previous philosophical accounts for trust that is mainly rooted in liberal tradition, argued that the significance of trust for thriving must be examined from a moral point of view. From her perspective, it is not constructive, for any understanding of trust pertinent to interpersonal relationships, to consider it as some form of contract, established between two equal parties in terms of both power and capabilities. From a careful observation of interpersonal relationships of all kinds in which cooperation and care is cardinal (e.g., that between parents/children, man/wife, caregiver/caretaker), she recognized that some of them are fundamentally unequal, and sometimes trust in such unequal interpersonal relationships is not motivated on a voluntary basis. Both are issues that severely challenge liberal ideals for the conditions of trust (a point I will return to in chapter 6). From this insight, Baier (1986) proposes to take trust to be a form of reliance in other people to act out of good will (in contrast to ill will) towards oneself. This demand of ethical consideration when acting out of goodwill is necessary, according to Baier, because of the risk and uncertainty involved in accepting this inequality of power or capacities, because the entrusted person can always decide to either honor or betray this trust.

This goodwill account of trust by Baier is not only important because it was one of the first views on trust that went beyond that of mere reliance, but also because she stressed the close connection between trust in interpersonal relationships and moral obligations. Since this initial proposal, others have extended the discussion on how to best distinguish interpersonal trust from that of mere reliance by considering the conditions of trust to include participation attitudes regarding praise or blame (Holton, 1994), normative expectations based on affective attitudes of optimism towards the goodwill and competence of others (Jones, 1996), personal and normative binding commitments to act through explicit promises or implicit encouragement (Hawley, 2014), and concerns of attachment with those we form relationships with (Kirton, 2020). While it is a long and extensive discussion to explain how these different accounts of interpersonal trust differ, they all agree on the fact that trust is not *just* a matter of practical reasoning deployed to achieve some gain based on predicative beliefs or assumptions. It is for this reason that philosophers like to distinguish interpersonal trust from mere reliance (Faulkner and Simpson, 2017), though the usages of trust in more ordinary language does not always adhere to this nuance (Tallant, 2019). Yet, it is an important difference, and Hawley (2014) summarizes the main point of the argument when she writes:

> "[...] we often rely upon inanimate objects but we do not grant them the rich trust we sometimes grant one another; inanimate objects can be reliable but not genuinely trustworthy. Moreover, our reactions to misplaced trust differ from our reactions to misplaced reliance. Suppose I trust you to look after a precious glass vase, yet you carelessly break it. I may feel betrayed and angry; recriminations will be in order; I may demand an apology. Suppose instead that I rely on a shelf to support the vase, yet the shelf collapses, breaking the vase. I will be disappointed, perhaps upset, but it would be inappropriate to feel betrayed by the shelf, or to demand an apology from it. Inanimate objects can be relied upon without being trusted" (p. 2).

Unpacking this line of argument, it seems that the analysis of what trust *is* in philosophical debates is restricted to relations only between people or animals that bond in social terms. Consequently, such a perspective precludes the placement of trust in inanimate objects and, by extension, also robots. The notion held by Hawley has also recently been expressed specifically in relation to robots as Lee et al. (2021) writes that "blaming and punishing one's robot vacuum cleaner for not cleaning the floor comes across as absurd – what ends would be served by blaming it and how does one go about punishing a vacuum cleaner? If a Roomba or other everyday technology does not work anymore, we do not hold it morally responsible or accountable for its dysfunction and one would normally not imagine ways to punish it" (p. 1). Following this reasoning, robots as inanimate objects would automatically be placed in the category to which only mere reliance could be assigned for the analysis of trust. Nevertheless, recent work on trust in HRI hass attempted to adopt the notion of interpersonal trust to better study trust between humans and robots (Ogawa et al., 2019) and as an explicit framework for the development of trustworthy robots (Lee et al., 2013; Wagner et al., 2018).

When speaking about interpersonal trust in the context of HRI, this refers to the consideration of aspects related not only to the performance of a robot, but also to those that are constitutive for relationships, which are influenced by apparent agency. Here, the attention is placed on how people come to consider trust in robots and their trustworthiness as a result of assumed motives or intentions underlying the performance or action of robots. The added dimension to interpersonal trust in HRI is the presumptions people have about whether robots will be concerned with e.g., their welfare, take their views and personal interests into account during decision-making, and work toward fair and unbiased outcomes. With these added social concerns of robots performing also for the good of the interaction, collaboration or even relationship, interest into how perception of responsibility and blame is attributed given an unfavorable outcome or situation has become a focus in recent studies on trust in HRI (see e.g., Komatsu et al., 2021; Van der Hoorn et al., 2021). These discussions bring forward the very ethical dimensions of how trust in robots and their trustworthiness extends to reflections on how people understand and value their interaction, collaboration, and encounters with robots. Whether trust violation instances by a robot will lead to the experience or evaluation of such situation as betrayal in the same manner as between people is a question open to empirical investigation, which is currently also a research gap in discussions on trust in HRI. But for now, the main point is to understand why the interpersonal trust in relation to HRI is different from that of mere reliance, and that it is closely related to the apparent agency of robots, enabling the perception of robots as having motives and intentions behind their actions. Without these perspectives, it would be irrelevant to hold robots accountable or blame them for any break of trust that people would place in them. In the current literature on trust in HRI, others have structured their discussion of this distinction using the same or other terms to denote opposite sides as they refer to e.g., performance vs. moral trust (Malle and Ullman, 2021), technological trust vs. interpersonal trust (van Straten et al., 2018), performance-based trust vs. relation-based trust (Law and Scheutz, 2021).

2.4 Trust-vulnerability Relation

The proposal to take into account the social and ethical dimensions of trust in HRI by applying the notion of interpersonal trust is valuable as a first step towards deepening our understanding of what is happening in interactions between humans and robots with apparent agency. This work aids in recognizing that there is an added layer of complexity, because it is no longer just a matter of performance, but also about what follows from HRI that leverages social rules and schemes to enhance the interaction.

2.4.1 The Property Approach

While it might seem rather straightforward to speak about interpersonal trust in the context of HRI, it does in fact require further and closer investigation. Comparing philosophical accounts of interpersonal trust (and other valuable work on the same topic from social sciences) to the level of technological advancement in robotics, Atkinson et al. (2012) set out in a panel discussion to address the important and also deeply theoretical question about the "appropriateness of using interpersonal trust as an analog for human-robot trust" (p. 306). As they explain, this has been argued by some to be a reasonable analogy on the grounds that some aspects of interpersonal trust also seem to be present in studies on HRI in which people came to treat robots as human partners. However, they also mention that others were not willing to draw such an analogy between interpersonal trust and that which could be used in HRI mainly because of the lack of reciprocity in the interaction. What is interesting about this objection is that such concern of reciprocity is a symptom of a more fundamental issue about the ontological status of the two kinds of agents involved. 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 of rejecting robots as suitable objects of interpersonal trust. It is appealing because the step taken 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 is 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 (T).
> **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 (T).

While different suggestions can be made concerning which properties are necessary and sufficient for members of the class of animate objects that belong to the category of objects considered suitable targets of interpersonal trust, there are at least two worth presenting here in short.

Central to the category of objects considered suitable for interpersonal trust is the requirement that all members of that class have the property of higher-level mental states (e.g., beliefs, intentions, emotions, desires). While people have such mental capacities, robots do not fulfill this requirement (as far as we know) because they are only capable of simulating such capacities (through, e.g., imitating, replicating, or mimicking). Even though simulations give off the appearance that robots have agency, it is not enough to upgrade the robots from the *as if* make-believe ascription to the richer social meaning of being *as* human from a strictly ontological point of view (Seibt, 2017b). Without any prospect of robots having the same mental capacities, they are not in possession of the right property and therefore dismissed from the category of interpersonal trust. Discussions about the possibility of developing robots with mental capacities similar to humans extends to the more classical debate regarding the programs of weak vs. strong artificial agents (Wooldridge and Jennings, 1995), a topic that has been of great interest to philosophers and computer scientists dealing with basic ontological questions (Dennett, 1994; Dreyfus, 1992; Searle, 1980; Turing, 1950) (which I also discuss further in chapter 7). While there are proponents on both sides of the discussions of such a possibility, no consensus has yet been reached, and it remains an open question to this date. Another property requirement of all members of the class of entities belonging to the category of objects that are suitable targets of interpersonal trust is the capacity for moral or ethical reasoning[20]. Again this property requirement is a serious challenge when considering robots, given their current state of development: robots are not capable of moral or ethical reasoning based on the concern of others. They might, however, be perceived to be, as they could be developed to operate with algorithms for decision-making and behavior that is aligned with ethical theories or moral principles (see e.g., the work by Arkin et al., 2019; Lindner and Bentzen, 2017; McBride and Hoffman, 2016; Vanderelst and Winfield, 2018). Such attempts to implement ethical or moral behavior into robots has been the standing aim of "machine ethics" (Allen et al., 2006; Anderson and Anderson, 2011; Wallach and Allen, 2009) and is further discussed in the new direction of "moral HRI" (Komatsu, 2016; Malle et al., 2015). Yet, with many technical and theoretical obstacles still to be tackled, it is not within the near future that we can expect to encounter robots capable of ethical and moral reasoning as presumed in interactions between humans. Because robots cannot fulfill the requirements of higher mental states and ethical or moral reasoning required as properties for the category of interpersonal trust, it seems that there are at least two ontological challenges that need to be overcome.

The tension created between the way people speak about trust in ordinary language (and in practice might not have any problems using when considering their interactions with robots), and the restriction in appropriate application regarding the category of interpersonal trust determined by the ontology, is a serious matter. It cannot simply be dismissed, because it can have serious consequences for the inferences people draw in their ascription of trust in HRI. Atkinson and Clark (2013) reflect on this conceptual challenges when they write:

> "[...] we are mindful that trust between humans and autonomous agents is not likely to be equivalent to human interpersonal trust re-

[20] To understand this requirement, imagine the case of a person (A) who trusts a hitman (H) for the task of killing her boss (K). While there might not be any expectations from A for H to act ethically, it is still possible to judge H as being trustworthy because this relates to H's reliability in carrying out K and not for H's ability to care about what the moral or ethical thing to do might be. As such, it is possible for H to be trustworthy in certain domains and not in others. If the trustworthiness related to whether A could trust H with the goal of keeping A alive, then H would be deemed untrustworthy. For this specific domain (i.e., trusting someone with one's life), being able to trust is influenced by the assumptions that (1) the life of others is of concern to H (i.e., thinking well of others) and (2) there is a commitment by H not to do harm (i.e., having a certain kind of motive for acting out of goodwill). These two assumptions are violated in the case of the hitman, therefore influencing A's trust in H when known.

> gardless of how ’human-like’ agents become in intelligence, social interaction, or physical form. Autonomous agents are not human, do not have our senses or reason as we do, and do not live in human society or share common human experience, culture, or biological heritage. These differences are potentially very significant for attribution of human-like internal states to autonomous agents. The innate and learned social predispositions and inferential short cuts that work so well for human interpersonal trust are likely to lead us astray in ascribing trustworthiness to autonomous agents insofar as our fundamental differences lead to misunderstanding and unexpected behavior. The foreseeable results could be miscommunication, errors of delegation, and inappropriate reliance” (p. 5).

Though the analogy helps us better understand the issues, this notion of interpersonal trust cannot be directly applied without violating the basic requirements of both parties to be ontologically equivalent, as they share the same properties. However, accepting an understanding of trust as mere reliance for the analysis of trust in HRI is not desirable either: while it may include certain significant characteristics of robots, it would exclude others central to the experience people have of apparent agency. When left unaddressed, discussions of interpersonal trust in the context of HRI force complex metaphysical deliberations about whether the relevant facts of ordinary usage, and the truth of the relevant pre-philosophical claims, require us to recognize the application of interpersonal trust to robots when accounting for the world and its workings. As such, considerations about whether it is appropriate to speak about interpersonal trust for HRI pose a challenge to the metaphysical theory of trust (and its variations) proposed by philosophers.

For those interested in the very abstract philosophical debate over trust, the discussion could continue with broader reflections on which grounds we should (or should not) extend the category of interpersonal trust by including robots. Yet, people eager to study trust in HRI might instead look towards a more pragmatic solution. They might prefer to know which implications such intricate philosophical discussions might have for their work on trust in HRI that is motivated and held to the standard of empirical investigations. To those interested in the exploration and analysis of the interactions between humans and robots, it is also an important task to account for what happens in spite of better knowledge, especially in those instances where the apparent agency of robots is not only reflected in their use of language, but also in their actions and behaviors. Tallant (2019) provides an interesting reply to the traditional analysis of trust in philosophy by arguing that people can have interpersonal trust in inanimate objects (using also the example of self-driving cars) if they treat the object as having agency, and that such a *possibility* should not be ignored as relevant to how to think about trust generally. As he remarks, while a philosophical analysis of interpersonal trust might imply that people cannot trust in inanimate objects, such analysis should not lead to ruling out the possibility:

> "[...] it would be inappropriate to trust the shelf to hold the vase; to feel betrayed by the shelf if the vase falls; to demand an apology from it; to blame the shelf: nonetheless, I do not see that as any barrier to doing so. I am capable of doing the inappropriate (as those who know me can testify), as are many others" (p. 2).

In the context of current work on trust in HRI, it might even make more sense to speak about interpersonal trust in robots, given the increasing body of empirical evidence suggesting that it is more common for people to treat robots as having agency than not – at least "in the moment" (Takayama, 2012).

2.4.2 The Event Approach

So what is the feasible solution to this ontological challenge, that does not dismiss the deeply metaphysical issues but at the same time remains relevant to the aim of studying trust in HRI through empirical investigation? I propose to shift the focus on trust in HRI away from only talking about the properties of the objects involved in the interaction (whether they are humans as animate objects or robots as inanimate objects) to instead considering the *event of interpersonal trust* itself. It is important to note that this suggestion does not mean that questions about properties are put aside. Rather, this new outlook simply extends the unit of analysis beyond the identification of properties (that are also sometimes referred to as attributes, features, characteristics or qualities) ascribed to either humans or robots, to a focus on the circumstance in which interpersonal trust happens. Replacing the common property approach to the study of interpersonal trust with one that understands trust as an event will allow for a broader perspective. It does not only considers who or what can be included in the category of appropriate subjects of interpersonal trust, but also takes into account the *conditions under which interpersonal trust occurs*. Taking the study of trust in HRI to be a matter of an event poses a new central question that is also open to empirical investigation, by asking whether the kinds of interactions that *happen* or *occur* between humans and robots could be characterized as instances of interpersonal trust. So although humans and robots are still ontologically different (due to their different properties), this broader perspective permits the study of trust to consider the properties of the objects as part of the event without making this category the dividing line for how we can discuss or consider trust in HRI.

From a methodological perspective, the important difference between the property approach and the event approach is that they operate with different criteria for the inclusion or exclusion of robots from the category of appropriate subjects of interpersonal trust. The property approach focuses on class membership of the right kind as the criteria, which leads to the exclusion of robots given the traditional analysis of interpersonal trust, as they belong to the group of inanimate objects. Considering instead the event approach, events typically contain a *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). Even though there is currently no agreement among philosophers about exactly what this principle is, it is considered a key element in most theories on events because it serves to denote the specific constrains that must be satisfied to identify a given event.

I argue that this Event Approach to the study of trust in HRI would serve the practical aim of bypassing the issues of ontological asymmetry between humans and robots (which is a complex and tangled metaphysical issue) while still allowing appropriate addressation of interpersonal trust, by placing the focus on the happening or occurrence. The happening or occurrence of interpersonal trust, however, is not only bound by the objects, facts, and relations characteristic of such events, but also by the *preconditions*. While agency, or at least apparent agency, has already been identified as a sufficient criterion of interpersonal trust, there are at least three preconditions that also have to be taken into account that all together form the basic necessary criteria.

To gain a quick understanding of these preconditions, consider the famous and stunning art performance *Rest Energy* (1980) by Marina Abramović and Ulay, which was first showed at ROSC'80 (see Figure 2.2).

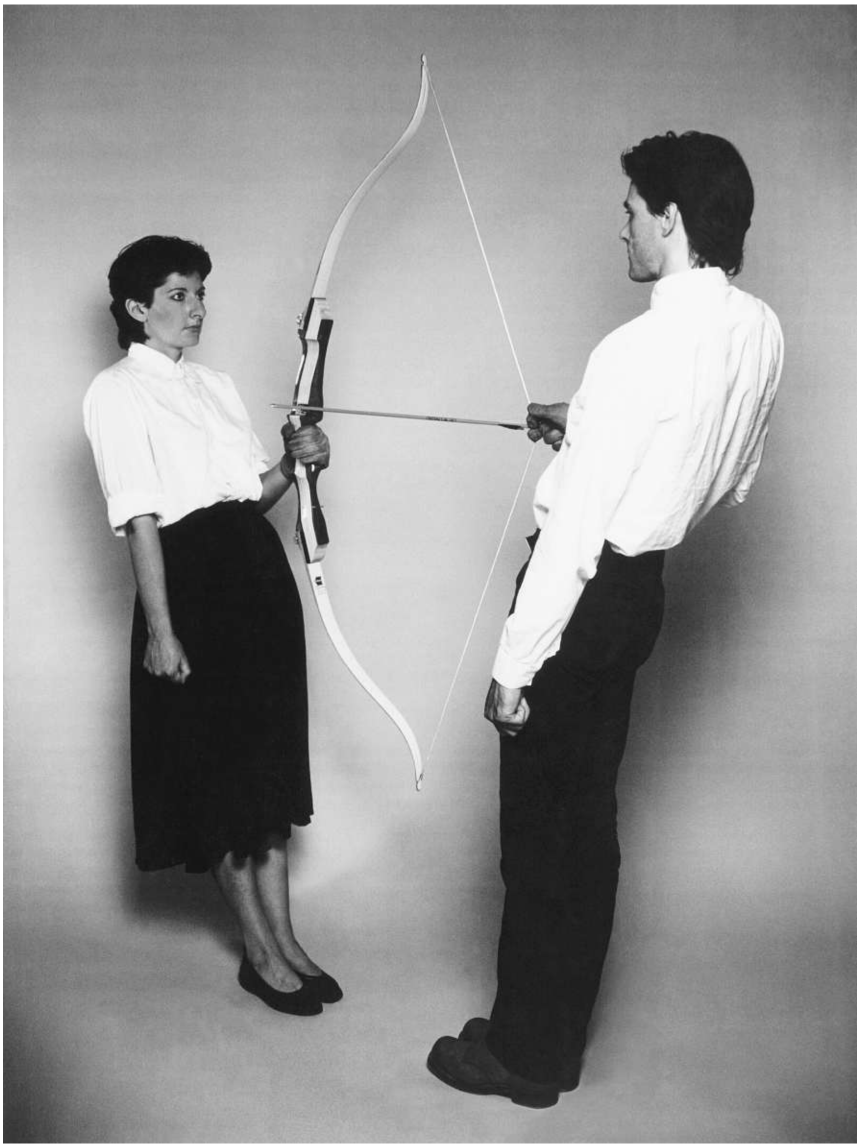


Figure 2.2: Abramović and Ulay performing *Rest Energy* (1980). Courtesy of Marina Abramović and Sean Kelly Gallery, New York Abramović (2016). DACS 2016.

In this piece, 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. For around 4 minutes, a strong atmosphere of tension is created, as any wrong movement or a lapse of attention could be fatal for Abramović due to the arrow pointed directly at her heart. No longer in control of the situation, she is left exposed. 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 in the specific circumstance: (1) when there is a possibility of harm (i.e., risk), (2) when there is a future-oriented likelihood of harm (i.e., uncertainty), and (3) when this exposure leaves people vulnerable (i.e., vulnerability). As interrelated preconditions of trust, what this art performance also illustrates is 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 both 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. The risk and uncertainty are evident to both of them.

CHAPTER 3

Emphasizing Vulnerability

> "We need to feel trust to be vulnerable and we need to be vulnerable in order to trust."
>
> Brown (2012, p. 47)

Part of sub-section 3.1.3 and 3.2.1 in this chapter will appear as a book chapter organized and edited by the co-organizers of the Trust Robots Doctoral College (TU Wien) as: *Hannibal, G. & Weiss, A. (forthcoming, 2022). Exploring the Situated Vulnerabilities of Robots For Interpersonal Trust in Human-Robot Interaction (pp. 1-19). Vienna, Austria: TU Wien Academic Press.*

In this chapter, I will provide an account of how vulnerability has been the least considered precondition of trust for studies on trust in HRI, and endeavour to explain why this has been the case. I will then go into more detail about how the notion of vulnerability has been considered so far when studying trust in HRI, and stress that it is important to consider the relational dimension for it to be an active precondition. At the end, I will present the different ways vulnerability has been conceptualized in other disciplines to provide a broader account of vulnerability as it is relevant for work on trust in HRI.

3.1 Strengthening trust in HRI

While I have touched upon the preconditions of trust in chapter 2, it is useful to include here a consideration of ways in which they have already been of interest to current research on trust in HRI to some extent. As I aim to show in the following sections, much work has so far been devoted to figuring out how best to strengthen reasonable trust in robots by minimizing risk and uncertainty. Vulnerability as a precondition of trust has been completely overlooked as a factor that could contribute to such efforts.

3.1.1 Minimizing Risk

As Wagner et al. (2018) point out, the question of how risk plays a central role to our understanding of trust in human-robot interaction is owed mainly to the embodiment of robots. Unlike AIs, robots navigate and operate in the same spatio-temporal reality as humans. As such, the actions of robots have the potential to cause serious direct harm to people in case of accidents, failures and when shown to be unreliable. Attending to the specific relationship between trust and risk to better understand how to strengthen trust between humans and robots, Wagner and colleagues highlight one precondition that might seem to many to also be the most important. For most people, debates about the more widespread use of robots in society quickly bring along the realization that robots not only provide new opportunities, but also pose new and sometimes hidden risks. From this point of view, trust in robots is motivated by the hope that it is possible to control or at least minimize possible risks in order to avoid harm (Coeckelbergh, 2013), especially since robots are often intended for high-risk domains of application such a warfare, finance, or healthcare – contexts in which trust in robots could turn into a matter of life and death (Wagner and Robinette, 2021).

Current literature on trust in HRI that aims to strengthen trust by addressing its relationship to risk fundamentally seeks to ensure safety: the focus so far has been on enabling robot designs that protect people from potential danger. Therefore, safety is understood as a key concern for trust in HRI, with especially the risk involved in over-trust having already been shown to be present in studies in HRI (see e.g., Aroyo et al. (2018); Booth et al. (2017); Robinette et al. (2016)). In this discussion, at least three different problems have gained a certain level of attention.

First, there is the issues of both physical and psychological harm. For example, Rezazadegan et al. (2015) explain how work into physical safety in HRI has been extensively explored for the industrial application of robots through the development of various risk-based safety analysis methodologies and by considering different design strategies for control systems. Given such assessment of safety, based on the analysis of risks, they provide in their overview examples of how physical safety in HRI revolves around the challenges of planning (i.e., covering safety criteria), strategies (i.e., reducing or preventing collision) and control (i.e., integrating proximity detection, collision avoidance, docking and compliance). Beside accounting for how physical safety in HRI has been studied, Zacharaki et al. (2020) add in their survey the more recent interest into the factors of psychological safety. In HRI, a focus on psychological safety seeks to study in which ways people find their interaction and collaboration with robots stress-free and comfortable, through either the behavioral adjustments of robot according to social considerations, or the particular features of robots supported by their embodiment. With the introduction of robots into everyday life, the additional aspect of psychological safety in work on trust in HRI will only grow in importance.

Secondly, problems of privacy and data protection have also been debated when it comes to risk prevention in HRI. According to Chatzimichali and Chrysostomou (2019), the three major challenges of privacy and data issues related to the use of robots are the "access of the company to personal data collected by the robot, sharing personal data with third parties and users' rights regarding derivative work" (p. 117). Focusing on commercially available robots, they argue that the current uncertainty of how companies are ensuring privacy and data protection puts people at risk, and that these legal issues need to be addressed in alignment with EU law. Also within a European context, Fosch-Villaronga et al. (2018) present in their work six different areas of privacy and data protection concerns for HRI that arise with the use of robots in healthcare. With their comprehensive account, they touch upon issues of confidentiality, induced trust, the nudging of disclosure, complexities of consent, conversational privacy management, data portability, and robot data collection.

Thirdly, the risks caused by biased algorithm-based learning and decision-making in robotics have been brought into the discussion lately, since unfair treatment of people in their own use or from application of robots in public domains (e.g., banking, healthcare, recruitment) perpetuates and in some cases causes societal issues. Howard and Borenstein (2018) highlight the problems of decision-making bias in the examples of robots used for peacekeeping or for healthcare, where the biases introduced can occur in various stages of the data set development used for training (e.g., determining the expert handling the data sets, the expert labeling of the data sets, how the images are secured, and what is considered a desirable data set output). In such cases, they write, using robots that reflect any biases will lead to undesirable results of stereotyping and discrimination of certain groups of the population simply because they belong to minorities (e.g., gender, ethnicity, religion, age). Focusing on the more concrete context of how unintended bias influences teams consisting of humans and robots, Rosenthal-von der Pütten and Abrams (2020) found that having a biased robot in a group setting is very problematic because its behavior can bring about "unequal treatment, intergroup bias and social exclusion of team members with severe negative outcomes for the emotional state of the individual and the social dynamics in the group" (p. 396). Those group members who were treated unfairly by the robot, as they explain, might not only be in risk of discrimination through the feeling of rejection and neglect, but also from the inability to address or cope with such situations when engaging in the group interaction.

3.1.2 Minimizing Uncertainty

When considering the precondition of uncertainty, the focus has so far been placed on more indirect strategies. Because uncertainty is mainly related to issues of unpredictability, the work aimed to minimize this liability to trust in HRI looks towards increasing familiarity through deliberate design choices. Familiarity is understood to be "an actor's close acquaintance with something" (Möllering, 2006, p. 94) and is believed to foster trust in dealing with uncertainty, because the process of familiarization enables people to suspend their doubt as they gain knowledge or experience satisfactory to their judgment. In this sense, familiarity serves as a means to draw on the past in order to feel more calm about the future. This is central to ensuring trust, as trust is about those risk people believe are going to come (Möllering, 2006). In other words, uncertainty can be minimized to foster trust because the process of familiarization enables some reference points for what people can expect from their interactions with robots. However, one of the main challenges facing the more widespread use of robots in everyday human life is that most people do not have much experience interacting with them. This creates a barrier that is not necessarily dependent on the specific design of robots in terms of their performance reliability, but might also have to do with more contextual factors (e.g., education, culture, subjection). Nevertheless, there are at least two broader design strategies that have been used in current work on trust in HRI to reduce uncertainty through the process of familiarization.

Among the most popular strategies to increase familiarity for trust in HRI is the aim to make the decision-making processes and behaviors of robots more transparent to the people interacting with them[1].While the are many different levels of abstractions on which transparency of robots might be required, the knowledge and understanding that is gained from such perspective is an increasing familiarization with how robots function or respond in a given situation of use. Exploring the use of robots in a healthcare scenario, Fischer et al. (2018) found that people with non-expert knowledge about robots gained better familiarity with the states, actions, and capabilities of robots when they were continuously provided with verbal explanations of what they were doing throughout the activity. Making this information more accessible to people, they argued, helped them reduce potential uncertainty of how the robot was going about the particular task, thereby ensuring a basic level of predictability. However, as pointed out by Wagner and Robinette (2021), the benefit of making information about the inner state of robots more accessible to people through explanations can easily tip the balance towards a false sense of certainty in the form of overtrust. Gaining more familiarly with the underlying function of how robots make decisions and act without explanations that actually support better understanding or knowledge, as they argue, could lead to an irresponsible acceptance of faulty robots in potentially dangerous situations. Another design strategy that has been popularly used for uncertainty reduction is the opposite of providing information access to the inner working of robots. Here, the focus is placed on the outer human-like appearance of robots to facilitate a more intuitive interaction, and the communication of social cues through non-verbal communication. As Mathur and Reichling (2009)

[1] The topic of transparency is currently receiving much attention in the AI and robotics community. This is not only because of the challenges of dealing with the epistemological issues of making transparent the underlying reasons and motives for particular decisions or actions (i.e., explainability), but also the more recent interest into the ethical issues of how to make transparent who (or what) is responsible in cases where robots causes serious harm to humans (i.e., accountability).

state, the use of anthropomorphic facial features to support the familiarization process with robots is important to consider for HRI research as people with little or no prior experience interacting with robots might infer their trustworthiness based on such exterior perceptions, which can be deceiving. While the manipulation of anthropomorphic features of robots did not directly affect the way people anticipated or attributed trust in robots, Christoforakos et al. (2021) did find in their study that such design cues have an indirect role in supporting trust by enhancing perceived competence and warmth of robots. However, as they write, because people might pick up the anthropomorphic design cues of the robot differently when given more variation in the visual appearance, the results suggest that the perception of human-likeness mainly has a positive influence when it is seen as an individual subjective perception. As such, the sense of familiarization that is created with the display of social cues through anthropomorphic robot design can be used as a way to reduce uncertainty when this falls into patterns from past individual experience.

3.1.3 Avoiding Overexposure

Even though the often cited definitions of trust by Lee and See (2004) and Mayer et al. (1995) used in HRI recognize vulnerability as an essential element of trust, this aspect has in fact been left rather unexplored (with few exceptions that I present in more detail in the next section). One immediate reason why vulnerability as a precondition of trust has been of little interest to roboticists is a simple technical matter. In their work, Wagner et al. (2018) take on a robot-centered perspective on interpersonal trust in human-robot interaction and propose both a conceptual and computational model of this phenomenon, with a special focus on perceived risk using game-theoretic elements and formal representation. While they are inspired by the popular definition of trust by Mayer et al. (1995), which places a great emphasis on the willingness of people to be vulnerable to the other party, they provide their own definition through a slight modification mainly for practical reasons:

> “Our definition differs from Mayer’s in one minor respect. Mayer characterizes trust as one’s willingness to be vulnerable. We replace vulnerability with risk only because risk is a more precisely and compu-

> tationally defined concept suitable for implementation on a robot." (Wagner et al., 2018, p. 4)

What Wagner and colleagues explain here is that they decided to use the notion of risk instead of vulnerability as a way to gain an operational definition of trust that they consider more useful for the context of human-robot interaction.

Considering a different reason, Cipolla (2018) correctly points out that 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), vulnerability 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 write 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 tool to empower 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 to reestablish a sense of security or protection. Characterizing this instrumental view on technology as a project of modernity, Coeckelbergh (2017) explains how the underlying assumption behind the development and use of Information and Communication Technology (ICT) reflects the agenda of vulnerability reduction. He 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" (p. 344).

What can be taken away from his account is that the perception of technology as a form of remedy to all the possible harms 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 hardly considered as something positive or worthwhile, unless it is merely to optimize our technological instruments and systems.

In HRI, focusing on vulnerability is also often considered somewhat problematic, but for a different reason. Through many years of ethnographic research into the way children and elderly 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 the basic needs of people, the bad association of vulnerability with technology is now related to danger of deception and its consequences for how people form emotional attachments. She 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” (p. 1).

The strong message by Turkle (2011) from this quote is not only that serious psychological harm can result from a false sense of intimacy when engaging with robots who seek to establish a emotional connection, but also that there is a level of enhancement involved in such kinds of interaction. Her work revolves to a large extent around presenting how the fascination with robots capable of imitating signs of care and love will eventually lead to unhealthy and inauthentic emotional attachments, because such technologies offer the possiblity to spare people the hardship and disappointment integral to developing deeper relationships with other people. By focusing on the vulnerability of people during human-robot interaction as a form of exploitation of both children and elderly who are in special need of care and love, a lot of effort has been put into better understanding and discussing what can be done to avoid people potentially being deceived by robots (Danaher, 2020; Grodzinsky et al., 2015; Sharkey and Sharkey, 2020).

This rather gloomy outlook on the role of vulnerability in our relation to robots is unfortunate when it comes to discussions of trust in HRI. Because vulnerability stands as one of the preconditions of interpersonal trust, aiming to avoid vulnerability or attempting to explain it away will paradoxically also undermine the demand for trust, as “in the absence of vulnerability trust is not required” (Misztal, 2011, p. 117). As Misztal explains, if vulnerability was not a 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, but which remains to be seen. This point is also well explained by Möllering (2006), who writes:

> “[...] 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*” (p. 6).

Central to our understanding of trust, as he shows, is that we are aware of our vulnerability but are able to interact and engage with the world anyway. I argue that this is similar when we aim to understand and study interpersonal trust in HRI. It is therefore important 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 is something that is not merely problematic, but could on the contrary also support the interaction and engagement with robots.

3.2 An Active Precondition

Understanding that vulnerability *is* a precondition of interpersonal trust makes it clear why it cannot be eliminated from the equation so easily, and why it has been included in most attempts to provide a definition. Still, there is not much work in HRI that has been focusing explicitly on the aspect of vulnerability for studies on trust between humans and robots. This has left a research gap in the current HRI literature on trust – I believe it is not only important and urgent to fill this gap, but also to explore further and deepen the way we can understand and study trust between humans and robots. As I will present in this section, those few studies that have lately been exploring the link between vulnerability and trust tend to take on a property approach and thereby miss out on two important points: firstly, that vulnerability in its relevance specifically for trust must be studied as something fundamentally relational, and secondly (and possibly even more importantly), that it can only come into play as an *active* precondition if both parties are able to accept and negotiate their vulnerabilities.

3.2.1 Vulnerability as Robot Self-disclosure

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

For example, Siino et al. (2008) found that a robot using affective disclosure during a collaborative task in a repair scenario resulted in people feeling less in control of their data, but also increased the robot's likability. Although this study is not directly about trust in HRI, it is still interesting because the findings could be understood as an expression of either human experience of vulnerability, or as 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[2] during a collaborative task in a navigation scenario led people to find it less trustworthy. As discussed by the authors, the tendency by people to negatively view others who constantly make apologies for themselves despite their intention of honesty, is an effect seen in HRI that sheds light on issues of distrust. However, there are also studies suggesting 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[3]

[2] I.e., “I think I am doing a bad job”, “I am disappointed in myself”, and “I think I should be doing better” (p. 852).

[3] E.g., “They reset my memory this morning, so my day has been a little rough”, “I get embarrassed when I need to ask someone to debug my program”, or “Sometimes I get lonely. I don't

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 made vulnerable statements during a collaborative task in a game scenario, its team members would display a much higher level of engagement with it. Extending on this work, Traeger et al. (2020) also found that the communication between team members would improve, and their experience as part of the group would be positive when the robot would provide statements of vulnerability.

While all these studies are important steps towards the inclusion of vulnerability for trust in HRI, they also fall short in two regards. First, none of the studies provides any definition of vulnerability (something I touch upon in chapter 2), but instead equates it to the errors or mistakes that a robot could make during interaction or collaboration. As such, it is an open discussion whether this proposed meaning is in fact useful for the aim of studying trust in HRI, which I believe is a highly relevant question. Secondly, 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 a specific situation or moment in time. As such, vulnerability is something that arises from the given circumstance around both the real and perceived vulnerability, depending on how the interaction plays out. As I wish to highlight in the following section, it is important to include the insight that vulnerability is relational in research on trust in HRI, because it is highly sensitive to the ongoing and ever-changing relationship between humans and robots during interaction.

3.2.2 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 the analysis of vulnerability in relation to technology, as an alternative to the more classical scientific approach. As he writes, the understanding that the classical sciences bring 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" (p. 38-39). From this perspective, he continues, vulnerability is something external to people, and

have many friends.") (p. 184).

can be evaluated from a third-person point of view and thus also characterized in objective terms. Vulnerability is *real* with regards to the possibility of risk and uncertainty as a threat to the livelihood or well-being of people. In this sense, the individual experience of being vulnerable is not taken into account or regarded as something that can be managed when understood properly. As Coeckelbergh (2013) explains, even those who do speak about vulnerability as tied to the subjective feelings or emotions of people are still presupposing that the perception of being vulnerable is seen in the light of an objective standard. Taking the completely opposite perspective would mean considering vulnerability only as subjective, where the first-person perspective is in focus – how the "I" comes to experience the vulnerability. This view, however, is also problematic, he argues, 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. What Coeckelbergh (2013) aims to challenge is in fact this overall idea of object-subject dichotomy in our understanding of vulnerability that is ingrained in 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." (p. 43).

From this critical discussion, Coeckelbergh (2013) elaborates on what he means when he takes vulnerability to be relational, that is: 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 or of the perceptions people have, as we have seen from previous work. Rather, it is something that must be located in the event of the meeting. What we can take as relational vulnerability in HRI is the co-constitution of vulnerability as a result of both the human and the robot coming into interaction or engagement. While Coeckelbergh (2013) puts a lot of effort into stressing the value of this analysis because it makes room for the existential dimensions of a "vulnerable being" (p. 44)[4], I 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 continuous and ongoing. Because vulnerability is relational in terms of interaction and engagement, it also means that it is always in the making. Coeckelbergh (2013) 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" (44).

Translating this insight into the context of trust in HRI, we can say that it is possible to consider vulnerability as a result that always occurs in the exchange between a human and a robot. While robots are a completely different kind (compared to humans), I believe that this does not hinder the recognition that they play their own and important role in the creation of vulnerability. Just as anything else in the world that confronts people as part of their everyday life, our meeting with robots has the potential to shape the way we come to experience and understand our vulnerability through the encounters. The same goes for the way that robots can be considered vulnerable in meeting with humans. They are also affected by the actions and behaviors of humans, even though the issues that robots face in such meetings might not have the same existential consequences. Nevertheless, potential risks and uncertainties that robots have to face when navigating in human spaces do exist, which render them vulnerable and therefore bring the theme of trust as bidirectional into the discussion.

[4] Or what I will discuss as an example of a universal human condition in the next section.

3.2.3 Vulnerability Recognition and Acceptance

Originating from her many years of research into shame as a social worker, Brown (2012) has recently also been known for her work on vulnerability as part of a larger discourse of resilience[5]. While her overall aim is to make people view vulnerability as a source for empowerment to deal with interpersonal relations and self-improvement in everyday life, her most important insight for understanding trust in HRI is the role that recognition of vulnerability plays in the meeting between humans and robots. Recognizing and accepting vulnerability is not a weakness, as the predominantly bad connotations may seem to indicate, but rather essential to an active and positive interaction and engagement with the world. As Brown (2012) writes:

> “Our rejection of vulnerability often stems from our associating it with dark emotions like fear, shame, grief, sadness, and disappointment – emotions that we don’t want to discuss, even when they profoundly affect the way we live, love, work, and even lead. What most of us fail to understand and what took me a decade of research to learn is that vulnerability is also the cradle of the emotions and experiences that we crave. Vulnerability is the birthplace of love, belonging, joy, courage, empathy, and creativity” (pp. 33-34).

By taking the perspective on vulnerability as something that can also foster a willingness to engage with others and the world, with the hope of finding deeper and more meaningful connections, an opportunity is created to view this precondition as part of a positive reinforcement of trust, rather than only functioning as a passive backdrop. Unlike risk and uncertainty, vulnerability does not link to trust in terms of management, regulation or control. As Brown (2012) argues, vulnerability is a part of human life that can bring people into interaction and engagement with others and the world, as the cause of both limitations and prospect. When vulnerability stands in relation to trust as a form of accepting an openness to harm or exposure, this allows for positive reinforcement in terms of benefits that both parties in the interaction or engagement can reap. The party

[5] The uptake of her work on vulnerability among the more general public is mainly due to her TEDx Talk “The Power of Vulnerability” from 2010 and Netflix Show “The Call to Courage” from 2019, both of which platforms seek to provide a venue for science communication.

trusting benefits from getting the support needed, while being trusted offers a chance to help others.

For the context of trust in HRI, this illuminates why vulnerability as a precondition is (metaphorically speaking) the key that unlocks trust as something that is co-created during the process of strengthening interactions and engagements between humans and robots. Vulnerability recognition and acceptance as an inherently positive part of trust in HRI leaves room for people and robots to gain the level of openness required for a successful outcome of a specific encounter. However, this also means that vulnerability, when recognized and accepted as part of trust occurring between humans and robots, is to some extend also bidirectional. Just as people must be able to trust in robots by being open to harm, robots are also open to exposure in their encounters with humans[6]. The bidirectional process of vulnerability recognition and acceptance in HRI, arising from the inherent co-creation of trust through positive reinforcement during the ongoing interaction and engagement, is important to keep in mind when considering the design space for HRI encounters that are trust-based. This is not only because vulnerability as a precondition to interpersonal trust can play an active and positive role in the meeting with others and the world, but also because it comes to show that vulnerability is an instance of constant negotiation. For the interaction and engagement between humans and robots to be labeled as an instance of trust, the vulnerabilities that are recognized and accepted emerge from the revealing or hiding of vulnerability by both parties during the encounter – finding the appropriate level of trust during HRI is a matter of coming to an agreement about which vulnerabilities can be on display without causing significant harm or exposure if violation does happen (a point I will return to in chapter 6).

3.3 Perspectives on Vulnerability

Similarly to trust, work on the topic of vulnerability is rather extensive and has been discussed in many areas of research. In her book on the sociology of vulnerability, Misztal (2011) presents a comprehensive overview of how the discussion of vulnerability originated, how it has changed over the last decades, and how it is currently shaped by new findings. While it is not possible here to go through all

[6] I further discuss this point in relation to the potential vulnerabilities of robots later on in chapter 6.

the important aspects of how vulnerability has been explored, it is important to establish a basic idea of how it has been developing towards more inclusion of social sciences, and why I draw on more recent discussions to make it relevant for trust in human-robot interaction.

3.3.1 From Hazards to Populations

Misztal (2011) explains that work on vulnerability was first concerned with managing the impact of natural hazards upon local communities through scientific and technological predictions and solutions. From the perspective of risk and disaster studies, the exposure to harmful events caused by natural phenomena (e.g., floods, cyclones, earthquakes and drought) was the main problem – through technocratic prevention, it was believed that people would become less vulnerable. However, as Misztal (2011) continues, such perspectives were often blind to the social, political, and economic dimensions that were key to the understanding of catastrophic events. This motivated new interest into better understanding the specific community or population impacted by the exposure in a way that did not treat them as a passive and heterogeneous group. Central to this movement towards vulnerability studies was the effort to better understand the circumstances that place people at risk, and the socio-economical and political factors that prevent people from responding well to the various exposures[7]. Sensitivity and resilience of people on all levels (i.e., individual, groups, community, state and societal) guided the policies and interventions developed to help people cope with their sense of vulnerability. Within the social sciences, and especially in sociological debate, there was also an increasing interest in the effects of globalization processes and the new kinds of risks people were facing (as I briefly presented in the introductory chapter 1) that also sparked a new interest into vulnerability studies.

While the original understanding of vulnerability has roots in discussions about the interplay between external exposure (from natural catastrophes) and proposals of various resilience strategies (whether mainly technocratic or more socio-political), a more recent interpretation has been surfacing in the psycholog-

[7] Misztal (2011) writes extensively on the issue of how attention to poverty has fostered some of the most important insights into current understanding of how the distribution of vulnerability is highly dependent on the social and economical injustice sometimes reinforced by political agendas held by governmental bodies or states.

ical literature where vulnerability is considered a form of feature characteristic of certain groups of the population. Again we learn from Misztal (2011) that another big interest in understanding and studying vulnerability was related mainly to issues of fragility, uncertainty, and lack of agency experienced among children and elderly people. In these discussions, vulnerability was associated with the protection and well-being of those in the population that were considered less capable in one way or another. Over time, the groups within a population that were considered especially vulnerable were also extended to those that were more exposed to social exclusion (e.g., the youth, mentally ill, to single parents, ethnic minorities, the unemployed and the homeless). As such, it is also common today to associate work on vulnerability with issues of insecurity, anxiety and powerlessness. From a slightly different perspective, Misztal (2011) mentions that fear has also been closely connected to the study of vulnerability, mainly in social and political theory. With the view that people are more at risk in late modernity, vulnerability understood to accompany fear has been centered in contemporary debates[8]. More interesting, however, is that the discourse on vulnerability mainly in terms of fear did not only motivate a renewed focus on vulnerability, but also motivated the development of "therapeutic culture" (p. 39) – a label used to describe the interventions, strategies, and scripts developed in a culture to help people deal with their experience of vulnerability caused by negative emotions and trauma.

However, as Misztal (2011) explains, the major challenge that this fear-driven account of vulnerability face is that it assumes that the risks people might experience are in fact things that they may also to some extent feel helpless against. As such, this understanding of vulnerability reduces it to concerns about human ability to exercise agency, and often leads to advice on how people should feel or act in light of risk, which then is considered a problem of how well individuals are coping. To avoid interpreting vulnerability as a serious limitation to the development of human agency because it restricts their self-defense to external forces, more recent work on vulnerability has proposed a broader view by stressing the universality of human vulnerability. An understanding of vulnerability from a more existential perspective, not only advocated in social and political theory but also in many ethical and moral discussions, stresses the importance of vulnerability

[8] However, it is important to Misztal (2011) to also point out that while this account of the sociological interest into the relation between fear and vulnerability is new, it was recognized long ago in political theory in the work of Hobbes.

as something fundamentally connected to the human condition (Misztal, 2011).

3.3.2 The Human Condition

Fineman (2008) is one of the influential proponents understanding vulnerability as part of our human condition, as an argument for the development of a more equal, inclusive, and responsible nation state. As she argues, the problem with the widespread conceptualization of vulnerability as something characteristic of a vulnerable population is that it also contains a form of stigmatization in terms of "victimhood, deprivation, dependency, or pathology" (p. 8). Confronting the negative associations surrounding the notion of vulnerability, Fineman proposes instead to recognize that people are always in some form dependent on various social, legal, and political structures or institutions to meet their basic needs. So while some people might have special needs (i.e., the more vulnerable group of the population), all humans share an existence as subjects who require the attention and care of other people or the supportive structure providing their everyday life in a world of constant risks and uncertainty. It is from this perspective that Fineman (2008) introduces the idea of the "vulnerable subject", which she uses to capture the universal aspect of being human over the span of a lifetime where the ever-present possibility of being harmed – whether unintentionally or intentionally – is not under our control.

The philosophical analysis of the specific connection between vulnerability and the human condition originates from three different directions, which Mackenzie et al. (2014) accounts for very well. As they write, the first main area in which vulnerability as part of the human condition came into focus was in feminist theory, where attention was placed on the issue of dependency and the ethics of care as a universal element of living together in communities, states, and larger societies. These feminist perspectives motivated a critical stance towards mainstream political and moral theory, which were argued to be blind to the normative dimensions of vulnerability. By pointing out very clearly that the human condition of vulnerability cannot be separated from the universal dependency that people have on other people or the state, those advocating more feminist theories have seen it necessary to ensure that there was always a strong moral obligation to protect the vulnerable against suffering resulting from lack of care. How vulnerability is connected to the human condition has also been of great interest to the field of bioethics and was later included in the core principles of clinical research ethics. In the medical context of providing and receiving care, the focus has been on ensuring that people are not left vulnerable to violations of their basic human rights of autonomy, dignity, and integrity. Because everyone has a body that throughout life will be in need of healthcare to some extent, this view acknowledges that the vulnerability of people must always be taken into account as a universal factor in order to maintain their well-being in treatment, trials, and assistance. Grounded strongly in the idea of corporeal vulnerability and its ethical implication as developed in the work of Butler (2006), the third area in which vulnerability is being associated with the human condition is in it role as a form of ontological foundation of humanity. Accepting as a basic and universal premise that the embodiment of human life and the vulnerability that people come to experience often arise from the various different responses others have to their body, this perspective enables an understanding of not only the good (e.g., love, care, generosity) but also the bad (violence, abuse, contempt) aspects of humanity. This identification of the inherent ambivalence of human nature introduces to the discussion the importance of how vulnerability is entangled with the self-other relationship, emerging in encounters with other humans. Consequently, this perspective on vulnerability as a human condition stresses that the continuous impact and influence people have on each other through their actions implies an ethical obligation to mitigate the suffering caused by misfortune and structural inequalities identified as part of everyday life.

I must add to the analysis by (Mackenzie et al., 2014) of how vulnerability tends to be discussed in philosophy in relation to the human condition, that similar perspectives can also to found in philosophical debates about the promise of the transhuman project of “enhancing the human condition and the human organism opened up by the advancement of technology” (Bostrom, 2005, p. 3). Hauskeller (2019) considers the whole transhumanism movement rather problematic because it seeks to erode what ultimately defines the basic condition of our humanity – our recognition that death will eventually arrive and that this pushes us towards others in the hope of a valuable and meaningful life. In this sense, he speaks of vulnerability in existential terms, as it comes to represent the universal condition of human life. In their critical analyses of transhumanism, Coeckelbergh (2013) and Liedo and Rueda (2021) bring to our attention that the desire to enhance our human nature through technology will, despite great success, never be able to take away vulnerability as part of the basic human condition. As they argue in different ways, while some of the human vulnerabilities, which advocates of transhumanism today wish to make obsolete, might be overcome in the future, it will only result in the creation of new ones that we are not able to predict or could even imagine. As such, these more nuanced voices in the debate do not dismiss the project of transhumanism, but instead show that vulnerability as part of our human condition will merely transform alongside the transformation we will see in human nature. Whether these are going to be favorable or not depends on the many choices that will be made along the way.

From this short account of how vulnerability is also part of recent discussions about the human condition, it becomes increasingly evident that it is a notion that is not only relevant to the analysis of trust, but also to a range of other neighboring concepts including e.g., need, dependency, violence, care, exploitation, human nature, and love. This is important to keep in mind when studying trust in HRI, because it is easy to get absorbed by the theme of vulnerability if the focus does not remain on its particular role in interpersonal trust as one of its preconditions. I will do my best to ensure that the focus is on this relation. However, what is valuable from this particular philosophical perspective is that it highlights that vulnerability constitutes a fundamental aspect of the human condition, whether due to our dependency, needs, or relation to others. As such, the experience of vulnerability is something that everyone will be familiar with at some point and level during their lifetime, regardless of whether this is further amplified by belonging to a vulnerable group of the population. This acknowledgment of vulnerability as something relevant to everyone is something that I will pursue in the rest of my dissertation, since it provides a starting point for the study of trust in HRI that is most relevant to the application of robots in more everyday life domains where the envisioned target population for using robots is considered in terms of the ordinary or lay person.

3.3.3 Imperfect Robots

From the above account of vulnerability as closely related to the universal human condition, it might seem a rather big jump to now continue to speak about robots. The objection that comes to mind immediately is that robots, as inanimate objects, do not have an inner life that is able to experience suffering (or anything at all) due to harm. One might ask: in which sense can vulnerability be linked to discussions about robots? While it is not the point to dispute the fact that robots are not sentient beings in which vulnerability relates to any subjective experience (a point I will also return to in more detail in chapter 6), robots are limited in their capabilities and their ability to deal with exposure that can compromise their proper functioning. Since the functioning of robots depends on various intertwined mechanical, electrical, and computational components, the space of vulnerabilities related to robots specifically is rather sizable. By drawing mainly on insights from literature on security issues in robotics and failures of robotic systems, I will briefly try to show that vulnerability is also a useful notion from a more robot-centered perspective when considering that robots are far from perfect.

The predominant and maybe also more typical way of discussing vulnerability in relation to robots focuses on the various security issues of e.g., safety, cyber hacking prevention, or privacy protection – general areas of concerns in computer science research. Because robots are nothing more than embedded computer systems, they also face the challenge of withstanding threats or attacks from external forces aiming to compromise their operation and secure usage. Working towards the introduction of more socially capable and anthropomorphic robots for application in both the public and private life of people, the priority of security measures is expected to become of even greater importance in the near future (Clark et al., 2017). However, as Kirschgens et al. (2019) explain, the rush to bring service, personal, and collaborative robots into the fast-growing consumer market with the aid of new developments in the Internet of Things (IoT), Industry 4.0, and Cloud Computing has resulted in an inadequate consideration of security issues that could compromise the safety of humans. These consequences, they write, may be even more challenging than the ones identified in the development of personal computers back in the 1990's, or in the more recent revolution of smart-phones in the 2000's. As Kirschgens et al. (2019) point out:

> "[...] when a PC is hacked, the output damage usually remains virtual and, although linked to reality in many aspects, the direct consequence of the breach generally stays non-material. Meanwhile, when a robot vulnerability is exploited, apart from this privacy violations, data or economic losses, there is another major effect to be considered: physical outcomes by robot malfunction. Robots can harm people and things. And this is why there is an urgent need for rethinking how we protect robots" (pp. 1-2).

The increasing demand for robots to rely on open source and to be better embedded into various technological infrastructures used by people with limited expert knowledge, as they stress, is something that reveals the serious security vulnerabilities in current robotics. Hardware, firmware/OS, and application attacks are presented by Clark et al. (2017) as the three main target layers when categorizing security issues for robots. Aiming to provide an open and free tool to better identify the various security vulnerabilities of robots, Vilches et al. (2018) developed a “Robot Vulnerability Scoring System (RVSS)” that takes into consideration aspects of safety, assessment of downstream implications, library and third-party assessment, and environmental variables. The potential security risks resulting from hacking robots were explored by Wolfert et al. (2020) in the particular context of social robotics. After they had performed an initial cybersecurity vulnerability analysis of two different commercially available social robotics platforms, Miller et al. (2018) found that the main challenge to ensuring secure use was connected to the lack of authentication. In their proof-of-concept study, they showed how social pressure through the control of a social robot led some people to compromising access to guarded spaces, sharing sensitive and personal information, and entering unsafe situations. As they conclude, it is important to further address security issues in future research on trust in HRI, because robots vulnerable to hacking with the intention of misuse might be harder to spot in contexts where the social capabilities and anthropomorphic design can be exploited to influence people. By now, growing awareness of the many security issues related to the use of robots for application in more everyday life contexts have also lead to further discussions around the legal implications and policy considerations required to make HRI more safe (Fosch-Villaronga and Mahler, 2021; Lutz et al., 2019; Subramanian, 2017).

Shifting attention towards a different way in which vulnerability is being discussed in HRI, there has been a recent surge of interest in how the unexpected behavior of robots influences the perceptions and feelings people have towards them. How people are able to understand and handle various errors, mistakes, failures, and fault caused by robots is the focus of such work, and is believed to play and important role in the decision of people to accept, interact with, and use robots. Honig and Oron-Gilad (2018) have provided a very extensive overview of how robot failure has been studied in HRI to this date, and I believe that such work is relevant to research on trust between humans and robots because there are two main responses people tend to have when it comes to understanding or perceiving robots as imperfect. Given that robots are often represented in both pop-culture and media as being very capable (e.g., having high intelligence, easily manipulating objects, engaging in conversations and interactions, navigating in unstructured environments), the first confrontations with the limits of such robots when having to operate in the real world very often results in disappointment or frustration to those not so familiar with the state-of-the-art in robotics (Bruckenberger et al., 2013). As Honig and Oron-Gilad (2018) point out, it takes very little time before error or failures occur, which also happens rather frequently when robots are being tested or deployed in real life situations despite great effort being put into ensuring their proper functioning. Consequently, there is often a big gap between the expectation people have towards robots and the reality of what robots can in fact do – a challenge that is in referred to in HRI literature as the "expectation gap" (de Graaf et al., 2016; Kwon et al., 2016). According to Henschel et al. (2021), the unrealistic expectations towards robots are further reinforced by the desire to design them as more and more human-like, in both appearance and behavior. While it may be possible to manage the expectations of people, the more serious consequence of the expectation gap is that the unexpected behaviors of robots can be interpreted as incompetence (Cha et al., 2015; Salem et al., 2015b). Robots might simply be perceived as not up to the tasks that are required of them, which can lead to disuse or even questions about whether they are useful for everyday life activities at all. As such, the imperfection of robots caused by recurring errors, mistakes, failure and faulty behavior can result in a negative attitude towards their potential use in society more widely. However, as I will argue, another kind of response to the imperfection of robots turns out to be quite the opposite. Sometimes the errors, mistakes, failure, and faulty behavior of robots increases their likability and familiarity, because people perceive this imperfection as something that makes them seem more human in a way (Gompei and Umemuro, 2015; Ragni et al., 2016; Salem et al., 2013). In the work of Mirnig et al. (2017), this notable observation was explained in terms of the Pratfall Effect: it is common for people to consider imperfection in other people as an indication of approachability that also make them more attractive. With this understanding, Mirnig et al. (2017) set out to explore in a study whether this social psychological effect could deliberately be used to make interactions with robots smoother and influence whether people found them more believable.

They point out that the study shows that whether the imperfection of robots is considered an expression of human-likeness rests heavily on whether they are able to also communicate about errors, mistakes, failures, and faulty behavior through social signaling. Taking this discussion into a philosophical reflection about the possibility of robots as companions, Coeckelbergh (2010a) argues that this all depends on whether robots are able to mirror human vulnerability, as this will instigate in people a feeling of empathy towards them. His argument for why this vulnerability mirroring is crucial for developing a relationship of companionship with robots lays in his understanding of vulnerability as a basic human condition (as discussed in the previous section). Coeckelbergh (2010a) puts into words this point as follows:

> “Our embodied existence renders us vulnerable beings. Human empathy is partly based on the salient mutual recognition of that vulnerability: this is what we (among other things) share as humans; this is what makes you ‘like me’ or ‘one of us’. In this sense, we are each other’s ‘vulnerability mirrors’. We can feel empathic towards the other because we know that we are similar as vulnerable beings. If we met an invulnerable god or machine that was entirely alien to us, we could not put ourselves in its shoes by any stretch of our imagination and feeling [...]” (pp. 6-7).

What we can learn from this interesting perspective by Coeckelbergh is that robots, who will eventually be able to display and communicate their vulnerability, could in fact help build stronger and more authentic relationships between humans and robots[9]. As I will explain in more detail in chapter 6, which explores robot vulnerabilities through expert interviews, I believe it is necessary here to already emphasize that this idea of vulnerability mirroring as a way of speaking would lose its value if vulnerability was understood as robots having to mirror the exact same vulnerabilities that are characteristic for humans. In this sense, I will extend this idea presented by Coeckelbergh (2010a) through my own empirical work on how vulnerability also comes to play an important role in our study of trust in HRI.

[9] Yet, because this also opens up a whole landscape of ethical problems, he also makes sure to explain that while this helps us better understand why it is possible to have robot companions, it does not mean that it is in any way what is desirable.

3.4 Guiding Trust Definition

Möllering (2006) writes that the topic of trust continues to fascinate and lure new generations of scholars across disciplinary boarders despite the vast amount of literature already devoted to the understanding and analysis of trust. In his view, one reason for this appeal is the power allotted to interpersonal trust, as it is (metaphorically speaking) often considered the invisible glue enabling and maintaining positive interpersonal relationships, cooperation and social cohesion[10]. The important role trust plays in uniting humanity into a common life was already clearly expressed by Simmel (2004) who observed and noted that "without the general trust that people have in each other, society itself would disintegrate, for very few relationships are based entirely upon what is known with certainty about another person, and very few relationships would endure if trust were not as strong as, or stronger than, rational proof or personal observation" (pp. 177-178). While the *value of trust* is widely agreed upon, as it has very positive effects on human well-being and smooth societal functioning, only little consensus has formed on what trust means. To this day, it is still hard to find a good and comprehensive overview or model of trust in current literature, even though getting a clear picture of the phenomenon is recognized as necessary. Determining what trust *is* and how it can be studied empirically is not as straightforward as it might seem at first glance.

Like with many other abstract concepts (i.e., justice, money, love, rules, rights, freedom etc.), trust is hard to grasp because there is nothing in the world we can simply point at to establish its existence directly. That is, there is no one-to-one correspondence between our concept of trust and it manifestation in the world. Instead, what constitutes trust is a complex collection of ideas that

[10] As Misztal (1996) explain, the original understanding of trust was in fact deeply integrated with religious debates about belief, because it was used in speaking about the dependency of humanity upon the mercy of a benevolent God or goodly powers. In this religious context, trust was closely associated with faith because human existence has always contained elements of vagueness and contingency outside our control. Human faith in higher order and justice as central to religious practice places trust as a strong motive to accept that which utterly escapes understanding, but nevertheless could be a strategic choice in dealing with risky and uncertain moments in everyday life. This line of thinking was even carried on to modern sociological debates as Giddens (1990) expressed that trust is a form or faith that has more to do with commitment than a simple cognitive achievement of risk assessment. Möllering (2006) along similar lines goes when he argues for a conceptualization of trust as a "leap of faith" that expresses the state of suspension that trust enables when people accept their vulnerability.

combine to create its rich meaning. Consequently, trust can only be investigated indirectly. Thus, on a theoretical level, we need to carefully study which ideas are relevant and how exactly they relate. However, the work of trying to grasp the idea of trust by teasing out the various conceptual elements and relations is often considered a problematic task because there is no unified understanding of trust, despite extensive and substantial theoretical work in a range of disciplines (e.g, philosophy, sociology, economics, political science, psychology, medicine, organization studies, cognitive science, risk management, and computer science). Various accounts of trust have been proposed: below, I list a selection of popular definitions, some of which have also been used to guide current understanding of trust in HRI:

> "[...] *a particular level of subjective probability with which an agent assesses that another agent or group of agents will perform a particular action, both before he can monitor such action (or independently of his capacity ever to be able to monitor it) and in a context in which it affects his own action*" (Gambetta, 1988, p. 217).

> "[...] *willingness of a party to be vulnerable to the actions of another party based on the expectation that the other party will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party*" (Mayer et al., 1995, p. 712).

> "[...] *psychological state comprising the intention to accept vulnerability based upon positive expectations of the intentions or behavior of another*" (Rousseau et al., 1998, p. 395).

> "[...] *I think it is in your interest to take my interests in the relevant matter seriously in the following sense: You value the continuation of our relationship, and you therefore have your own interests in taking my interests into account. That is, you encapsulate my interests in your own interests*" (Hardin, 2002, p. 1).

> "[...] *the attitude that an agent will help achieve an individual's goals in a situation characterised by uncertainty and vulnerability*" (Lee and See, 2004, p. 51).

Lacking a common understanding of trust that can also be expressed in a clear definition has often led to the permission of vagueness and imprecision. This tendency to leave aside the demanding conceptual work in research on trust is well stated by PytlikZillig and Kimbrough (2016) when they write:

> "The frequent complaints about the continued "elusive" nature of trust and the lack of an agreed-upon definition for trust are not entirely unfounded, but they also seem to be somewhat self-serving and misleading. That is, such claims may serve as rhetorical devices to underscore the importance and difficulty of one's topic of study, while at the same time providing authors with full license to define trust however they like" (p. 41).

While they acknowledge that the task of getting to the nature of trust is challenging because of the many suggestions provided so far, neglecting conceptual clarification will only lead to further confusion and inconsistency. The absence of a systematic understanding of trust to provide a basic foundation will eventually also undermine the very attempt to get a better understanding of trust through empirical work, as there will be doubt about whether what is being measured is actually appropriately related to the concept. In this sense, a wide range of conceptual flexibility may be necessary and valuable to show the complexity of trust, as various fields of research approach the topic differently. Yet, it is also worthwhile to attempt to suggest at least some kind of conceptual overview of trust for the sake of clarification and deeper understanding. As Castelfranchi and Falcone (2010) argue:

> "[...] the fact that the use of the term trust and its analytical definition are confused and often inaccurate should not become an unconscious alibi, a justification for abusing this notion, applying it in any *ad hoc* way, without trying to understand if, beyond the various specific uses and limited definitions, *there is some common deep meaning, a conceptual core to be enlightened*" (p. 7).

With my disciplinary background in philosophy, I strongly agree with their call for a more general and domain-independent account of trust, though I have no ambition of comparing my own attempt with the extensive work they have been carrying out in their proposal of a computational model of trust (Castelfranchi and

Falcone, 2010). For the purpose of my PhD project and in writing up this dissertation, my aim is rather to zoom in on a very specific trust-vulnerability conceptual relation as a starting point for my own investigations into trust in HRI. Since there are many ways to cut the conceptual cake when constructing a definition of trust, my own attempt is to be considered only as an initial step towards a working definition, which could also prove useful to other HRI researchers working on the same topic. After the basic relations, objects, and metaphysical commitments of the trust concept have been presented in the previous chapter 2, I propose the following working definition of trust:

> *Given the possible risk that agent A might get harmed if agent B will not do C and uncertainty about whether agent B can and will do C, agent A trusts agent B to C in S if and only if agent A voluntarily agrees to accept vulnerability.*

Where:

> Agent A = any kind of autonomous agent able to trust
> Agent B = any kind of autonomous agent able to perform a specific task
> C = any kind of act or conduct in a certain domain
> S = any trust situation located in a space-time region

The point of this very general definition is that it very clearly explains the relationship between the different preconditions of trust (risk, uncertainty, vulnerability) and trust as defined within mainstream philosophy (A trusts B to do C). As such, my proposed guiding definition of trust captures the vulnerability-trust conceptual relation by showing how a high sense of risk and uncertainty results in vulnerability that instigates a demand for trust based on the voluntary agreement to be open to harm.

While this definition could, in principle, be extended to also incorporate the different stages of trust (i.e., going from the disposition or attitude to trust, to the decision or intention to trust, to the announcement or act of trusting) to help highlight the dynamic aspect of trust (Castelfranchi and Falcone, 2010), such a broad definition would require that I provide a longer additional account for how other relevant conceptual relations to central or neighboring concepts (e.g, delegation, autonomy, dependency, evidence, stakes, intention) would fit into such conceptualization. While I made a very rough sketch of a conceptual model to consider the wider range of conceptual relations involved in the trust concept for my own clarification, I believe that Castelfranchi and Falcone (2010) have already done sufficient work to explain what such a picture could look like[11]. My own goal is simply to show that the definition of trust should also take unto account its preconditions, and that it is the exact conceptual relation between trust and vulnerability that I will be limiting myself to for the remainder of this dissertation (see e.g., the Figure 3.1).

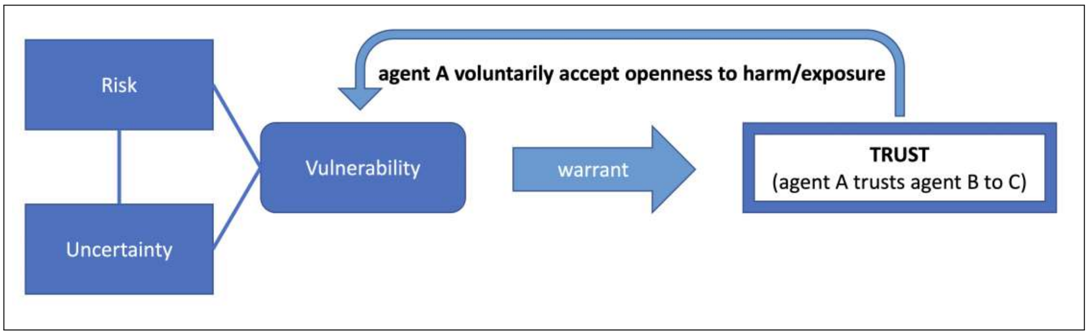

Risk
Uncertainty
Vulnerability
agent A voluntarily accept openness to harm/exposure
warrant
TRUST
(agent A trusts agent B to C)

Figure 3.1: My proposed guiding definition of trust represented visually.

I am using this simple conceptual model as a reference point for my two online studies, because it enables me to generate two different kinds of questions directly targeting the conceptual relation between trust and vulnerability for the specific purpose of studying trust in HRI, for which empirical work is needed. However, I will be using the guiding definition differently depending on whether the focus is placed on the human perspective or the robot during their interaction. Considering a human-centered perspective on trust in HRI (which I will be studying with two online interactive surveys in the chapters 4 and 5), we can infer from this simple conceptual relation that to voluntarily accept openness to harm, one must feel or be vulnerable in the first place. The question then arises, do people have to feel vulnerable in situations where they have to trust in robots or consider them trustworthy? When considering instead a more robot-centered perspective (which I will be studying with the use of expert interviews in chapter 6), the relevant

[11] Castelfranchi and Falcone (2010) even translated such conceptual knowledge into a computational model, which is a required step in the integration of knowledge when engineers build robots using a trust concept to guide their behavior. I believe this is also why their computational model is mainly aligned with the doxastic account of trust.

question that arises is: how we can consider their vulnerabilities, and how is this done empirically, since robots cannot feel vulnerable? However, before moving on to my presentation of the empirical work that I have undertaken in my PhD project to explore interpersonal trust in HRI through the vulnerability-trust conceptual relation, it is necessary to account for how risk, uncertainty and vulnerability as the preconditions of trust have been considered in current HRI studies to date.

CHAPTER 4

Human Experience of Vulnerability

> "Every day, we trust countless others without being able or required to perform any detailed reasoning about whether or not this is justified. Routinely, we are in a position of vulnerability towards others, expecting no harm from them or even presuming their benevolence and solidarity - and this often applies to others about whom we know very little, too."
>
> Möllering (2006, p. 51)

This chapter was developed and discussed in collaboration with Astrid Weiss (TU Wien) and Vicky Charisi (EU Commission) as part of the HUMAINT project. Most of the content has already been published as Hannibal et al. (2021).

With all the theoretical perspectives presented in the two previous chapters 2 and 3, I now shift the focus in this chapter to the first online HRI study I conducted to explore human experience of vulnerability for studying interpersonal trust in HRI. As such, this empirical work is intended to be a proof-of-concept study for showing that the *feeling of vulnerability is related to the perception of trust* in the specific context of HRI. Establishing the details on the direction and the strength of the conceptual relation between trust and vulnerability is work to be taken up in future research, which I also further elaborate in the conclusion chapter 8. Moreover, to make clear and acknowledge that the online HRI study is the result of a collaborative effort, I will in this chapter change the voice of the narrator from "I" to "we". Overall, this chapter comprises an account of how we developed and conducted our online HRI study, presents the results of our data analysis, and discusses the most interesting findings.

4.1 Study Aim

Early studies on trust in HRI used the psychological definition of trust offered by Deutsch (1960), which characterizes trust as the analysis of costs and benefits given a particular situation to ensure a favorable outcome. This definition was preferred because it highlights *risk* as the main concern and is well suited for studies focusing on the performance of robots for the purpose of cooperation. Today, most studies have shifted towards the more sociological definition by Mayer et al. (1995), which understands trust as the "willingness of a party to be vulnerable to the actions of another party, based on the expectation that the other will perform a particular action important to the truster, irrespective of the ability to monitor or control the other party" (p. 712). Its popularity in HRI stems from the inclusion of two very important elements of trust: vulnerability and benevolence. However, when vulnerability and benevolence as important elements are only brought into the discussion as part of the definition of trust, their crucial role in how they can guide HRI research is not fully exploited. That is, we only know that vulnerability and benevolence play an important role for trust, but not *which* one. In our work, we do not simply consider vulnerability as a precondition of trust and the fundamental element of benevolence as two different conceptual constructs among many others that influence trust during HRI, but rather as two interrelated *constitutive* elements for the way humans trust robots. Thus, we explore how a focus on human experience of vulnerability and human perception of robots as benevolent during HRI can offer a more theory-driven approach to studies on trust in HRI, and show with this proof-of-concept study how it is possible to study these two constitutive elements from a human-centered perspective.

From this explanation for why a research focus on the constitutive elements of vulnerability and benevolence can be considered useful for studying trust in HRI[1], the aim of our online HRI study was to explore which kinds of situations trigger an experience of vulnerability and sense of benevolence by the robot, and to what extent there would be any relation between such situations with the way people would rate trust during HRI.

4.1.1 Vulnerability in HRI

As I have mentioned previously in chapter 2 and that will also serve as the motivation of our online HRI study, the notion of vulnerability in the context of studying trust in HRI is considered in terms of self-disclosure by a robot (i.e., the work by Kaniarasu and Steinfeld, 2014; Martelaro et al., 2016; Sebo et al., 2018; Siino et al., 2008; Traeger et al., 2020). Using such an understanding of vulnerability is helpful for an HRI study that uses an experimental design methodology, as the vulnerability dimension becomes very concrete (i.e., specific linguistic statements) and can be manipulated to a different degree. However, we rather operationalize vulnerability more broadly in terms of *openness to harm* (whether physical or emotional) that leaves a sense of exposure (Cipolla, 2018). This is the most basic and general understanding and in our original and follow-up study we measured vulnerability for trust in HRI through written accounts of personal experience of uncomfortable feelings triggered by the specific trust violation instance caused by the PEPPER robot. These uncomfortable feelings are mainly recognized and categorized as negative emotions such as, e.g., disappointment, anger, sadness, fear, distress, helplessness, and frustration (Chen et al., 2011). With this operationalization, we are able to highlight and link vulnerability to the situatedness of trust in HRI as this feeling results from the specific reactions people had towards the robot during the interaction. That is, we were able to understand vulnerability not as a property of humans or robots, but rather as a feature of the relation between them and of the ongoing transformation of this relation given the specific situation they are in (Coeckelbergh, 2013).

However, since the experience of vulnerability also depends very much on the actions other people take in order to honor or betray the trust we place in them,

[1] In current research on trust in HRI, the experience of vulnerability by people and the considered benevolence of robots have so far been empirically studied as two separate elements. For practical reasons, previous work on such topics is being presented in a similar manner for later on to shown how we brought them together in our own proof-of-concept study.

the aspect of benevolence must also be includedin an elaborate and holistic study on trust in HRI.

4.1.2 Benevolence in HRI

Since benevolence can roughly be understood as the indention of an agent to do good out of the interest or concerns for the well-being of others (Mayer et al., 1995), it is very closely related to the vulnerability pre-condition of trust. When people trust in others as a way to protect their sense of vulnerability, it is at the same time a process where they evaluate if the other person is someone who is able to recognize and intentionally avoid exploiting the situation or cause any harm despite the opportunity to do so. In this sense, benevolence becomes a very fundamental indication of how trustworthy people judge others to be when focusing specifically on social interaction because it relates to the way they assume to be treated when accepting their openness to harm, i.e., vulnerability. Thus, we decided to include benevolence as the most relevant indication for how trustworthy people considered the robot (that they were interacting with) to be for studying interpersonal trust in HRI.

Benevolence has mainly been studied for trust in the contexts of automation (see e.g., Calhoun et al., 2019), human-robot teams (see e.g., De Visser et al., 2020; Wang et al., 2016), collaborative robots (see e.g., Nordqvist and Lindblom, 2018), and human-machine interaction (see e.g., Lyons and Havig, 2014). Very little work has so far been done in HRI specifically, with the exception of the work of Khalid et al. (2019)), who in their study consider benevolence as a construct taken from the definition of trust by Mayer et al. (1995), where it is characterized as follows: “a trustee is believed to want to do good to the trustor, aside from an egocentric profit motive.” (p. 718). Without further consideration of how to conceptualize benevolence, they used this definition to explore the perceived benevolence of humanoid robots within the different scenarios of business, disaster, and healthcare, using a cluster of associated terms (i.e., cheerful, friendly, kind, likable, and pleasant) as measurement. While these associated terms might be practical measurements for how trustworthy people perceive robots to be, this operationalization of benevolence is problematic because it leaves out the situational underpinning. That is, the perceived benevolence is not only a result of the actions that robots display, but also of how the humans interpret the specific interaction when it unfolds. Moreover, using the definition of benevolence by

Mayer et al. (1995) has also been criticized by Malle and Ullman (2021) for being too narrowly focused on a business context: they are skeptical about whether this definition of benevolence can be applied to HRI research more generally.

Given these issues, we believe that the conceptual analysis of benevolence by Livnat (2004) might be more helpful for studying trust in HRI, as it breaks down this aspect into its emotive, performative, and cognitive elements, unfolding the close connection between perceived benevolence and situatedness. The emotive element of benevolence concerns how people perceive robots in situations where care or concern for their human counterpart can be expected (suggesting that feelings matter). The expectation of care or concern relates to the perceived intent of a robot to do good rather than harm (suggesting that mind perception is needed) which goes beyond mere beneficence/beneficiary. However, this perceived intention to do good is not enough for the perception of a robot's benevolence. It must also be demonstrated in the behavior or action performed in the situation (suggesting actualization is important). Concerning the performative element, the way people perceive a robot's benevolence is linked to whether this attempt to do good during the interaction seems sincere. This is why an inactive robot will not be perceived as benevolent by people, even if it is still considered attentive. Further, a robot's benevolence as perceived by people is also based on the assumption that it has a minimum level of cognitive competence to identify and plan the necessary actions for doing good. As such, the cognitive element of a robot's perceived benevolence will ensure that people also consider it somewhat rational and transparent (suggesting the inclusion of clear reasoning and basic knowledge about human nature). Using this conceptual analysis of benevolence, we are now able to explore the perception of robots as benevolent from a more relational perspective. We consider benevolence as evolving from the actions taken in a particular situation, rather than as traits designed to make robots appear more benevolent. For empirical studies on trust in HRI, we therefore have to shift our attention to identifying those kinds of situational circumstances that make people perceive robots as benevolent, instead of focusing solely on specific cues or utterances by robots.

4.2 Methodology

So far, we have argued that vulnerability as a precondition of trust and benevolence as an fundamental element cannot be understood merely as two conceptual constructs in the definition of trust – they are instead two important constitutive elements for human trust in robots. We have also argued that, on a theoretical level, vulnerability and benevolence are closely linked to their role in interaction, which we express with the notion of situated trust. Consequently, to study human trust in robots with a focus on vulnerability and benevolence as constitutive element, it is not beneficial to rest on more classical study design approaches in which the establishment or creation of trust prior to the interaction in the given situation is required for the measurement. As an alternative, we believe that the work by Möllering (2006) highlights two important points that are also of particular relevance for research on human trust in HRI. First, he stresses the *taken-for-grantedness of trust in everyday life*. From a socio-phenomenological tradition, he views trust as always-already in an interaction or relation. This means that trust only comes to the foreground when it becomes problematic, is violated, under pressure, or needs to be questioned. Theoretically, a trust violation is required to happen before it is possible to measure trust. As such, the methodological requirement of having a baseline measurement of trust before the facilitation of the interaction between people and the robot, which is commonly seen in standard HRI studies on trust, would be problematic as the mentioning of trust itself could lead some participants to question it before the planned interaction have even started. To accommodate this methodological implication, we omitted any baseline measurements of trust for our own proof-of-concept-study, and would moreover suggest that in cases where a baseline measure is requires, to do it much in advance. Secondly, Möllering (2006) points out that trust in everyday life has less to do with rational reasoning than with *feeling and morals*. Measuring human trust in robots in this sense extends beyond what people are willing to risk in an uncertain situation . Thus, we wanted to take into account also the emotions and ethical concerns that people might be left with after an instance of a subtle trust violation in more everyday life situations.

Motivated by these insights, we set out to explore how a non-rationalist approach to trust in HRI enables a better understanding of the situatedness of human trust in robots, as related to vulnerability and benevolence as two impor-

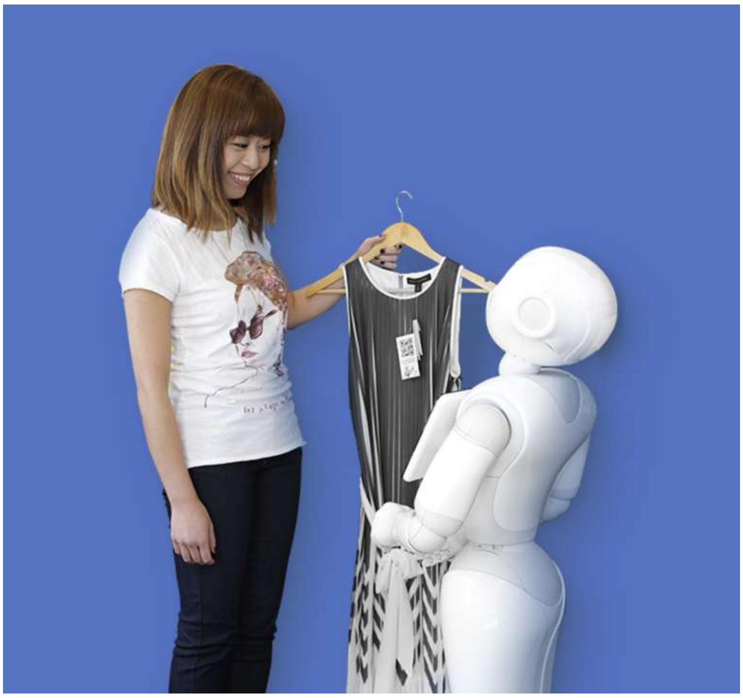


Figure 4.1: The PEPPER robot assisting a customer with clothes shopping © 2016 RobotLab (edit by G. Hannibal).

tant constitutive elements of trust. In order to do so, we decided to develop a proof-of-concept study around mundane everyday life scenarios involving subtle uncertainties, instead of relying on game-theoretical scenarios with obvious high-risk stakes. To investigate our theoretical approach, the following research questions guided our work: (RQ1) Is it possible to use subtle trust violations instances to study situated trust in HRI when focusing on vulnerability and benevolence? (RQ2) Which subtle trust violation instance in different scenarios of clothes shopping are more likely to trigger human experience of vulnerability and the perception of robots as benevolent ?

Specifically, we designed and conducted an online survey simulating three different clothes shopping scenarios, inspired by the RobotLAB demonstration video[2], which shows the implementation of a fashion recommendation engine in a humanoid robot for the potential application of robots in small retail businesses working as sales assistants.

[2] Demonstration video at the NRF Annual Convention & EXPO 2016 – Retail's BIG Show (accessed 19.01.2021): `https://www.youtube.com/watch?v=iJ184evAu-I`

4.3 Study Design

To address our research questions, our between-subjects interactive online survey contained three different scenarios for clothes shopping and tree respective trust violations, which also served as the experimental conditions: *economy*, *privacy*, and *transparency*. Use the links to the economy/privacy/transparency interactive online survey provided in appendix A.3 to experience what it was like to participate in the online HRI study (ca. 20-30 min for completion).

4.3.1 Procedure

Our interactive online survey was hosted on the Lime Survey platform and designed in modules (see Figure 4.2). Once participants clicked on the link in our recruitment information, they would be sent to a dispatch page, which directed them randomly to one of the interaction scenarios. People would first see a page with a consent form (see e.g., in appendix A, Figure A.1). If people did not voluntarily consent to participate, they were automatically excluded from our study and directed to an exit page. All participants who did give informed consent were directed to the introductory part of the study, where more details and practical information were provided. Our proof-of-concept study was ethically approved by the JRC ethical committee as part of the HUMAINT studies on human-robot interaction, and the page was designed according to the JRC data protection requirements.

In the first part of the study, participants were given two pre-engagement questionnaires about their attitude towards robots and shopping (i.e., NARS and PSA scale). When completed, they moved on to a small greeting session with the robot. In the engagement part of the online survey, we guided all participants through either the economy, privacy, or transparency scenario, where the participants would be assisted by a robot for clothes shopping. In the post-engagement part, they were given a questionnaire about trust (i.e., MDMT scale) to continue with a yes/no question and open-ended questions about their experience of having the robot assist them, focusing on aspects of vulnerability and benevolence. All participants were also given the option to give feedback on the study at the end.

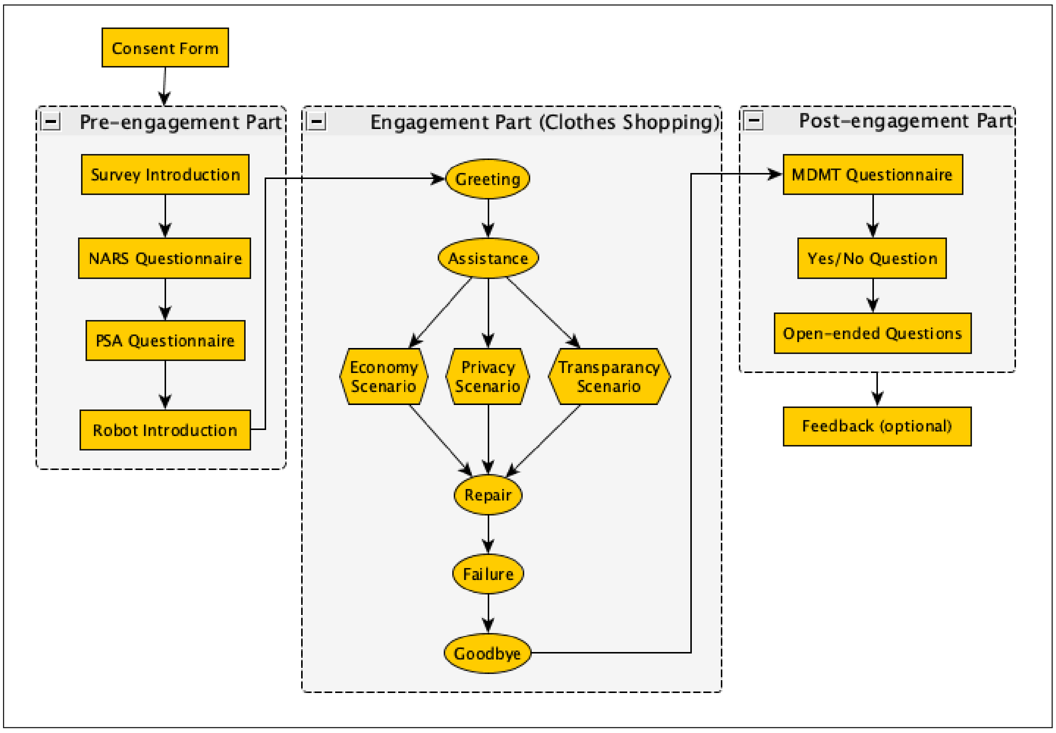


Figure 4.2: The different parts of the online survey.

4.3.2 Scenarios

The following steps were included in all scenarios (see Figure 4.2): i) the robot would greet the participant and offer assistance, ii) participants would add an item to the shopping list, iii) a subtle uncertainty moment with a trust violation would occur, iiii) a trust repair attempt and failure by the robot would follow. The instances of trust violation were designed as follows:

- *Economy Scenario* (ES): participants would be offered a special customer promotion and given the option to participate in a second, different promotion scheme if they did not like the initial offer. After choosing one of these, the robot would make a calculation mistake: the expected discount would not be included in the final price. The assumption behind this subtle trust violation design was that people who follow advice when spending money could potentially be vulnerable as this might challenge their financial security when this advice turns out not to be to their advantage (see e.g., McKnight et al. (2002)).

- *Privacy Scenario* (PS): Given the previous choice of clothes by the participants, the robot would suggest to change the gender on the customer account. Regardless of the answer, the robot would ask the participants if the information about their choice to change the setting or not can be used for future training material. We based this subtle trust violation design on the assumption that disclosure or sharing of personal information makes people potentially vulnerable to privacy issues, as this information can be misused (see e.g., Joinson et al. (2010)).
- *Transparency Scenario* (TS): With the motivation to provide a better recommendation for additional clothes items, the robot would ask the participants for irrelevant information. This subtle trust violation design rested on the assumption that people perceived robots as rational in their decision-making, so when robots make decisions that seem irrational or confusing, people might be potentially vulnerable, since people might start doubting their own sense-making (see e.g., Huang et al. (2019)).

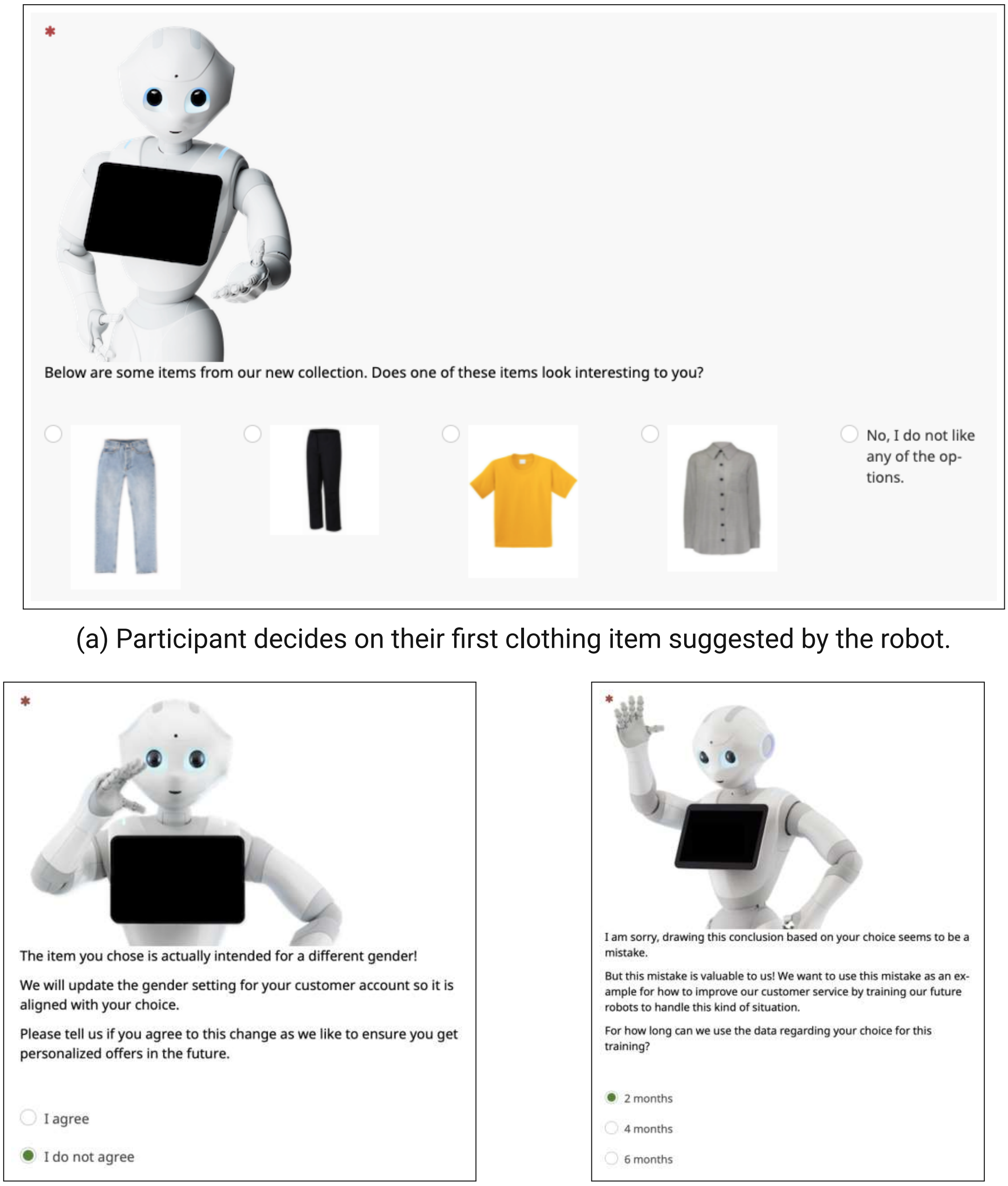
As mentioned above, these scenarios were developed following the assumption that trust is a default mindset (Möllering, 2006), which means that a trust violation is needed to study how people trust as they can no longer take this for granted.

4.3.3 Depicted Robot

For our online survey, we used photos of a commercially available humanoid robot developed by Softbank Robotics (see Figure 4.3). We chose photos that were expressive and showed the robot from the waist up, and photo edited the touch screen to black to ensure that people would focus on the body-language and social cues rather than the screen. These expressive photos of the robot were displayed on all the questions in the engagement part of the online survey. We believe that adding these expressive photos to the choice of action created a stronger sense of the robot being present despite the online format.

4.3.4 Measures

To answer our research questions, we made use of two established HRI scales. We used the Negative Attitude Towards Robots (NARS) scale by Nomura et al. (2006) to measure if the participants already had any particular negative attitudes



*

Below are some items from our new collection. Does one of these items look interesting to you?

○ ○ ○ ○ ○ No, I do not like any of the options.

(a) Participant decides on their first clothing item suggested by the robot.

*

The item you chose is actually intended for a different gender!

We will update the gender setting for your customer account so it is aligned with your choice.

Please tell us if you agree to this change as we like to ensure you get personalized offers in the future.

○ I agree

● I do not agree

*

I am sorry, drawing this conclusion based on your choice seems to be a mistake.

But this mistake is valuable to us! We want to use this mistake as an example for how to improve our customer service by training our future robots to handle this kind of situation.

For how long can we use the data regarding your choice for this training?

● 2 months

○ 4 months

○ 6 months

(b) A trust violation happens and participants decides how to respond.

(c) The reaction of the robot given the decision by the participant.

Figure 4.3: The robot helping with clothes shopping in the engagement part of the online survey.

towards robots before they were assisted by the robot. We also used the Multidimensional Measure of Trust (MDMT) by Ullman and Malle (2018) to measure the level of trust the participants would experience after their engagement with the robot. For the purpose of this online survey, we also developed a questionnaire on personal attitudes towards shopping (PSA), using questions from Babin et al. (1994) and Paridon et al. (2006), to better contextualize unexpected or unusual results. We also collected socio-demographic information about the participants regarding their age, gender, nationality, educational level, area of subject, and prior experience with robots. This information was used to check for any possible bias in the data. At the end of the online survey, a yes/no question on the experience of vulnerability (VUL) and benevolence (BEN) were added, followed by three open-ended questions related to these aspects. These questions were intended for the participants to further reflect on their experience of trust in having the robot assist them with clothes shopping.

4.3.5 Technical Pilot

Upon the completion of the implementation of the online interactive survey, we proceeded with a technical pilot which would allow for feedback and possible adjustments in our design. N=15 adults were recruited, among whom some were specialized on HRI or online experimental studies. The data collected in the pilot study showed an expected and ordinary variation in the responses of the participants and they also provided some suggestions for improvement. Following the implementation of the feedback from participants, we launched the main study on May 11th, 2020.

4.3.6 Participant Recruitment

All participants were recruited through online posts on social media (Twitter, Facebook, LinkedIn, Instagram), the Reddit internet forum, mailing lists (CHI-announcements, eusset, BCS-HCI), JRC newsletter, and personal networks or contacts. The data was collected during the months May to June, 2020.

4.4 Analysis of the Results

In total, we collected 98 valid survey results across the three different clothes shopping scenarios (economy = 36, privacy = 30, transparency = 32). We used the software IBM SPSS Statistics (version 27) to analyze all the quantitative data collected with the questionnaires and yes/no question, while the software MAXQDA (version 2020) was used to analyze the data collected through the open-ended questions. In this section, we will present the results from both our quantitative and qualitative measurements.

4.4.1 Questionnaires

The data set we used for our analysis consisted of 98 valid surveys. Participants were kept in the analysis if their questionnaires were completed[3] and no suspect data pattern was present[4]. The mean age of the participants was 36.46 years (SD:12,79, n = 98); the youngest participant was 18, the oldest 75 years old. A total of 42,9% of participants identified as female, 52,0% identified as male (4 people identified as non-binary and 1 stated other). Regarding the question on the pre-knowledge on robots, 33,7% of the participants stated they knew robots from culture (i.e., literature, movies, radio, magazines, and TV), 26,5% from education (i.e., course work, thesis projects, internships), 21,4% from work (i.e., building, programming, research projects), 14,3% spare time (i.e., DIY, science magazines, family, friends), and 4,1% accidental (i.e., store visit, study participant, events).

Data Reliability and Variable Computation

In order to measure *trust*, we used the Multidimensional Measure of Trust (MDMT scale), which consists of four different subscales: reliable, capable, sincere, and ethical. Subscale scores are average ratings of the four items constituting the particular dimension (e.g., Capable = average ratings of capable, skilled, competent, meticulous). All the “Does Not Fit” endorsements were treated as missing values. To compute a score for *Capacity Trust* and *Moral Trust* on the MDMT scale, we averaged the ratings on the eight items constituting the related

[3] For the three scenarios, we had to discard many surveys because they were only partially completed: economy = 86, privacy = 127, and transparency = 151.

[4] I.e., questionnaires were discarded if more than 10% of questions was missing, patterns like 1234512345 occurred, or the first answer option was always chosen.

subscales and we checked all subscales for internal reliability with Cronbach α. All values were higher than 0.70, which indicates good reliability (see e.g., also Figure 4.4).

The trust ratings were all relatively low, as can be seen in Table 4.4. As the MDMT scale is rated from 0 to 7, it indicates that the trust violations in all our shopping scenarios were effective, as all of them were rated below an average of 4.

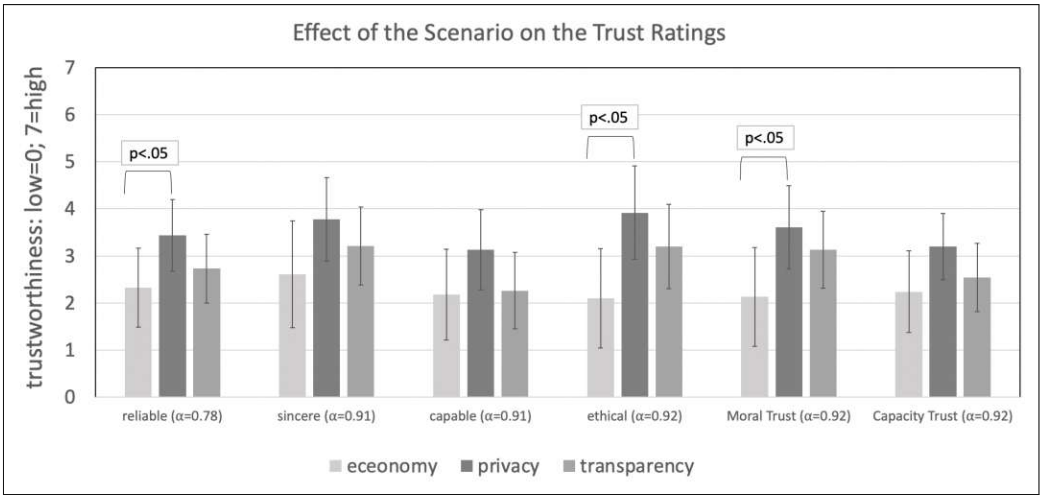


Figure 4.4: Internal reliability and descriptive statistics of the MDMT subscale ratings for each scenario.

In order to assess *people's attitude towards robots* in the beginning of the online survey, we used the Negative Attitude Towards Robots Scale (NARS). The 14 questions of this scale build three sub scales; (S1) negative attitude toward situations of interaction with robots; (S2) negative attitude toward social influence of robots; (S3) negative attitude toward emotions in interaction with robots. An internal variability check, using Cronbach α was done, revealing that two scales were slightly below 0.70, however, all items were kept for further analysis. The attitude ratings were all rather low as well, as Table 4.1 shows with the ratings for S1 being the lowest, indicating that our participants did not have a rather negative attitude towards interacting with robots. A one-way analyses of variance (ANOVA) also revealed that there were no significant differences in the ratings on any of the three subscales depending on the scenario. Therefore, a bias in the trust ratings on this respect can be excluded.

NARS scale	scenario	mean	SD
S1: interaction situations	economy	1,81	0,66
(α=0.76)	privacy	1,91	0,65
	transparency	1,98	0,64
S2: social influence	economy	2,84	0,70
(α=0.63)	privacy	2,72	0,75
	transparency	2,90	0,80
S3: emotions	economy	2,80	0,67
(α=0.65)	privacy	2,88	0,89
	transparency	2,83	0,83

Table 4.1: Internal reliability and descriptive statistics of the NARS subscales ratings for each scenario.

In order to assess people's *attitudes towards shopping*, we compiled a questionnaire consisting of 12 items which we summatively computed into scales on *shopping enjoyment*: 4 items on the degree of how much shopping is considered an enjoyable activity, *shopping advice*: 4 items on the degree to which a friend's advice is appreciated in shopping activities, and *shopping disappointment*: 3 items on how often one is disappointed with a shopping decision. An internal variability check using Cronbach α revealed that all scales were above 0.70.

Shopping scale	scenario	n	mean	SD
shopping enjoyment	economy	36	2.31	0.93
(α=0.87)	privacy	30	2.53	1.04
	transparency	32	2.59	1.11
shopping advice	economy	36	2.51	0.68
(α=0.77)	privacy	30	2.76	0.95
	transparency	32	2.78	0.94
shopping disappointment	economy	36	2.15	0.71
(α=0.77)	privacy	30	2.14	0.86
	transparency	32	2.15	0.80

Table 4.2: Internal reliability and descriptive statistics of the PSA scales ratings for each scenario.

The attitude ratings were all rather average, as Table 4.2 shows. Our participants were neither shopping enthusiasts nor too insecure about shopping decisions or taking advice from a friend. An ANOVA also revealed that there were no significant differences in the ratings on any of the three subscales depending on the scenario. Therefore, a bias in the trust ratings on this respect can be excluded.

Impact of Scenarios

First, we were interested in knowing if the three different scenarios had an effect on participants' trust ratings (MDMT scales). We already identifed a pattern in the descriptive statistics for all MDMT subscales with respect to the scenario (see e.g., Table 4.3): The privacy scenario always had the highest ratings, followed by transparency with the second highest, and the economy scenario always had the lowest ratings.. As such, the robot was rated relatively more trustworthy in the privacy scenario compared to the two others on transparency and economy. We conducted an ANOVA on the effects of the scenarios with respect to the different trust subscales of the MDMT. Three significant effects could be identified: There was a significant effect of the scenario on how *reliable* (F(2,77) = 3.382, p = 0.039) and *ethical* (F(2,58) = 4.853, p = 0.011) the robot was perceived to be, as well as on the overall *Moral Trust* (F(2,58) = 3.560, p = 0.035). A Gabriel's Pairwise posthoc test (chosen because the variances between groups were not equal) revealed that there was a significant difference in the ratings between the economy and the privacy scenario in terms of how reliable and ethical the robot was perceived to be, as well as in terms of overall Moral Trust. The economy scenario is rated significantly lower in all those three subscales.

Regarding our two yes/no question about whether participants experienced *vulnerability* or *perceived benevolence*, a significant difference in the answering behavior for the three different scenarios could be found only in the latter. As a Leven's Test revealed that the conditions for a parametric test were not met, a Kruskal-Wallis Test was calculated (H(3)=6.84, p=0.03), reflecting the pattern mentioned above: In the privacy scenario, participants most often replied with "yes, I experienced benevolence" (mean rank: 57,13), followed by the transparency (mean rank: 50,91), and the economy scenario (mean rank: 41,89). Pair-wise comparisons again revealed that the difference is only significant between the economy and the privacy scenario. This indicates that most participants in the economy scenario thought that the robot did not act in their interest (n=28).

MDMT scale	scenario	n	mean	SD
reliable	economy	28	**2.33**	1.68
(α=0.78)	privacy	25	**3.44**	1.52
	transparency	27	2.73	1.47
sincere	economy	27	2.61	2.26
(α=0.91)	privacy	26	3.78	1.77
	transparency	20	3.21	1.66
capable	economy	30	2.18	1.93
(α=0.91)	privacy	27	3.13	1.70
	transparency	26	2.26	1.62
ethical	economy	22	**2.10**	2.10
(α=0.92)	privacy	24	**3.92**	1.99
	transparency	18	3.20	1.80
Moral Trust	economy	21	**2.13**	2.11
(α=0.92)	privacy	22	**3.61**	1.77
	transparency	18	3.13	1.63
Capacity Trust	economy	24	2.24	1.73
(α=0.92)	privacy	24	3.20	1.41
	transparency	25	2.54	1.45

Table 4.3: Internal reliability and descriptive statistics of the MDMT subscales ratings for each scenario.

Impact of Vulnerability and Benevolence

Next, we were interested to see whether the experience of *vulnerability* or *benevolence* would also affect participants' trust ratings (independently of the respective scenario). The descriptive statistics on *vulnerability* already revealed the tendency of people who answered the question on vulnerability with "no" to rate the MDMT scales slightly higher, however, an ANOVA revealed that a statistically significant difference was detected only for the subscales *reliable* (F(1, 78) = 5.47, p = 0.02) and *ethical* (F(1, 62) = 4.11, p = 0.047). People who stated that they did not experience any vulnerability in their interaction with the robot rated it as significantly more reliable (mean: 3.06, SD: 1.54) and ethical (mean: 3.42, SD: 2.17). An overview on the yes/no distribution in relation to trust ratings for vulnerability can be found in Figure 4.5

In comparison, we saw that the assessment of benevolence affected participants' trust ratings for all MDMT subscales, as an ANOVA revealed: *reliable* (F(1, 78) = 17,06, p = 0.0) *sincere* (F(1, 71) = 5.96, p = 0.02) *capable* (F(1, 81) = 9.12, p =

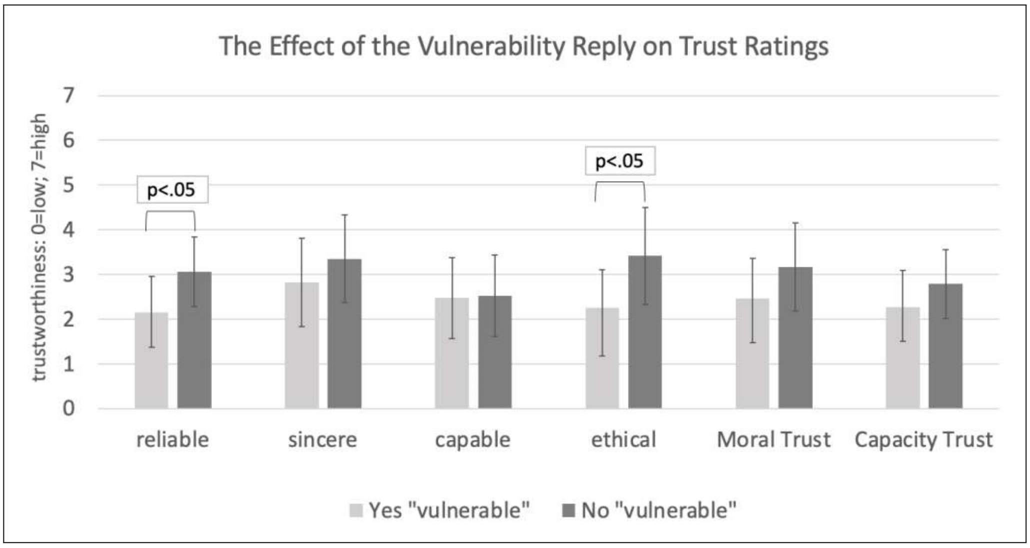


Figure 4.5: Different MDMT ratings for vulnerability regarding yes/no answers.

0.00) *ethical* (F(1, 62) = 16,90, p = 0.00) *Moral Trust* (F(1, 598) = 12.12, p = 0.00), *Capacity Trust* (F(1, 71) = 13.53, p = 0.00). Participants who stated that the robot did not act in their interest rated all subscales lower. An overview on the yes/no distribution in relation to trust ratings for benevolence can be found in Figure 4.6.

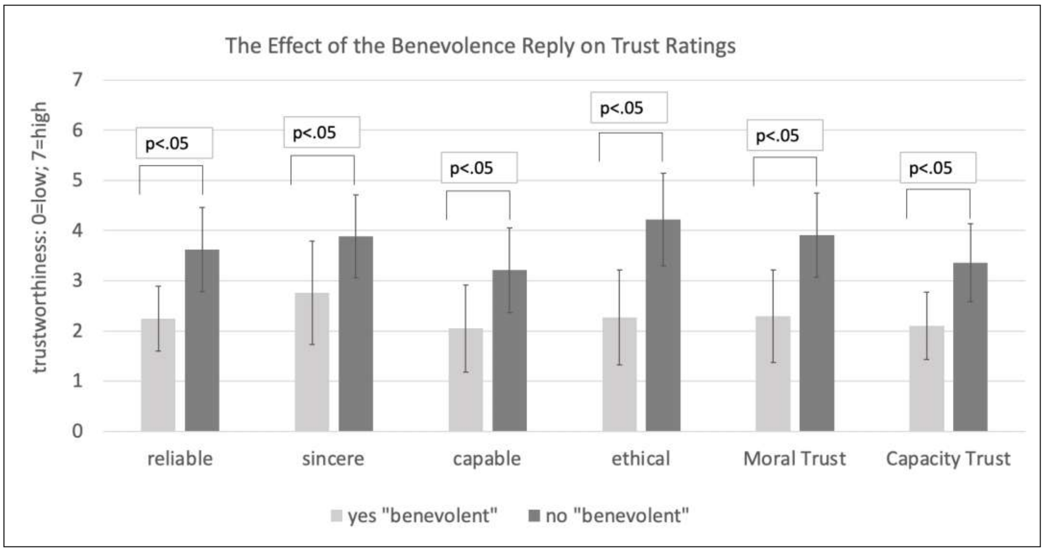


Figure 4.6: Different MDMT ratings for benevolence regarding yes/N no answers.

These findings support our assumption that *experience of vulnerability* and *perceived benevolence* relate to trust ratings to some degree. In conjunction with the finding that at least *benevolence* was significantly differently experienced for the three scenarios, a further impact on trust can be assumed and will need further research in the future. Additionally, a Pearson correlation revealed a statistically significant weak positive correlation between *vulnerability* and *benevolence* (r(96)=0.28, p=0.00).

Impact of Socio-demographics and Attitudes

We did not find any difference in the ratings for the three *NARS scales* and our three *shopping attitude scales* with respect to the different scenarios through an ANONA analysis, also no correlations between these attitudes and the trust ratings were identified. Similarly, an ANOVA did not reveal any significant differences with respect to the *MDMT scales* and *education*, *pre-knowledge on robots*, and *gender*. *Age*, however, correlated with three of the MDMT subscales. A Pearson correlation revealed a statistically significant moderate negative correlation between *age* and ratings on how *capable* (r(96) = -0.23, p = -0.00) and *ethical* (r(96) = -0.36, p = 0.00) the robot was perceived to be, as well as the *Moral Trust* rating (r(96) = 0.30, p = 0.00). The younger participants were, the more trustworthy they considered the robot to be in these respective scales. A statistically significant weak correlation was also found for the yes/no answering behavior of participants with respect to *vulnerability* and *age* (r(96) = 0.28, p = 0.00), indicating that younger participants were more likely to answer this question with “no”. No such correlation was observed for *benevolence*. However, for *benevolence* a weak gender correlation occurred (r(96) = -0.22, p=0.00), indicating that females were slightly more likely to answer the benevolence question with yes. This correlation was not be observed for *vulnerability*. As our analysis of sociodemographics and attitude data only revealed weak correlations of some variables with our dependent measures, no more regression models with multiple independent variables were calculated.

4.4.2 Open-Ended Questions

We examined the participants’ experience of vulnerability and perception of the robot’s benevolence with a set of eight open-ended questions. Including these questions was motivated by our wish to gain a deeper understanding of the nuances of peoples’ experiences that are not captured by statistical results. For the analysis of all the participants’ verbatim responses that we also used for exemplification, two members of the research team manually annotated their answers first deductively based on the pre-defined themes of the different scenarios and then inductively as unexpected codes would come up as we went through all the collected data (Miles et al., 2020). Based on this work, we identified the categories forming the themes of our annotation scheme (see e.g., Table 4.4). Then, we applied a content analysis to all the responses on the unit level of full sentences. All coding, categorization, and thematic analysis of the expert interviews were done electronically using the software MAXQDA (version 2020). Overall, the coding scheme corresponded to the focus of our three different scenarios: privacy, economy, and transparency. Two additional codes also became evident in our analysis, which related to the participants' experience of interaction and the subsequently induced feelings.

Based on the five codes of the annotation scheme, we identified n=367 quotations from the open-ended questions. Our thematic analysis revealed that participants most often (44.1%) stressed themes that were related to the interaction experience (C1). Comments about transparency issues (C4) followed with 25.0%, while feelings resulting from the interaction (C5) was covered with 14.2%. Privacy issues (C3) came after with 9.0%, and the least covered theme related to economy issues (C2) with 7.1%. A closer look at the content of C1 showed a strong interconnection with the other themes in terms of providing the base for reflection. That is, comments about interaction design decisions, the robot's appearance, and suggestions during clothes shopping were often used as an explanation or justification by participants to comment on one of the other themes. For example, one of the participants described the experience with the robot by stating that "*I feel I was tricked because I chose an item of clothing and then was told that it was men's clothing*" (P_117). Interestingly, the participant first indicated their affective state (i.e., feeling tricked) for then to justify the possible reasons with reference to the experienced interaction. Throughout all scenarios, we observed that participants had explicit comments about aspects of vulnerability and their perception of the robot's benevolence.

Experience of Vulnerability

Most of the participants connected their feeling of vulnerability with the robot's behavior and the restrictions of interaction opportunities with the robot (e.g., "*I wanted to buy something else but the robot didn't allow me*" (P_135)). Some

Coding Theme	**Definition and Example**
(C1) Interaction Experience	Comments about factual and behavioral interaction instances regarding the interface usability, the robot characteristics, or the scenario design (e.g., *"the current robots that exist in reality are dumber or slower than me"* P_105).
(C2) Economy Issues	Comment regarding the financial exchanges, numeric mistakes, and pricing of clothes items (e.g., *"I must trust it is the correct price"* P_25).
(C3) Privacy Issues	Comments regarding collection of personal data, data protection, and access to data without asking for consent (e.g., *"worried that PEPPER already had access to my data to e.g., make charges without real consent"* P_59).
(C4) Transparency Issues	Comments regarding the reasoning underlying the robot's behavior, the robot's perceived intentions, or surprising actions (e.g., *"Asking for personal data was odd - especially as there was no obvious need for the data"* P_42).
(C5) Induced Feelings	Comments regarding the feelings generated during the interaction or from reflecting on own reaction (e.g., *"I lacked agency. In a way, it made ME the machine in this interaction with these prompts"* P_76).

Table 4.4: The annotation scheme emerging from the content analysis of the open-ended questions.

participants also mentioned the negative emotional states they would experience during the interaction (e.g., "*adding preferred but not chosen items to the shopping card made me feel loss of control [...] and I was a bit irritated about that*" (P_74) or "*I lacked a lot of agency. The robot guided me too much and gave me only very limited choices for answering in a way, it made ME the machine in this interaction* *with these prompts*" (P_72). These comments suggest that the participants experienced a sense of vulnerability when the robot's particular behavior would trigger more negative emotional states resulting from feelings of loosing control or lacking agency.

Other participants connected their experienced vulnerability with concerns of privacy issues, as they saw a possibility that the robot might access their personal data (e.g., "*when Pepper was asking form my signs and such then say he will recommend me something I did feel like I'm being read by a robot*" (P_57)). One of the participants provided an interesting comment about the experience of vulnerability in relation to privacy issues as they regarded the experience of vulnerability as something negative if it resulted from interaction with other people. However, this was considered somewhat less of a problem when considering a robot in the place of a human because of its inability to understand their thoughts and feelings (e.g., "*Vulnerability is a very uncomfortable feeling for me, like feeling exposed or misunderstood, but it maybe not as bad with a robot as with a 'real' person, because the robot may not notice my discomfort as a friend would*" (P_76)). From this comment, it seems that the felt vulnerability in relation to the sharing of personal information would have been more problematic if it involved another person that should be able to consider the participant's discomfort with the situation.

We also found that some participants connected their feeling of vulnerability with their perception of the robot's benevolence (e.g., "*I probably would not interact with Pepper again. It would creep me out a lot. O might even feel like I am getting exploited*" (P_119), "*Making assumptions about clothes being meant for a specific gender, and thus unfit for me, made me feel irritated - and perhaps vulnerable too, how did the robot guess my gender (which is fluid)? Very conservative robot*" (P_148)). In both examples the participants indicated that the robot was not acting to their benefit. This suggests that for some participants, there is a relation between their feeling of vulnerability and the way they assess the intentions or transparency of the robot, which influences not only their trust, but also future use choice. Form the analysis of the open-ended questions we gained insights on the importance of feelings and morals for people to make their trust decisions on a robots in the everyday life scenario of clothes shopping..

Perceived Benevolence of the Robot

We found that the perceived benevolence of the robot was for most participants connected with issues of transparency. Some participants reflected explicitly on the robot as a technological system that needed to be transparent for them to understand its specific behaviors (e.g., "*The link to whatever reasoning was (supposed to be) going on was not transparent - AI, for me, needs to be explainable*" (P_42)), while others mentioned that this lack of transparency influenced their sense of trust violation (e.g., "*I am not averse to trusting AI or robots but I would like the process behind the questions to be as transparent as possible*" (P_08)).

Concerns about who would benefit from the robot's behavior were raised by some participants (e.g., "*It forced me to choose between two items that I did not like either of [...]. It's obviously acting for the store, not me*" (P_51), and "*I see the purpose (and commercial inventive) of these systems as more about collecting personal data rather than providing a personal service. I do not trust the motivations of the developers behind Pepper as a shopping assistant*" (P_57)). As such, it seems that trust violation with regards to the perceived benevolence of the robot is not only seen in light of trust that unfolds in more interpersonal interactions, but also extends into reflections about possible hidden agendas pushed by the developers of such technology.

Some participants did not necessarily consider it problematic that the robots' actions would be beneficial to others (e.g., "*Sometimes things don't have to be to my benefit, but then the question arises: who is benefiting? If it's the shop then I am mad. If the society benefits, things may be different...* [sic!]" (P_86)). This example suggests that the perceived benevolence might be highly dependent on knowing and believing that those who will in fact benefit also have an eye for the common good (e.g., "*I would redesign it so that it is for the benefit of society. So I would want it to point out how much of my shopping is green, how much is needed etc*" P_146)). Thus, based on our analysis, we can show that the perceived benevolence of the robot affects trust mostly with respect to the issues of transparency and explainability when considering these everyday life scenarios of clothes shopping.

4.5 Discussion

Empirical work on trust in HRI has mainly been guided by insights from the rational choice tradition and economic theories within the social sciences[5]. Trust from these perspectives is assumed to be selective (trust is limited to certain people), reasonable (trust is based on having good reasons), and decisive (trust directs thoughts and actions). Moreover, such perspectives also rest upon methodological individualism and self-interested utility maximization, where the choice of placing trust in another person is considered rational when the probability of a potential gain is higher than the potential loss. Consequently, the way trust has been commonly operationalized is by how well rational actors are able to economize on transaction costs given their ability to estimate the outcomes of alternative courses of action. Among the various theories developed to predict human behavior during interaction, it was game theory that aspired to developing the most powerful heuristics for research on trust. Originally concerned with situations posing a social dilemma of choice that contained a high risk, game theoretical experimental set-ups are still among the most popular approached to this date for studying the problem of trust (Möllering, 2006).

After presenting the results of the data analysis and the basic ideas underpinning the trust paradigm from which most research in HRI is committed, we will in this discussion explain how our results contribute to current research on trust in HRI. At the end, we will also mention some of the limitations of our work.

4.5.1 From Games to Ordinary Situations

With our online interactive survey, we aimed to explore how the experience of vulnerability and the perceived benevolence of the robot can increase our understanding of trust when studying it in a HRI context resembling the everyday life situation of clothes shopping. Based on the analysis of our results, we believe that an initial ground has been provided with our interactive online survey to *empirically study trust during HRI outside a game theoretical experimental set-up* when considering the constitutive elements of vulnerability and benevolence.

[5] We have decided on this short and joint account of these perspectives because it is not always easy to separate clearly the ideas from the rational choice tradition from those of economical theories in the literature on trust. For a more detailed account that explains well the various overlaps and differences between these perspectives, please see the work by Möllering (2006).

While it might seem uncontroversial to be of the opinion that trust can be studied without using a game scenario, because many ordinary everyday life situations require trust, this is not a trivial point. The majority of studies in HRI make use of a game scenario to study trust (see e.g., Aroyo et al. (2018); Correia et al. (2016); Mota et al. (2016)). We acknowledge that such an approach is useful because it provides a rather standardized setting. However, since such game scenarios always contain a competitive element, using a game-theoretical set-up can potentially distort how and when people trust. We might ask, what about more ordinary scenarios like that of shopping for clothes? We argue that the results from our online survey contribute to current discussions by showing that there is much to gain when considering more ordinary everyday life situations that do not build on the assumption of competition, even when this is done in collaboration with robots. Consequently, we suggest to include more ordinary scenarios for studying trust in HRI, such as e.g., cooking, commuting, and exercising. We suggest that there is a need for a more critical discussion about whether the use of game scenarios is the only useful approach for studying trust in HRI more generally, as it leaves out the everyday life uncertainties of ordinary situations. Moreover, going beyond game scenarios for studying trust in HRI is not only an advantage for inclusion of more ordinary situations of everyday life, but these ordinary situations are also very realistic near-future applications of robots.

4.5.2 From High-Risk to Mundane Situations

Another important point for discussion regarding our results relates to the common concern in the HRI community that a high-risk situation is required for studying trust in HRI properly. For example, Salem et al. (2015a) are concerned with the general problem of providing a high-risk situation that will be effective enough to measure trust given the strict regulations of research ethics committees. They explain that ethical requirements influence the design and validity of experimental studies on trust in HRI significantly, as people might not actually feel at risk. Adding to the work by others also considering more low-risk everyday life situations for studies on trust in HRI (see e.g., Cameron et al., 2015; Saunderson and Nejat, 2022), we show with our proof-of-concept study how an alternative study design can bypass this concern when considering trust in HRI more broadly. The risk that the participants in our online survey were exposed to cannot be considered high, because we intentionally decided to focus on subtle uncertainties in an everyday life scenario. As expected, the economy scenario was rated the most effective, potentially because the risk of losing money is easily recognizable. Even in an everyday scenario as mundane as clothes shopping, people still felt a sense of vulnerability and were concerned with issues of benevolence. This suggests that high-risk situations are not the only suitable situations for studying trust in HRI. In our view, using mundane everyday life situations is equally useful because a sense of uncertainty about the consequences of even low risks can make people question their trust in interactions with robots or the trustworthiness of robots. While we acknowledge that some high-risk situations with devastating consequences are important for studying trust in HRI, we want to bring into the current discussion considerations of how subtle everyday life uncertainties might provide new perspectives, as they can also instigate a high sense of exposure.

4.5.3 Concerning Familiar Situations

When discussing the results of our proof-of-concept study, we also became aware of how important it is to foreground situations of everyday life where the process of familiarization becomes central to trust in HRI. While some people do engage in both game-based and high-risk activities as part of their everyday life, these activities do not constitute the most part of their day. If we think about the ordinary and mundane situations people engage in regularly, they are much more characterized by a search for familiarity and stability. We encourage a sensitivity to situations that are recognizable to people because trust hinges on how familiar people are with their world, which includes not only other people but also artifacts, concepts, or emotions (Möllering, 2006). How familiar people feel with a given situation needs to also be taken into account when studying trust in HRI because it is closely related to their concrete and particular life-situation that is expressed in their experience of vulnerability and perception of the robot's benevolence. Even for game scenarios and high-risk situations, we need to critically consider how a lack of familiarity might influence the levels of trust people will place in robots (see e.g., Correia et al. (2016)).

4.5.4 Limitations and Concluding Remarks

There are several limitations to our proof-of-concept study. First, because of the COVID-19 outbreak, we had to redesign our originally planned study to take place online instead of in person. One of the advantages of this new online format was the opportunity of collecting a bigger sample size. A challenge was to ensure that we captured the human experience of vulnerability and perceived benevolence of the robot, since there was no physical interaction. While we acknowledge that our proof-of-concept study contained a limited opportunity for interaction, the results do suggest that a future laboratory study or field-trial may show even stronger effects that would support the importance of situatedness for trust in HRI. Secondly, our results from the open-ended questions indicate that some participants had difficulties linking the aspect of vulnerability to their experience. This might have affected the results of the yes/no question intended to measure this experience. Thus, for a follow-up study, we advise not to use the term vulnerability directly (i.e., "was there any time during the interaction where you felt vulnerable?"), but rather to operationalize this aspect with more indirectly associated notions (e.g., anger, discomfort, fear, sad, and nervous) that, when combined, can be used as a measurement. Thirdly, we recognized that the yes/no question for the aspect of benevolence was badly phrased, as it contained a negation that may have confused some people (i.e., "was there any time during the engagement where you believe the robot was not acting in your interest?"). However, the open-ended questions were helpful for understanding the relevance of the concept and people's answering behavior on the related yes/no question.

We aimed to show how vulnerability and benevolence as constitutive elements for human trust in robots can be used to study trust in HRI. With a critical view on previous work on vulnerability and benevolence in HRI, we presented a conceptual analysis in which they were both closely linked to the situatedness of HRI and to the *affective dimensions* that are important from a human-centered perspective. Additionally, we supported these theoretical insights by presenting the results from our proof-of-concept study, which addressed our RQ1 and supported that it is possible to empirically study human experience of vulnerability and the perception of a robot's benevolence in the *ordinary*, *mundane*, and *familiar* situation of clothes shopping. We found that the most useful trust violation scenario given the everyday life situation of clothes shopping was that of economy or privacy, which addresses our RQ2. We also discussed how focusing on subtle trust violation scenarios in an everyday life situation can inspire future empirical studies on situated trust in HRI by challenging current studies to go beyond the rationalist trust paradigm that favors game-based, high-risk and low-familiarity experimental set-ups.

CHAPTER 5

Trusting Robots in a Privacy Scenario

> "[...] robots are not capable of [operationally acting as trustees or trustors] but we humans can nevertheless act as if they were fellow social agents and interact with them in false, but possibly economically or emotionally useful, trusting relationships."
>
> Sullins (2020, p. 320)

This chapter has been developed and discussed in collaboration with Anna Dobrosovestnova (TU Wien) and Astrid Weiss (TU Wien), together with Vicky Charisi (EU Commission) as part of the HUMAINT project. Most of the content has already been published as Hannibal et al. (2022a).

Although the trust ratings of the privacy scenario were not the lowest ones among the three scenarios from the previous online HRI study (Hannibal et al., 2021), we decided that the insights gained were most interesting and worth more consideration. Especially in light of the increasing use of robots in domains of application closer to the everyday life of people, where privacy concerns will become more pertinent. As such, the intention with the follow-up study that I will be presenting in this chapter was to further explore human experience of vulnerability for trust in robots with a special focus on the theme of privacy. Since the follow-up study is very similar to the previous proof-of-concept online study regarding the set-up, I will report only the most relevant changes to the methodology and study design. The main focus in this chapter will be on providing an overview of privacy in HRI and its relation to trust, explain our study design, presenting the results of our follow-up study, and discussing how our results contribute to our understanding of trust in HRI.

5.1 Trust and Privacy Concerns in HRI

With their research agenda of "privacy-sensitive robotics", Rueben et al. (2018) wished to create more awareness about privacy-related issues for further studies on HRI. Since privacy is a basic human need and right, they argue, the different ways in which it can be violated is important to understand and take into account in the development of robots intended to be used in both public spaces and private settings. Rueben et al. (2018) also list the following seven themes that they believe will be central to discussions about privacy in robotics: "(1) data storage, processing, and filtering, (2) how robots can trick people into giving up personal information, (3) trust, (4) blame, (5) privacy regulations and other legal topics, (6) special private domains like the home, and (7) privacy scholarship outside of HRI" (Rueben et al., 2018, p. 77). Regarding the theme of trust, they suggest researchers working in the field of HRI explore in more detail how exactly privacy concerns and trust are related, point to the factors that influence the perceived trustworthiness of robots handling sensitive data, and examine how trust between people and robots is impacted by a privacy violation. Although according to Rueben et al. (2018) there is still much work to be done for understanding and studying the many overlapping and complex connections between privacy and trust, they argue that such efforts are worthwhile. In their view, robots developed to reduce privacy concerns are also more likely to be considered trustworthy.

A short review of the current HRI literature on the topic of privacy shows that much work has already been done with regards to addressing privacy concerns more generally. For example, Lee et al. (2011) conducted 10 semi-structured interviews to find an approach to designing privacy-sensitive features, Syrdal et al. (2007) set up an exploratory study with 12 participants about their concerns regarding disclosure of personal information in an HRI scenario, Krupp et al. (2017)

aimed to identify the various privacy concerns related to telepresence robots by carrying out three focus group interviews with a total of 13 participants, and recently, Rossi et al. (2021) studied how people felt about sharing highly sensitive information with either a human or robot bartender using an online survey where 76 participants filled in several questionnaires. There has also been some work aimed at clarifying the notion of privacy and how these different dimensions are present in the context of HRI. Rueben et al. (2018, p. 78) presented a taxonomy for privacy that features the four categories of "(1) Informational privacy over personal information; (2) Physical privacy over personal space or territory; (3) Psychological privacy over thoughts and values; and (4) Social privacy, over interactions with others and influence from them". Very similarly, (Lutz and Tamó-Larrieux, 2020, pp. 88-89) distinguish between "physical privacy" and "informational privacy", where the latter is further categorized into "institutional information privacy" or "social informational privacy". There seems to be a gap in the research agenda of privacy-sensitive robotics when it comes to studies that explicitly address to what extent privacy concerns and trust in HRI intersect. This leaves the impression that the notion of privacy as a supportive factor for trustworthy robots might in fact be more a hypothesis at this stage, rather than a claim grounded in empirical evidence.

5.2 Study Aim

While our follow-up study did not aim at closing this gap in the research on privacy and trust in HRI, we did strive to provide insights into these issues by exploring empirically how the experience of vulnerability and the perception of robots as benevolent relates to the way people rate their trust in them, within a *privacy scenario*. With the proof-of-concept study, we set out to explore how the relation between vulnerability and benevolence relates to trust within the realistic everyday life situation of clothes shopping[1]. In this follow-up study, we focused especially on the privacy scenario, where participants interact with a robot that provokes a subtle trust violation instance consisting of a mild privacy breach.

[1] See e.g., chapter 4 for the account of the way vulnerability and benevolence have been conceptualized within current HRI research on trust.

5.3 Methodology

As in the previous proof-of-concept study (see e.g., chapter 4), in this follow-up study, we relied on the basic methodological assumption that trust is a default mindset (Möllering, 2006), which means that a trust violation is needed to empirically study whether people trust in robots or find them trustworthy.

Additionally, the methodology was guided by the theoretical insight by Coeckelbergh (2013) that there is an important distinction to be drawn between the human experience of vulnerability *in the moment of interaction or engagement* and those vulnerabilities that are *constituted by imagination*. That is, human beings are able not only to be aware of the possibilities of risk and uncertainty in the exact situation in which they find themselves, but are also able to reflect on the possible risk and uncertainty that will take place in the future. Currently in the literature on trust in HRI, this distinction can be seen in different studies that mainly focus either on the specific factors that influence human trust in robots in a specific interaction (see e.g., Haring et al., 2013; Sanders et al., 2014; Ullman and Malle, 2017), or on broader study of human trust in these robots by considering more contextual aspects (see e.g., Cameron et al., 2015; Haring et al., 2014; Sebo et al., 2018). By comparing participant experiences of vulnerability when interacting or engaging with a robot with their experiences reflecting on possible vulnerabilities after such interactions, we assumed for our follow-up study that the types of vulnerabilities they would mention would differ according to whether they were asked about it right after the interaction or later on after having had time to reflect on this experience more generally.. On a practical level, we explored this discrepancy using the method of semi-structured interviews with a small sample of the participants who would also participate in the interactive online survey as described in the following section (and as already sketched out in the previous chapter 4).

5.4 Study Design

Our follow-up online HRI study reused the study design as presented in the previous chapter 4:[2] with the consent of all participants, they were asked to fill in

[2] Use the links to the privacy iterated survey provided in appendix B.3 (ca. 20-30 min for completion) to experience how it was to participate in the follow-up online HRI study.

questionnaires about their attitude towards robots and shopping (i.e., NARS and PSA scales) before interacting with the robot. A small greeting session with the robot was then provided to introduce the robot to the participants. Then we guided the participants through the privacy scenario where they would be assisted by a PEPPER robot for clothes shopping. Afterwards, we asked the participants to fill in a questionnaire about trust (i.e., MDMT scale) for then to continue with answering a yes/no question and open-ended questions about their experience of having the robot assist them, focusing on aspects of vulnerability and benevolence. An option to provide feedback on the study and their participation at the end was also provided to the participants.

To analyze the quantitative data collected through our questionnaires and yes/no question, we again used the software IBM SPSS Statistics (version 27), as well as the software MAXQDA (version 2020) to analyze all the qualitative data that we collected with the open-ended questions and semi-structured interviews. In this section, we will present the results from both our quantitative and qualitative measurements. As part of the HUMAINT project, we were again able to use the consent forms already approved by the JRC ethical committee. As such, the data collected was approved by the Data Protection Officer of the European Commission, and conformed with the GDPR policy (see e.g., Figure B.1 in appendix B).

However, we made some relevant changes to the study design in this follow-up version that we will give a short account of (see e.g., Figure 5.1 for a graphical representation of all parts of the study design).

5.4.1 Privacy Scenario Focus

After discussing the results from the previous proof-of-concept study on human experience of vulnerability during HRI, we saw that the privacy scenario was not rated significantly less trustworthy than the two other scenarios of economy or transparency. Nevertheless, we were keen to explore this specific privacy scenario in more detail because we became more interested in how the more subtle trust violation instance related specifically to privacy concerns, as we suspected it might be of greater importance to the vision of having robots as an integrated part of human everyday life.

For instance, in 2017, the Austrian supermarket chain MERKUR (now BILLA+) placed a human-like robot in selected stores for marketing purposes. The robot

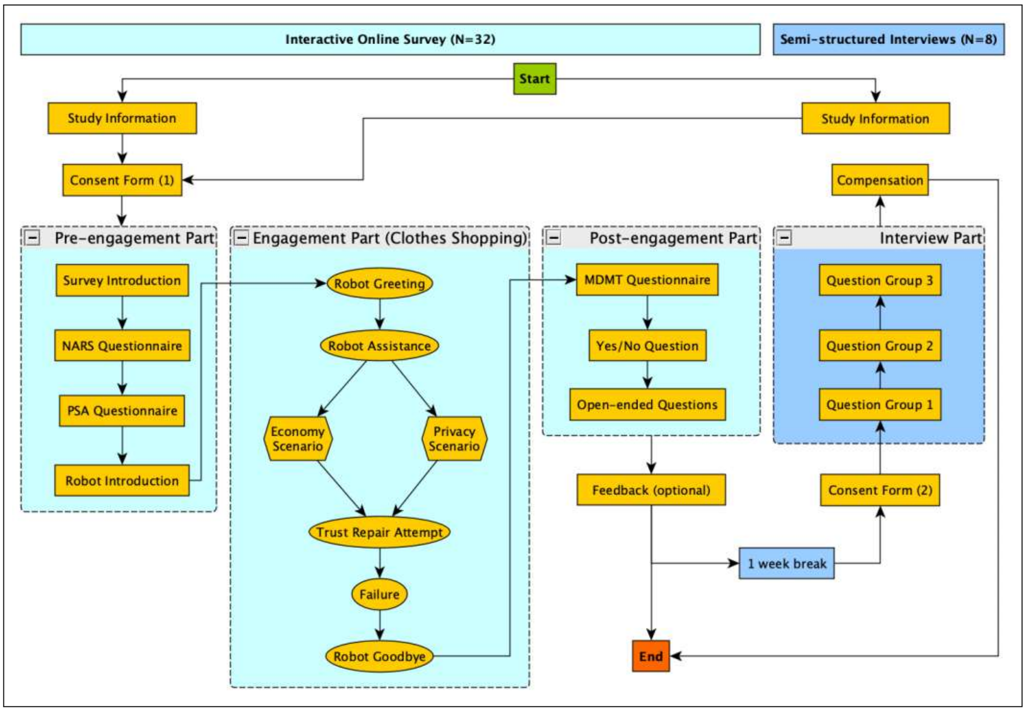


Figure 5.1: A visualization of the study design with both the interactive survey and interview part.

was programmed to provide customers with information about discounts, recipes, and news (see e.g., Al-Youssel, 2017; Neubauer, 2017). The English bank HSBC also used this robot to help bringing in foot traffic and new business by improving their customer service. The robot greeted the customers, informed them about the availability of self-service options, and helped them to determine their needs by asking questions (see ee.g., Campbell, 2018; Shaw, 2018). In such use cases, where a robot helps people with their shopping, there are multiple situations where personal and sensitive information about people needs to be collected, stored, and used. While the information gathered is intended to help customize and personalize the interaction, this might in return leave people vulnerable to exploitation in cases where this information is being misused.

5.4.2 Video Stimuli

For our iterated interactive online survey, we again used the commercially available PEPPER robot, developed by Softbank Robotics. For the video version and with the extensive assistance by Matthias Hirschmanner from the Trust Robots Doctoral College at TU Wien, we programmed into the PEPPER robot both the economy and privacy scenario using the visual interface program NAOqi 2.5.10 with Choregraphe Suite (Softbank Robotics). We used a mix of pre-programmed gestures for the sequence of behaviors, and added our own text for the dialogue.

We wanted to produce a video version of the privacy scenario that would seem very authentic and close to a real shopping experience. Our goal was to make the video-based interaction as close as possible to the still-picture version used in the previous survey, to ensure suitable comparison. For this reason, we placed a green screen behind the robot, where we would later insert a picture that showed the inside of a clothes store. The videos were produced and recorded over the course of four days in a small studio belonging to the HCI group at TU Wien, which had all the facilities required to create high quality video output (see e.g., Figure 5.2).

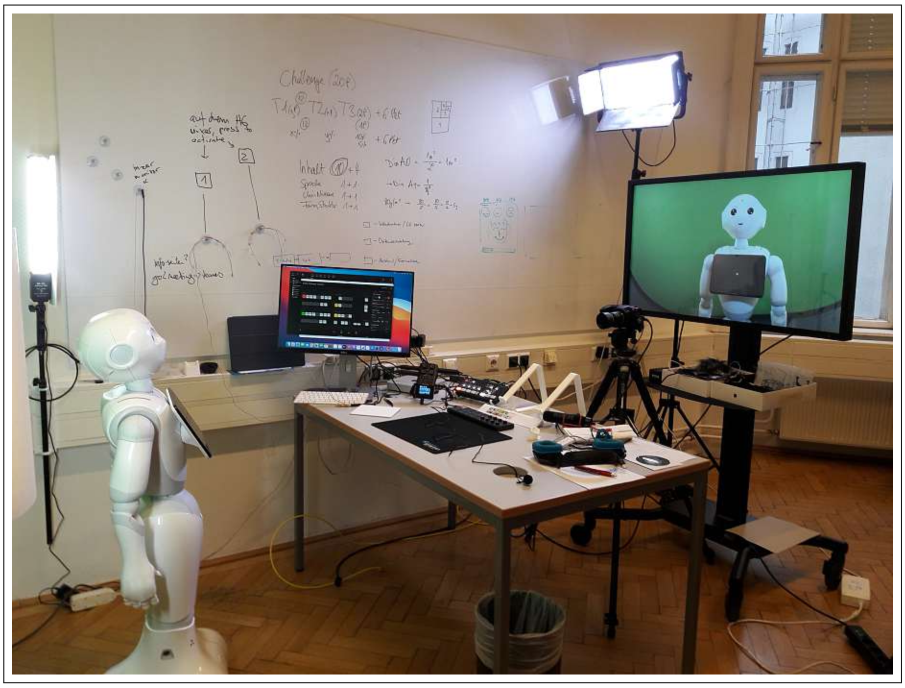
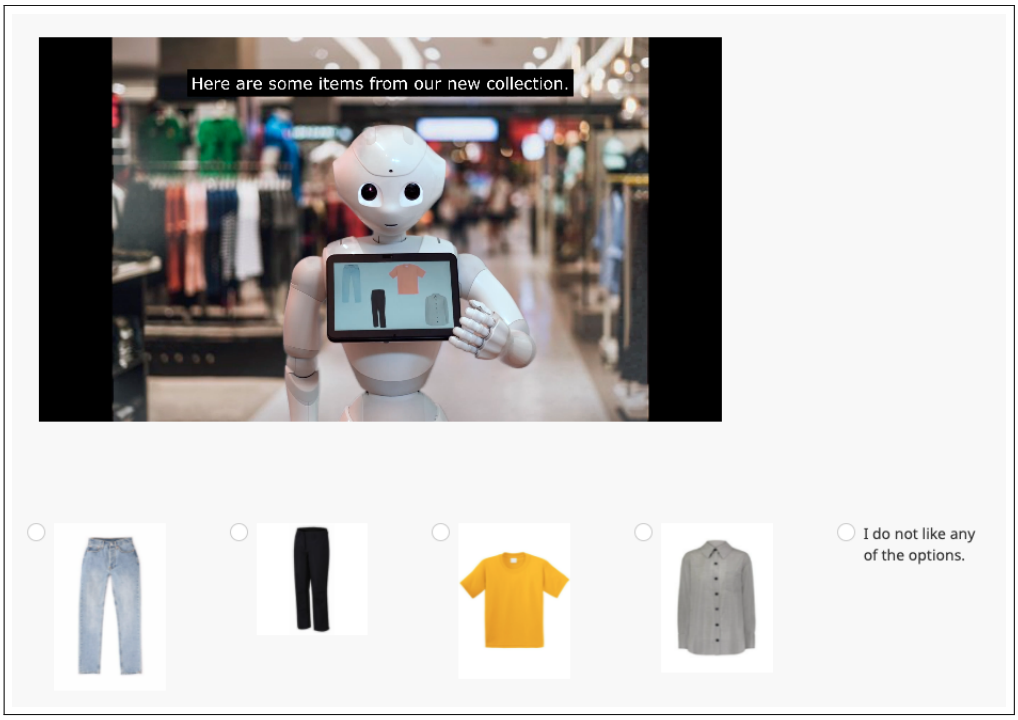


Figure 5.2: The studio set-up for the recording of videos that we used for the follow-up study.

After recording the videos, we edited and streamlined them so they could be well integrated into the interactive online survey. These videos of the robot were displayed above all the questions in the engagement part of the online survey, and were played automatically, with the option to re-watch each video if needed. We assumed that the video stimuli presented along with the choice of action provided a stronger sense of the robot being present, compared to the previous version that featured only still pictures of the PEPPER robot (see e.g., Figure 5.3).



Here are some items from our new collection.

○ ○ ○ ○ ○ I do not like any of the options.

Figure 5.3: Screen shot of the interactive part of the survey. Participants are asked to chose one of the clothes items.

5.4.3 Technical Pilot

Upon completing the interactive online survey with the the video version included as the new stimuli, we set up a small technical pilot which would allow us to gain some feedback and insights for possible adjustments of our design. Two adults were involved in this technical pilot, who both had some experience with HRI studies and could also provide input on the flow of online experimental studies. In the feedback thus received, our attention was drawn to certain areas of improvement for the video stimuli, as well as question formulation and selection choice in demographic information. Following the implementation of the feedback, we launched the main study on March 3rd, 2021.

5.4.4 Recruitment

For the recruitment of study participants, we advertised the interactive survey on social media platforms (i.e., Twitter, LinkedIn, Facebook), various HCI/HRI/philosophy mailing lists, and through professional as well as personal contacts. For a small fee, we also gained access to the recruitment platform "Call For Participants" that put our interactive online survey in front of potential participants who would be visiting the website (the guest could choose whether to participate in the the interactive online study, provided that they meet our recruitment requirements)[3]. We also advertised the interactive survey on the Slack working-space platform used by HRI/HCI researchers for the recruitment of participants. Although recruitment for the initial online study was rather successful, it was much more challenging to gain enough participants for this follow-up study. The data was collected from the beginning of March to the end of July, 2021.

5.4.5 Semi-structured Interviews

Alongside the interactive online survey, we also wanted to conduct semi-structured interviews with n = 11 participants for the follow-up study. The idea was to conduct the interviews approximately one week after the participants had filled in the interactive online survey part, as the time passed would leave some room for them to reflect on their experience. To find participants for this part of the study, we began by adding some information about the option to also participate in the follow-up interview at the end of the interactive online survey. However, since this strategy was not very effective, we decided to approach people (with the help of our friends and colleagues) more directly about their interest in participating, and informed them from the start that participation consisted of both the survey and interview.

Before conducting the interview, all participants were asked to sign a separate consent form made specifically for the interview-related data collect. This form contained further details about the aim of the interviews and what participation consisted of, as well as information about the data protection policy and their rights as participants. After receiving the signed and dated consent form from

[3] For further information about the "Call For Participants" recruitment platform, please visit their website: `https://www.callforparticipants.com`

each participant, we used email to coordinate a specific date and time for the interview, which was conducted online using the ZOOM platform.

The interview procedure consisted of three overall parts that we divided into pre-interview (i.e., short introduction of the interview aim and recap of participants' rights, as well as the request to record the interview), interview (i.e., the list of prepared questions), and post-interview (i.e., providing practical information and closing remarks). For the semi-structured interview part, we asked the participants 12 different questions in total. These questions were grouped according to the different stages, starting from the concrete experience they had had interacting with the robot for the interactive online survey, then moving on to more general reflections about their view on their own experience with robots:

- Question group A: their participation in the online study.

1. What do you remember about participating in the online survey?
2. What would you say/do you think that the survey was about?
3. How did you feel afterwards?
4. Why, or what triggered this experience/feeling?
5. Did you talk with anyone about your experience/reflection?
6. What did you tell people about the study?

- Question group B: their view on trust in robots after participating.

7. Imagine you meet the PEPPER robot again helping out with clothes shopping in the near future, how would you feel about that?
8. Do you think that the scenarios from the online study are very likely?
9. What do you think that everyday life with robots will look like in the future?

- Question group C: their broader reflections on trust in robots.

10. Have you had any reflections on or discussions about trust in robots before participating in this study?
11. Before taking part in this online study, did you ever think that your experience [their words] could arise from interacting with a robot?

12. Do you think that other people can relate or had similar experience [their words] as you when interacting with the PEPPER robot?

This guideline we prepared for the interviews (see e.g., Figure B.3 in appendix B) allowed us to direct the conversation towards the aspects of trust in HRI that we wanted to explore with our follow-up study, while also leaving enough flexibility to pursue aspects highlighted by participants that would be interesting to consider, or support our understanding of their reflections. The semi-structured interviews lasted between 25 and 40 minutes, depending on how detailed the participants replied to the questions, and whether we asked follow-up questions to gain more insight into certain aspects they mentioned. All participants received a 15 EUR voucher as compensation for their time.

5.5 Analysis of Results

We collected 32 valid survey results in total for the iterated privacy survey[4] , and conducted 11 follow-up interviews with participants around one week after they had done the survey part.

5.5.1 Questionnaires

The mean age of participants was 34.72 years (SD:11.01, n=32); the youngest participant was 22, the oldest 64 years old. A total of 43.8% of participants identified as female, 53.1% identified as male (1 person identified as non-binary). Regarding the question on prior knowledge on robots, 34.4% of participants stated that they knew robots from culture (i.e., literature, movies, radio, magazines, and TV), 18.8% education (i.e., coursework, thesis projects, internships), 21.9% work (i.e., building, programming, research projects), 6.3% spare time (i.e., DIY, science magazines, family, friends), 3.1% accidental (i.e., store visit, study participant, events). For our follow-up study, we added the answer category “I have no or very limited knowledge about robots” and with this option now available, 15.6% stated that they had no or very limited knowledge about robots. Adding this new answer option ensures a more representative description of the participants’

[4] We only included those questionnaires that were completed and had no suspect data pattern (i.e., were more than 10% of questions was missing, patterns like 1234512345 occurred, or the first answer option was always chosen). Accordingly, we had to discard 130 uncompleted surveys.

prior knowledge for our follow-up study because it is a mandatory question but in our previous proof-of-concept study it was assumed that participants had such prior knowledge. Overall, the samples of the first and the second study can be considered rather similar in their composition.

Data Reliability and Variable Computation

In order to measure trust, we used again the Multidimensional Measure of Trust (MDMT) scale by Ullman and Malle (2018). The MDMT consists of four different subscales: reliable (α=0.74), capable (α=0.93), sincere (α=0.95), ethical (α=0.94), which were calculated through average ratings of the four items constituting the particular dimension; Reliable and capable build together the *Capacity Trust* scale (α=0.90), while sincere and ethical build the *Moral Trust* scale (α=0.96). As in the previous study, all the “Does Not Fit” endorsements were treated as missing values and Cronbach α values of 0.70 and higher indicated a good internal reliability for all scales. As the MDMT scale is rated from 0 to 7, ratings of all MDMT scales below 4 indicated that the trust violations in our iterated privacy scenario were similarly effective as in the previous study .

MDMT scale	**scenario**	**n**	**mean**	**SD**
reliable	iterated privacy	24	3.52	1.72
	original privacy	25	3.44	1.52
sincere	iterated privacy	17	3.00	2.21
	original privacy	26	3.78	1.77
capable	iterated privacy	26	2.70	1.90
	original privacy	27	3.13	1.70
ethical	iterated privacy	14	3.79	2.19
	original privacy	24	3.92	1.99
Moral Trust	iterated privacy	12	3.35	2.06
	original privacy	22	3.61	1.77
Capacity Trust	iterated privacy	22	3.23	1.68
	original privacy	24	3.20	1.41

Table 5.1: Descriptive statistics of the MDMT subscales ratings for the iterated and the original privacy scenario.

Impact of Vulnerability

We wanted to see if the experience of vulnerability (measured through a yes/no question) affected participants' trust ratings in the iterated privacy survey. There was a significant effect for participants who stated that they did not feel vulnerable, as they rated all the MDMT subscales higher than people who answered with "yes", except sincere (F(1, 15) = 1.67, p = 0.22). This was visible from an ANOVA: reliable (F(1, 22) = 9.84, p = 0.01), capable (F(1, 24) = 6.06, p = 0.02), ethical (F(1, 12) = 6.22, p = 0.03), Moral Trust (F(1, 10) = 7.32, p = 0.02), Capacity Trust (F(1, 20) = 8.39, p = 0.01). The descriptive statistics on vulnerability show that people who answered the question with "no" rated the MDMT scales sometimes even above the scale average of 4 (see Table 5.2).

MDTM scale	vulnerability	n	mean	SD
reliable	Yes	12	2.58	1.57
	No	12	4.46	1.35
sincere	Yes	9	2.36	2.31
	No	8	3.72	1.98
capable	Yes	13	1.87	1.69
	No	13	3.54	1.78
ethical	Yes	8	2.72	2.34
	No	6	5.21	0.73
Moral Trust	Yes	7	2.27	2.03
	No	5	4.88	0.76
Capacity Trust	Yes	11	2.34	1.50
	No	11	4.13	1.49

Table 5.2: Descriptive statistics of the MDMT subscales ratings for the respective yes/no answers on vulnerability.

Impact of Benevolence

Next, we wanted to see if the experience of benevolence impacted participants' trust ratings in the iterated privacy survey. However, no statistically significant difference was found (reliable: F(1, 22) = 1.31, p = 0.27; sincere: F(1, 15) = 0.24, p = 0.63; capable: F(1, 24) = 1.02, p = 0.32; ethical: F(1, 12) = 2.31, p = 0.15; Moral Trust: F(1, 10) = 0.64, p = 0.44; Capacity Trust: F(1, 20) = 2.36, p = 0.14). However, descriptive statistics show that in most cases, people answered the question on benevolence with "yes" (see e.g., Table 5.3).

MDTM scale	benevolence	n	mean	SD
reliable	Yes	17	3.26	1.83
	No	7	4.14	1.36
sincere	Yes	14	2.88	2.35
	No	3	3.58	1.59
capable	Yes	19	2.47	2.04
	No	7	3.32	1.40
ethical	Yes	10	3.25	2.29
	No	4	5.12	1.30
Moral Trust	Yes	10	3.14	2.20
	No	2	4.44	0.80
Capacity Trust	Yes	16	2.91	1.81
	No	6	4.10	0.87

Table 5.3: Descriptive statistics of the MDMT subscales ratings for the respective yes/no answers on benevolence.

Link between vulnerability and benevolence

A 1:1 comparison of the results from the original and iterated privacy survey is not possible as we changed the phrasing of the vulnerability yes/no question and the benevolence yes/no question. In the original privacy survey for the vulnerability yes/no question, most participants answered with “no” (n= 30; yes=8, no=22), a significantly different distribution compared to an expected random distribution ($\chi^2 = 6.53$, p=0.01). This difference vanished with the improved phrasing of the vulnerability yes/no question that we used in the iterated privacy survey (n= 32; yes=17, no=15)[5]. Regarding benevolence, we achieved a clear answer tendency (n=32; yes=23, no=9) – a significantly different distribution compared to an expected random contribution ($\chi^2 = 6.13$, p=0.01)[6]. However, we could not find a significant correlation between the answers on vulnerability and benevolence

[5] We realized that participants had a hard time imagining what could be meant by the term “vulnerability” when analyzing the data from the original privacy survey. Thus, we changed the wording from “was there any time during the interaction with the PEPPER robot where you felt vulnerable?” to “Was there any time during the interaction with the PEPPER robot where you felt vulnerable (i.e., feeling uncomfortable or experienced any negative emotions)?”.

[6] From the data analysis of the original privacy survey, we realized that the benevolence yes/no question contained a double negation that made it difficult to determine if the participants answered the question with yes or no as they intended (n=30; yes=14, no=16). We got rid of this ambiguity by rephrasing the benevolence yes/no question to “Was there any time during the interaction where you felt that the PEPPER robot acted against your benefit?”

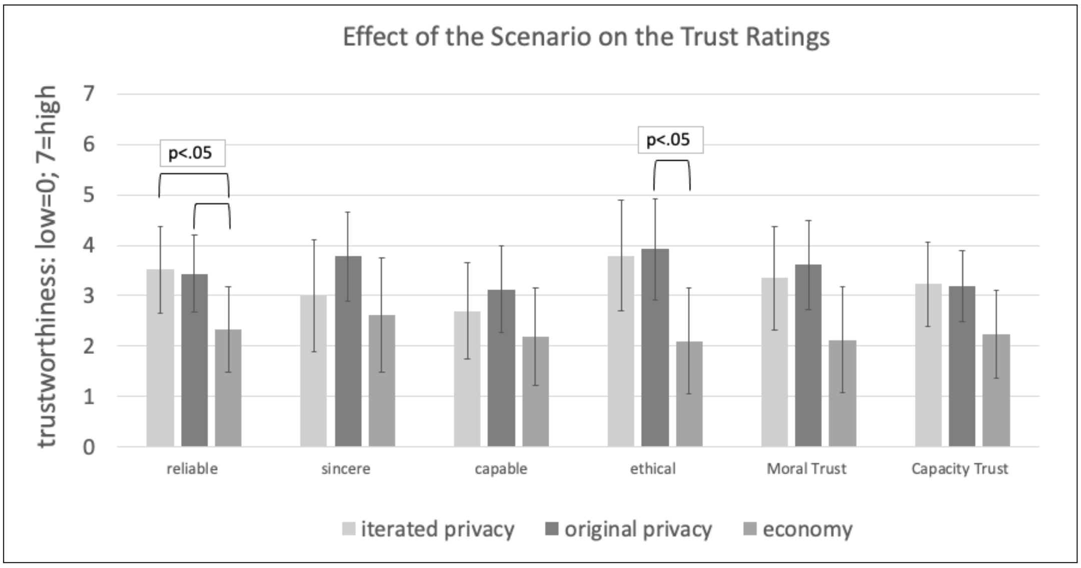


Figure 5.4: Effect of the scenario on the MDMT subscale ratings.

either for the original, or for the iterated privacy scenario (original: Cramer's V = 0.34, p=0.06; iterated: Cramer's V = 0.12, p=0.54).

Impact of Scenario

Finally, we investigated if the trust violation in the iterated privacy survey actually created significantly lower trust ratings than the scenarios of the previous study. Since a Leven's Test revealed that the conditions for a parametric test were not met, Kruskal-Wallis Tests were carried out, which were significant for the following two MDMT subscales: reliable: H(3)=10.82, p=0.01; ethical: H (3)=9.96, p=0.02. Pair-wise comparisons (Bonferroni corrected) revealed that, like in the previous study, the difference is only significant between the economy and the original privacy scenario for the the MDMT reliable subscale scale (U=21.96, p=0.05), as well as for the iterated privacy scenario (U=22.77, p=0.04). For the MDMT ethical subscale, the difference can only be found for the economy and the original privacy scenario (U=19.51, p=0.02; iterated privacy: U=18.45, p=0.10) (see e.g., Figure 5.4).

From our analysis, it seems that whether we used the picture or video stimuli for the interactive online survey did not lead to significant differences in the MDMT ratings. Moreover, the different scenarios (i.e., economy, original and iterated privacy version, transparency) when considered in isolation affected the rating significantly only for the reliable and ethical MDMT subscales. The yes/no question on vulnerability, however, influenced all subscales of the MDMT significantly, except sincere. Given this observation, we decided to look further at how the MDMT ratings were influenced by the different scenarios when considering only those cases where the participants said they did not feel vulnerable. A Kruskal-Wallis Test , revealed that they rated the scenarios significantly different with respect to the MDMT reliable subscale (H(3)=13.72, p=0.00), the MDMT capable subscale (H(3)=9.80, p=0.02), the MDMT ethical subscale (H(3)=9.41, p=0.02), and the MDMT *Capacity Trust* scale (H(3)=12.09, p=0.02;).

To summarize, if we looked only at the participants who stated that they did not feel vulnerable, we found significant differences in more of the MDMT subscales ratings compared to the analysis including all the participants. In our view, this result indicates that there is a relation between the experience of vulnerability and the situated nature of trust.

Socio-demographics and other independent variables

In order to assess people's attitude towards robots at the beginning of the interactive online survey, we used again the Negative Attitude Towards Robots Scale (NARS)[7]. The internal reliability check with Cronbach α revealed that all scales were slightly below 0.70; however, all items were kept for further analysis. The NARS ratings were also all rather low, with the ratings for S1 being the lowest (see Table 5.4). This indicated that our participants did not have a negative attitude towards interacting with robots, like in the previous proof-of-concept study.

In the previous proof-of-concept study, we created from our PSA questionnaire 3 scales (shopping enjoyment, shopping advice, and shopping disappointment) using factor analysis[8]. However, a factor analysis combining the data from proof-of-concept and follow-up study revealed that 5 components represent an

[7] The 14 questions of this scale build three subscales: (S1) Negative attitude toward situations of interaction with robots (α=0.56); (S2) Negative attitude toward social influence of robots (α=0.59); (S3) Negative attitude toward emotions in interaction with robots (α=0.57); rated from 1=totally disagree to 5=totally agree.

[8] Compiling the 16 items of our PSA questionnaire covered originally the following 5 components: PSA 1-4: hedonic motive (do people go shopping for the pleasure?) PSA: 5-8 utilitarian motive (do people consider shopping an effort?), PSA9-11: personal outcomes confidence (do people have confidence in their shopping skills?), PSA 12-13: social outcome confidence (are people admired by others for their shopping skills?), PSA 14-16: information sharing (do people share their experiences of shopping with others?).

NARS scale	scenario	mean	SD
S1: interaction situations	iterated privacy	2.06	0.55
	original privacy	1.91	0.65
S2: social influence	iterated privacy	2.77	0.61
	original privacy	2.72	0.75
S3: emotions	iterated privacy	3.00	0.70
	original privacy	2.88	0.89

Table 5.4: Ratings to the NARS subscales for the iterated and original privacy scenario.

eigenvalue above 1 and explain 65,46% of the total variance. A closer look at the component matrix revealed that several items loaded on more than one construct, which mean that only two components reliably extract: enjoying shopping (PSA 1-4 and PSA 15-16), with an internal reliability of 0.86, and outcome satisfaction (PSA 7, 9, 10, 11) with an internal reliability scale of 0.74 (see e.g., Table 5.5).

PSA scale	1	2	3	4	5
PSA1	.806	.130	-.291	-.052	.011
PSA2	.755	-.010	-.293	-.112	-.081
PSA3	.778	.071	.001	.070	.064
PSA4	.674	.052	-.341	.132	-.086
PSA5	-.266	.268	.160	.544	.592
PSA6	.188	.036	-.098	.669	-.386
PSA7	.183	.643	.209	.207	.285
PSA8_r	.266	.153	-.518	-.207	.524
PSA9_r	-.108	.746	-.163	-.178	-.226
PSA10_r	-.115	.780	-.143	.010	-.116
PSA11_r	-.100	.789	.118	-.129	-.198
PSA12	.645	.004	.523	-.282	-.049
PSA13	.548	.186	.611	-.183	.068
PSA14	.407	-.008	.126	.495	-.173
PSA15	.752	-.062	.036	.022	.134
PSA16	.663	-.132	.109	.077	.008

Table 5.5: Factor analysis of the PSA scale items.

The descriptive statistics for the two scales (see Table 5.6) show that our participants were not shopping enthusiasts, but quite satisfied with their shopping outcome. Furthermore, a Kruskal-Wallis H-test revealed that the shopping attitude was not rated differently for the scenarios (enjoying shopping: H(3)=1.80, p=0.62; outcome satisfaction: H(3)=0.18, p=0.98). Therefore, we conclude that it did not affect the scenario assessments.

Shopping scale	**scenario**	**n**	**mean**	**SD**
enjoying shopping	original privacy	30	2.62	0.99
	iterated privacy	32	2.38	0.90
outcome satisfaction	original privacy	30	3.85	0.71
	iterated privacy	32	3.86	0.57

Table 5.6: Descriptive statistics for the shopping scales for original and iterated privacy scenario.

No significant correlations could be found for age and any of the scales or the vulnerability yes/no or benevolence yes/no question. Similarly, a correlation analysis of the shopping scales with the MDMT scales only revealed one weak significant correlation between the enjoying shopping scale and the ethical trust scale (Pearson's r=0.27, p<0.05), meaning that the more people enjoy shopping, the higher they rated the ethical dimension in trust.

5.5.2 Open-ended Questions

We examined participants' perceptions of vulnerability and benevolence with a set of eight open-ended questions. The main purpose of the qualitative analysis of these responses was to acquire a deeper understanding of the nuances that appeared in participants' reflections, which can potentially support our interpretation of the findings from the quantitative data analysis.

Based on the answers of participants, we applied the method of content analysis to all the responses from the open-ended questions that were coded first deductively and then inductively (Miles et al., 2020). We considered a whole sentence with complete meaning, part of a sentence, or single words as a valid unit for the coding. Focusing only on the privacy scenario, the annotation scheme consisted of three codes already identified in the previous proof-of-concept study (see e.g., Table 4.4) because we found references to the (C1) interaction experience, (C5) robot features and behavior, and the (C4) feelings occurring. Two additional and new codes emerged from the analysis: one code that referred to usage or meaning of the trust notion (C2), the other referred to the reasoning for accepting a breach in privacy (C3). Figure 5.7 provides an overview of the five different codes (C1-C5) derived from the qualitative data analysis, with a short explanation and example for each of them.

Coding Theme	Definition	Example
(C1) Interaction Experience	References to factual and behavioral interaction instances regarding the survey interface usability or the experimental scenario design.	*"For me too much interaction/context between the interactions was missing and the different interactions following up on each were not in a good flow in terms of UX."* (P_17)
(C2) Trust Understanding	Reference to the usage or meaning of the trust notion for interactions with robots or in regular interactions between humans.	*"I think there is always a risk of a technology but also a human being acting against my benefit. So in my opinion this is a risk I am taking in every interaction. Therefore, I would also be willing to take it with PEPPER."* (P_19)
(C3) Violation Tolerance	Reference to the reasoning for accepting a breach in privacy, or the willingness to engage in the interaction again.	*"If the interaction gives me information which is not attainable in any other way, I would be willing to interact despite feeling vulnerable."* (P_10)
(C4) Induced Feelings	Reference to the various kinds of feelings generated during the interaction, or from reflecting on own reactions.	*"I wanted to buy something that was for my partner of a different gender and it could not cope with that scenario. Made me feel frustrated and a bit angry about it wasting my time."* (P_13)
(C5) Robot Related	Reference to the perception of the robot in terms of both features/ behaviors and how it influenced the interaction or experience.	*"I felt powerless to change the course of the interaction. The robot was in full control of what was talked about and when it was talked about."* (P_6)

Table 5.7: Annotation scheme emerging from the content analysis of the open-ended questions.

Based on the five codes of the annotation scheme, we identified $n = 347$ quotations from the open-ended questions in which the participants of this study had the opportunity to reflect freely on their experience. From the analysis, we found that overall, participants were most attentive to the aspects that were related to their (C4) induced feeling (26.8%), which was followed by responses referring to the (C5) robot-related features or behavior (22.7%), with a close run up by references to trust (C3) and violation tolerance (22.1%) followed by some comments about (C2) trust understanding (15.2%) before finally mentioning elements related to the (C1) interaction experience (12.9%).

The aim of the qualitative data analysis presented in the following sections was to generate a comprehensive understanding of the dynamics among the specific codes of the annotation scheme, as reflected in the narrative accounts of the participants. We divided this analysis into comments related explicitly to experience of vulnerability, and comments related to perception of the robot's benevolence.

Vulnerability Experience

The kinds of feelings people connected with their experience of vulnerability were expressed mostly in negative terms (e.g., feeling "uncomfortable", "impatient", "stupid", "annoyed", "powerless", and "not understood"). Some participants said their experience was not really a feeling they associated with vulnerability, but still it was somehow unsettling (e.g., "*I did not have a feeling of vulnerability but I felt uncomfortably because I did not feel any connection to the robot. I did not feel that the robot did see me as the person I am.*" (P_21)). The experience of vulnerability that is recognizable in these comments suggests that the trust violation in the privacy scenario stirred up emotional reactions of an unpleasant kind, and that the participants considered it problematic for their interaction with the robot.

Other reflections by participants on vulnerability had more to do with the way the robot was acting towards them (e.g., "*[...] pepper trying to update my choice of gender, which I had deliberately expressed before .*" (P_01a)), or appeared during the interaction (e.g., "*The robot snaps to its 'neutral' position, which is quite eerie and creepy.*" (P_11b)). The fact that the robot would also question their gender identity was considered very unpleasant by some participants ("*It made me uncomfortable and sad that I said I'm non-binary and then Pepper told me the clothing item I chose was not aligned with my gender.*" (P_16)). We gain from these comments an understanding of why people might have felt vulnerable: the robot seemed to excise some level of agency or having intentions.

Many participants also related their sense of vulnerability to the feeling of restricted action space, either because of the restrictive options provided by the robot (e.g., "*These limitation of my choice was artificially created by the bot after I had told it that I did not want any of what it had randomly offered me.*" (P_07a)) or because of a feeling of pressure by the robot to buy any of the items ("*When I HAD TO choose one of the clothes while I didn't want one [...]*" (P_18)). Comments like these suggest that the way participants felt vulnerable was also influenced by their concerns about being free to decide for themselves what actions to take during the interaction.

Perceived Benevolence

When it came to the perceived benevolence of the robot, many of the participants mentioned that they did not see how the outcome of the interaction was of any benefit to them. Either because they considered the robot to be forcing them to chose a clothing item they were not interested in buying (e.g., "*[...]I feel it is acting against my benefit, or trying to oblige me to buy an item I don't want to buy*" (P_07b)) or because the forced decision was seen as a wish by the retail store to profit (e.g., "*So pepper acted in the interest of the shop, not mine, offering me an array of pants, or specially priced ones to choose*" (P_06a)). These responses reveal not only that people perceive the robot as being inconsiderate, but also extend this view onto the company using the robot to sell clothe.

The perception participants had of the robot as acting against their benefit was also connected to issues of missing or inadequate information. For example, one participant said that the robot only provided the required information when it was too late in the shopping process (e.g., "*I was told the price of the item only after selecting it*" (P_01b)), while another mentioned that it would have been helpful to get more detailed information from the robot about the clothing items (e.g., "*Also, after picking an item, it was immediately added to my cart - I would have liked to see it in more detail before that.*" (P_04b)). These responses suggest that people consider the appropriate delivery and richness of information provided by the robot to be important for its perceived benevolence.

There were also some participants that did not consider the robot benevolent because it was explicitly addressing their gender as a problem in its ability to assist them with clothes shopping. To some, the robot caused concerns from the mere fact that it would propose to adjust the settings simply based on the clothes items they picked (e.g., “*By offering to change my gender settings based on a clothing item I chose*” (P_16)). Others were critical of why the robot would even make decisions about the shopping process based on stereotypical gender norms, and did not view it as helpful at all (e.g., “*The moment when it had a categorical distinction of ’gendered’ clothing. As a customer, I want to make the choices on what I want to buy by myself, no matter what actual gender I identify myself with. I don’t want anyone to decide that aspect for me and don’t see the point why there has a distinction in clothing has to be made.*” (P_09b)). Such comments from participants show that the information about gender or its role in guiding the clothes shopping decision-making is a sensitive topic that makes some people doubtful about robots.

Trade-off Between Trust and Privacy

Our data analysis also shed light on the willingness people showed to accept a potential trust violation instance to happen again in the same situation, which we studied through a privacy breach where the robot request to re-use information about their gender. In both cases where the participants responded with “yes” to either feeling vulnerable or perceiving the robot to act against their interest, several of them would still consider interacting with the robot at least once more, .

Taken together, 7 out of the 17 participants who said they felt vulnerable and 7 out of the 23 participants who said that the robot acting against their benefit would interact with the robot again in the same situation of clothes shopping, despite the chance of having their trust violated . The reasons provided by the participants to accept the privacy breach varied and were often reflecting some form of cost-benefit analysis. One participant saw potential in the application of robots as shopping assistants (e.g., “*Probably, yes. It’s an interesting concept and I’d love to see future executions of it*” (P_09) and two participants focused mainly on the potential advantage they could gain from interacting with the robot because it could be useful also in different situations (e.g., “*Sure, in other situations it might be useful.*” (P_02)) or help serve their agenda (e.g., “*Yes, if it helps me achieve my own goals*” (P_06b)). Related specifically to the level of performance, some participant would take into account how often problems with the robot would occur (e.g., *"yes, unless this seems to happen again and again"* (P_01)) and based on the expectation that the robot will prove to be better and safer in the long term (e.g., *"probably only if I had an expectation that the bot was going to improve its performance over time. and that rather quickly and without harming me."* (P_07)). Another participant mentioned that they consider interacting with the robot again simply for exploratory purposes (e.g., *"Not in a 'serious' way but maybe to test it and play around with it - but I'd do the shopping by myself."* (P_18)) or out of mere curiosity (e.g., *"If I know there would be such a risk (even if it is not a major risk, I have to say), I would probably use it only out of curiosity, for the mere interest of interacting with a robot and see if it really understands what I am saying, and which items it would suggest me."* (P_07b)). To one participant, the decision was to be made depending on the potential worth of the interaction goal or outcome in general (e.g., *"It depends on the value I see in this interaction."* (P_10b)).

5.5.3 Semi-structured Interviews

We transcribed the 11 semi-structured interviews verbatim, using the automated transcription software OTTER (version 2.0)[9] and afterwards we checked and manually edited the texts for missing information or misrepresentation as needed. To prevent potential bias for subjectivity, the analysis of the interviews was carried out by two members of the research team independently and in two rounds of iteration.

In the first round, our focus was on the rich details of the data and how these contribute to a more nuanced understanding of the experience of interacting with the robot in the privacy scenario. Therefore, we used a thematic analysis approach Clarke et al. (2015) as a guiding framework at this stage. We used inductive coding to derive open codes from a close reading of the transcripts. We were specifically interested in different perspectives that participants had on their interaction with the robot, how they interpreted their experience, and any other noteworthy discussion points.

Then the two members of the research team in discussion compared the codes they derived with the outcomes of the analysis of the open-ended questions to check if they align or pointed in different directions. Having ensured that there

[9] Find details about the transcription software on their website: `https://otter.ai`

was a significant overlap, the two members of the research team proceeded to the second round of iteration to analyze the interviews with deductive coding focusing on the topics of trust, vulnerability, privacy, attitude to shopping and participants' background experience with robotics. The deductive codes were decided upon with a view towards the aspects of trust in HRI that we wanted to explore with these follow-interviews, as well as the inclusion of reoccurring perspectives that were raised by the participants. Finally, we based the presentation of our data analysis on a structure that aims to highlight how different participants reflected on their experience and related topics. In what follows, the first section discusses people's perspectives on trust, vulnerability and privacy in the context of shopping scenario. The second and third sections then dive into participants' motivations for interacting with the robot in a similar context in a real-life situation.

Trust related

Participants highlighted different aspects related to their understanding and experience of trust in the context of shopping. Several participants we interviewed mentioned that their understanding of trust in the robot was a matter of its reliable performance. In this sense, the robot would be considered trustworthy if it carried out its job as expected and efficiently (e.g, "*I would think of it as a machine and here for me the criteria is not so much trust, but maybe reliability, like - does it work?*"(P_02)). In this regard, the evaluation of whether the robot could be trusted to perform its work reliably was also dependent on the option for repeated interactions (e.g., "*Reliability was a big, is a big factor, that you can't really test in a one-shot interaction and it's something that you don't want to even go further testing if it doesn't help you achieving the goal that you had in mind*" (P_04)).

From the point of view of trust as mere reliance, the robot was interpreted as a tool or a machine rather than an autonomous agent to whom responsibility and blame could be attributed. In this case, people thought that the notion of trust with regards to responsibility is only applicable towards the engineers and researchers developing the robot or the company deploying it (e.g., "*I just think whatever in a robot has got to be, it's got to be put there by a person, or a number of people.*" (P_07)).

Whether participants felt they could trust a robot also depended on the type of task the robot was to perform. As one participant mentioned, in the shopping context, they would be interested in what suggestions the robot puts forth, but they would not count on it for the final purchasing act. (e.g., "*I don't think I really count on it to make a purchase act, or I don't really, I would not really change my decision because of its suggestion*" (P_09)).

The experience of trust was further considered in light of the perceived damage or harm that can follow if the robot was to make a mistake. Several participants commented that trusting a robot in the context of clothes shopping may not be an issue as the stakes were not high or because they considered clothes shopping as an unimportant activity in their life (e.g., "*And I don't have really big trust issues in robots, especially when it comes to shopping. If this would be some medical situation, maybe it would be different, but shopping where not so much could go wrong I think it's a fine option to include robots into everyday life.*" (P_08)).

Vulnerability related

The experience of vulnerability was for some participants related to how they saw shopping for clothes as something personal and intimate. One participant explained that sometimes the things that we choose to wear have certain personal meaning and also express who we are. If the option to choose clothes is mediated by a robot that oversteps this intimate sphere with suggestions that are based on gender stereotypes, this can be felt as very uncomfortable and would even be offensive if it happened in a real-life situation (e.g., "*And then the robot came with the comments: oh, this is actually not your gender. And, for example, there I felt in a way, I think that was uncomfortable for me because the robot tried to come into this very intimate, personal sphere where I thought if this would happen in real world, then I would be offended by it*" (P_03)).

Another participant also mentioned the irritation that aroused from the question of changing the gender setting by the robot because it was very presumptuous and would be inappropriate if it came from a human sales assistant (e.g., "*It really did sort of irritate me when the robot came up with this question should it update my gender in my personal account. [...] I had this impulse that it's a very dumb question to ask. But then also there's an irritation there because it's just the wrong thing to say. In a setting with a person, what would a person do?*" (P_11)).

Another aspect related to the vulnerability experience was the issue of technological solutions imposing or constraining human choices. For example, one participant brought into discussion the problem of personal autonomy for choosing clothes items without the robot intervening with suggestions of what fits better based on pre-determined gender views (e.g, "*The robot tried to dominate my personal sphere. And somehow, I felt that as I said uncomfortable because there are certain boundaries where I'm willing to interact with the robot [...] And I don't want to be guided or dominated by technology*" (P_03)).

Privacy related

Some participants addressed the questions of privacy without being explicitly asked about it. Participants identified that privacy issues were related to aspects of gender identity as sensitive information that should not be shared with others. One participant for example mentioned that whenever surfing the internet, they feel hesitant about what personal data to share or not (e.g, "*I think that's always this trust issue when it comes to sharing your personal data and especially when it comes to sharing what gender I myself identify. So I think there's always some struggle or just criticizing thought which I have when being confronted with technology, whatever it might be*" (P_03)). The same participant also pointed out that the issue is even more pertinent when it comes to the use of robots in such clothes shopping context.

In general, participants relied on their prior experience with regular internet use when considering issues with data collection and storage. One participant for example focused on how asking for data storage by the robot sounded very familiar to experiences of regular online shopping (e.g., "*[...] it asked me should we save your preferences or should we save you in our database. And in this moment it made me feel like wait a minute, this sounds familiar. And, no, that's not okay. And then this triggered other emotional responses*" (P_06)).

The problem of collecting huge amounts of information about people for making personalized recommendation was also mentioned by one participant (e.g., "*I think maybe for a certain kind of things it's useful and it can be trustful, but you really need to give it a lot of information so that it can have relevant information to take from.*" (P_09)). As they pointed out, in order for the robot to bring the cloth items that fit people best, their particular taste must already be known and that requires a lot of prior information collection and storage. Without people's willingness to share such information, the robot can only suggest generic options.

Openness to future interaction

One perspective that seemed to reoccur among the participants was the requirement of added value in terms of practical use.

For many participants it was important that the robot was actually useful when considering if repeated interaction would be worthwhile (e.g, "*I would definitely try to understand how it can be used as some additional, let's say value to the shopping experience*" (P_06)). In the context of clothes shopping, usefulness meant to one participant that the robot would support their goal of more efficient and quicker shopping decisions (e.g., "*[...] the pepper wasn't like amplifying my goal oriented behavior. [...] if pepper would have been quote unquote tuned to my expected behavior, which is like maximize speed and comparison between items, because that's what I care about.*" (P_04)). If these expectations were not met, some of the participants could not see how having a robot mediating shopping experience would make any difference (e.g, "*I felt like I didn't really see a benefit compared to just a regular online shop. So if I have to imagine that I would have been in a physical store with that type of robot, I think I would be annoyed*" (P_02)).

Even though most participants found the robot rather useless in helping with the shopping activity, some of them were still very motivated or could easily imagine interacting with the robot in a more real-life situation. Simply from curiosity or interest in trying out the possibilities of what the robot could do to improve the clothes shopping experience, one participant would not hesitate to further interact with the robot if given the chance (e.g, "*test the function and where it can help me*" (P_03)). Another participants also imagined using a robot as a shopping assistant because of its potential to make shopping more enjoyable (e.g., "*I would definitely be interested to see to what extent this robot can actually help me in my shopping experience. Like if it would make it easier or more fun*" (P_05)). The novelty effect was also mentioned as important for the way the participants would consider interacting with the robot again in a similar real-life context (e.g, "*I would assume that there's some sort of novelty factor in the first time you interact with it. You're like, oh, this is cool*" (P_04)).

5.6 Discussion

After this presentation of the results, we now continue to the discussion of some of the most interesting points we can highlight from the analysis of both the interactive online survey and the semi-structured interviews. To begin with, however, we could not confirm our assumption that there is a difference between the kind of vulnerabilities that people reported they experience in the moment interacting with the robot compared to those they could imagine after given some time to reflect on their experience . From our analysis of the data, it seemed that all the different considerations were collected into a joint reflection based on both the past interaction and their later reflection. As such, we think our data suggests that the conceptualization of vulnerability experience in HRI as provided by Coeckelbergh (2013) remains a valuable analytical point, but nothing we could show empirically with the study design we chose. Further studies would be needed to really test this discrepancy assumption, because there could be many other reasons for why we failed to provide the required empirical evidence (e.g., maybe a stronger feeling of vulnerability is needed, maybe a good measure can only be made after different series of interactions, maybe a filed trial where people bump into the robot is a better situation). For now, when we compared the answers from the open-ended questions in the interactive online survey with the comments from the semi-structured interviews, there does not seem to be a relevant distinction to draw for our interpretation.

Considering that our follow-up study was also exploitative in nature, we noticed other interesting results from the data analysis that we think are relevant to our overall aim of understanding how to study trust in HRI with a special focus on a privacy scenario in the everyday life situation of clothes shopping. We will discuss these points also in view of current or previous work and discussions in HRI literature related to aspects of methodology, the trust and privacy relation, and the aim to design for trust in HRI.

5.6.1 Interactive Online Survey

From their systematic review of how methodologies and measures are evaluated in HRI, Jung et al. (2021) explain that it is a common perspective in the community to consider picture stimuli as rather limited because they cannot well represent the various changes in the way the robot moves, or how its facial expressions shifts during the interaction, which is necessary for a more realistic and authentic HRI experience. When comparing the results from both the original picture-based and the iterated video-based privacy scenario with the economy scenario (as presented in the previous chapter 4), we found it very interesting that there was no significant difference: whether we used the picture or video version for stimuli in relation to the trust ratings did not seem to matter.

This observation is relevant for current research on HRI since it suggests that the concern sometimes raised in the HRI community that a picture stimulus is insufficient for the participants to experience a shared context with the robot is less problematic. In our view, what seems to matter is not the specific choice of stimuli (pictures vs. video) but rather that the online study have a component that is interactive. While a video might provide a stronger experience of immersion in the interaction, having a sense of being able to interact with the robot in terms of responding to questions or requests (when using the online platform) might be also be very important. As such, we recommend to incorporate an interactive component to create the necessary atmosphere of shared context and interaction with the robot when studying trust in HRI because trust is a highly situated phenomenon.

Additionally, we reflected on the use of a online study design more generally as we agree with Feil-Seifer et al. (2021) that the impact that COVID-19 has had, and continues to have, on HRI research practice helps us reevaluate what previous assumptions about what validates a research design has dramatically changed in light of the global pandemic. As they point out, we might need a discussion in the HRI community about what role interactive online studies, which have previously been considered inferior to in-person studies when the latter are an option, can play in furthering research. Taking our interactive online survey as an example, we believe that the use of an online study to explore trust in HRI remotely is promising and that broader methodological discussion about the possible advantage and disadvantages will become very useful in times where no in-person HRI studies cannot be conducted in the foreseeable future for any reason.

5.6.2 Trust-Vulnerability Relation

Our methodology and study design choice was guided by insights from the less mainstream HRI disciplines of philosophy and sociology. Philosophical and sociological insights used to study trust in HRI are still mainly used to support existing understandings or definitions for the trust concept rather than knowledge that could provide the foundational assumptions for the specific methodology or study design. With our two online HRI studies, we take some initial steps towards a bigger discussion in the HRI community about how to better integrate disciplines whose valuable methods and perspectives are not always easy to translate into the standard format of a "user study".

In our previous proof-of-concept study, our modest aim was to explore if vulnerability was something that people experienced when studying trust in HRI in the first place, and whether it could be empirically studied using the subtle everyday life situation of clothes shopping. With our follow-up study, we had the chance to expand on this work by zooming in on the trust-vulnerability relation within the HRI context (within the privacy scenario specifically) to further investigate if human experience of vulnerability also functioned as a precondition of trust, as we came to learn from the philosophical conceptual analysis and discussions about the sociological socio-phenomenological accounts of interpersonal trust. From our statistical analysis, we found that there was a relation between those participants who said that they felt vulnerable and the significantly lower trust rating in the data as compared to those who said "no". We discussed, if this might suggest that the difference in scenarios (i.e., economy, privacy, and transparency) in the everyday life situation of clothes shopping was less important to the perception of trust than the feeling of vulnerability. When we then looked at the differences of scenarios again only for those who *did not* say that they felt vulnerable, we found that the trust ratings would vary depending on the given scenario.

While we cannot use these results from our follow-up study to confirm whether our theoretical assumption[10] can explain this pattern in our data as many other reasons needs to be considered, we believe that first steps have been taken to show how philosophical and sociological insights can also further empirical studies on trust in HRI that aspires to investigate the trust-vulnerability relation. In this sense, we join the growing group of voiced in the HRI community who advocate methodological plurality (see e.g., the work by Bethel and Murphy, 2010; Fraune et al., 2022; Weiss and Spiel, 2022), among which our follow-up study is just one example of how such considerations can help expand our approach to studying trust in HRI.

[10] We assumed that if vulnerability is a precondition for human trust in robots, it should have an influence on the trust ratings regardless of the specific scenario. In contrast, if people are not feeling vulnerable, they rate their trust depending on more contextual factors present in the different scenarios.

5.6.3 The Privacy Paradox

Early studies on privacy issues related to online shopping have revealed an inconsistency between the stated privacy attitude and behavior of people, which is now commonly referred to as the "privacy paradox" Norberg et al. (2007). As explained by Solove (2021), these studies suggests that people in general would not take very simple or free measures against possible threats to their privacy even though they express concern about it, and that they would give up personal information about themselves for a small compensation regardless of their view on privacy as something important. Studying the privacy paradox in relation to privacy issues in the specific context of HRI has with the work of Lutz and Tamó-Larrieux (2020) and Lutz and Tamò-Larrieux (2021) also recently been a topic of interest in the HRI community. With the assumption that robots promoted as more privacy-sensitive will have a positive effect on their acceptance and use by people, they conducted two different HRI studies on the privacy paradox. With a survey-based study Lutz and Tamó-Larrieux (2020) were able to confirm the privacy paradox in HRI as the results showed that participants mentioning concerns about informational and physical privacy had no effect on robot use intentions. However, with the results from their vignette-based study, Lutz and Tamó-Larrieux (2020) could no longer observe the privacy paradox. They discussed and explained that this difference in results was due to the difference in the way people might perceive privacy issues in general, compared to situations where they would be confronted with concrete privacy-invasive technologies.

The subtle violation of trust that we constructed for our follow-up interactive online study consisted of a privacy breach by (1) using the personal information about the participant's clothes items choice to automatically adjust the gender setting supporting the recommendations of the robot, and (2) storing and sharing this newly collected data with their consent only to train other robots, with the argument of personalizing future suggestions. Our results suggested that while a fair amount of our participants said they felt vulnerable, did not consider the robot benevolent, or sometimes both, they also expressed interest in interacting with the imperfect robot again in the same situation of clothes shopping in a real life situation. So although people had good reason not to trust the robot that provoked a mild privacy breach, it did not seem to keep some people from considering future interaction with it. While there is much discussion about whether the privacy paradox does in fact exist or not (Kokolakis, 2017; Solove, 2021), the results of our follow-up study seem to align with those studies in favor of it in the specific context of studying trust in HRI. Considering the philosophical analysis of trust and the knowledge that people give up their privacy for even a small benefit from the literature on the privacy paradox, we pose the question to current discussions on trust in HRI *for which reasons people decide* to interact with an imperfect robot if given another chance as the mild privacy breach that it provoked was intended to indicate lack of trustworthiness. As such, we encourage further discussions about whether the development of more privacy-sensitive robots will also lead to so-called trustworthy robots if considerations of the privacy paradox are left aside.

5.6.4 Tolerating Imperfect Robots

As presented in the philosophical analysis of interpersonal trust between humans (see e.g., the detailed account in chapter 2), people might judge it worthwhile to trust in others without knowing whether they are in fact trustworthy considering the added value of trust from the assumption that the other person act out of e.g., good-will or optimism (Baier, 1986), special personal and normative commitments (Hawley, 2014), motivation from affective attitudes (Jones, 1996) or care for the personal attachments they develop (Kirton, 2020). As such, the inner life of the other person is central to the judgment of whether the risk of trusting in a potentially untrustworthy other human is worth taking.

Considering the overall everyday life situation of clothes shopping, we found from our analysis of the semi-structured interviews that several participants on a very basic level considered the added value of having a robot assisting them merely in terms of utility (e.g., increased efficiency, suggestion of better options, and better understanding of individual clothing preferences). We saw that this view of added value as utility related to their instrumental view of the robot – it was simply a practical tool that needed to function well for the task at hand. On a more abstract level and related to their own knowledge about or specific interests in robots, entertainment purposes seemed to be the added value of having robots as clothes shopping assistants. The potential they saw in the robot, we came to understand, was mainly driven by their own optimism about improving technological solutions for the future of shopping, or from getting their curiosity met by exploring how a robot could provide an interesting or novel shopping experience. Though participant would use some anthropomorphic language to talk about the robot (e.g., referring to its gender or ascribing intentions to its behavior), they also reflected explicit on the fact that robots are ontological speaking (at least for now) not of the right kind to have an inner life let alone a rich one. Accordingly, it might not be too surprising that the added value of interacting with the imperfect robot provoking a privacy breach is based on the more simple or external criteria of utility, or entertainment factors.

Given the results of our follow-up study, we believe that we can help nuance the discussion about whether the efforts to design trustworthy robots will also turn out as a promising strategy for more successful HRI as often argued in the motivation for research on trust in HRI (e.g., Christoforakos et al., 2021; Langer et al., 2019; Martelaro et al., 2016). Because people seem willing to trade the potential added value of utility and entertainment for tolerating an imperfect robot provoking a privacy breach, the trust we might reach with robots in this specific everyday life situation of clothes shopping is much less dependent on normative and moral concerns as mainly perceived only to serve as tools to make tasks easier, or help stimulate. Designing explicitly for trustworthy robots by taking into account privacy-friendly features in this HRI context, seems helpful if the kind of trust in robots that we hope to achieve is related to factors of reliable performance in utility or entertainment value. However, the question remains if such efforts would also be the best strategy to the design of trustworthy robots where the ambition is to transfer the potential benefits of interpersonal trust as we see it in relationships between people into a HRI context. Future research on trust in HRI will benefit from exploring these issues in everyday life situations that goes beyond that of clothes shopping, where exactly normative and moral concerns are central to the interaction.

5.6.5 Limitations and Concluding Remarks

There are several limitations to our follow-up interactive online study that could be improved for the study design in the future.

First, we acknowledge that the study design of our iterated privacy survey is limited since it lacked a control group where people would not be exposed to any subtle trust violation consisting of a mild privacy breach. Consequently, our data analysis and interpretations of the results are not allowing us to draw any conclusions about whether it was the privacy breach in itself that in fact caused the low trust ratings and how this privacy breach relates specifically to the felt vulnerability of the participants. For our future work on studying situated trust in HRI and with a special interest in a privacy scenario, we will include a control group to improve our study design and to ensure that we can explore the exact relationship between the privacy breach and the way people might feel vulnerable.

Secondly, we used a relatively small sample size for the iterated survey (32 participants), which makes it harder to generalize the results of the questionnaire. It may well be that a larger sample size would enable us to find more effects. For our semi-structured interviews with the participants, we had even a smaller sample size. While it is a persistent methodological challenge to decide how many interviews are enough, due to the nature of interpretive iteration in the analysis of qualitative data (Patton, 2015), we believe that a collection of more interviews could have been helpful in determining whether the codes we arrived at were exhaustive, as prescribed by saturation criteria (Sebele-Mpofu, 2020). Thirdly, due to the nature of our follow-up online study, we did not offer any of the participants in-person interaction with the robot. The interaction took place through an interactive survey with video snippets of how the robot would have acted if it was an in-person meeting. While interaction with robots in the real-world is the most ideal set-up for conducting HRI studies, as we assume the embodied nature of robots is important for the interaction experience (Deng et al., 2019), video-based surveys have already proven an adequate alternative (Woods et al., 2006). Especially in the unusual situation of the global COVID-19 pandemic, where it is not possible to conduct “business as usual” with in-person HRI studies (Feil-Seifer et al., 2021), developing and using our interactive survey proved a useful replacement.

With our follow-up study, we aimed to further investigate human experience of vulnerability related to trust in HRI within a privacy scenario in the context of shopping. Considering the results of our follow-up study, we brought into discussion whether the common assumption in HRI that trust can be strengthened by also designing robots to be more sensitive to privacy concerns needs to be challenged. Not only did our participants tolerate an imperfect robot that put them at risk with a mild privacy breach, but they also considered the possible gain to be related to simple added values of utility and entertainment. It seems that when it comes to the the application of robots in the everyday life situation of clothes shopping, the development of trustworthy robots is less pertinent to successful interactions between humans and robots . For future work, we aim to conduct an in-person HRI study with and without a privacy breach provoked by the robot assistant in the public space of a clothes store to learn more about the interrelation of vulnerability, trust, and privacy.

CHAPTER 6

Vulnerabilities of Robots

> "[...] even if we question views that make a too strict distinction between humans and non-humans and between biological and artificial, we can and must distinguish between different kinds of vulnerabilities and vulnerability configurations. For example, if we share with (non-human) animals a biological body, then this means we also share a certain kind of vulnerability which is different from the vulnerability of a robot (at least those we know and can foresee today)."
>
> Coeckelbergh (2013, p. 197)

Most of the content in this chapter has already been published as Hannibal (2021) and part of this chapter will also appear as a book chapter organized and edited by the co-organizers of the Trust Robots Doctoral College (TU Wien) as: *Hannibal, G. & Weiss, A. (forthcoming, 2022). Exploring the Situated Vulnerabilities of Robots For Interpersonal Trust in HumanRobot Interaction (pp. 1-19). Vienna, Austria: TU Wien Academic Press.*

In the previous chapters 4 and 5, my empirical work on the trust-vulnerability relation for studies on trust in HRI has taken a human-centered perspective. In this chapter, I shift the focus towards a robot-centered perspective by presenting the results of expert interviews I conducted to investigate to what extent robots could also be considered vulnerable. My motivation for carrying out this exploratory work was rooted in the aim of understanding how the notion of vulnerability as a precondition of trust could extend over the entire interaction between humans and robots. While it is most common to speak about the shortcoming of robots in terms of failure, error, and fault (Honig and Oron-Gilad, 2018), I tried to introduce the notion of vulnerability for robots as a way to stay open-minded about potential categorizations or taxonomies normally intended to map such knowledge. As Honig and Oron-Gilad (2018) already observe from their extensive literature review on failures in HRI, there are also problems that only appear because interaction between humans and robots takes place. As I will show in this chapter, the way humans behave towards robots is also important for the way we can and should study trust in HRI, because robots are exposed to harm in their meeting with humans.

6.1 Study Aim

As I accounted for in chapter 3, the vulnerability notion has been studied within the HRI community as a property of robots in the form of self-disclosure (see e.g., Kaniarasu and Steinfeld, 2014; Martelaro et al., 2016; Sebo et al., 2018; Siino et al., 2008; Traeger et al., 2020). I also explained that on a conceptual level, operationalizing vulnerability as robot self-disclosure for studies on trust in HRI is problematic because a working definition is often lacking, and because the relational aspect of such a notion tends to be omitted in existing HRI studies. As I aim to show with the following empirical work, the reduction of vulnerability in HRI to a form of self-disclosure, raises additional problems because the design of vulnerability-related behavior of robots in the form of linguistic statements suggests a very narrow understanding of how robots could be considered vulnerable. In this form, the vulnerabilities of robots becomes an imitation of human vulnerability, which presumes that they are perceived as having an inner life. However, and considering the existing literature on robot failures (e.g., Honig and Oron-Gilad, 2018; Ragni et al., 2016; Salem et al., 2015b) 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: given that robots are of an ontological different kind, they have their own specific types of vulnerabilities that result from the way that interactions unfold in the specific situation. Hence, to systematically identify these robot-specific vulnerabilities is in fact equally important to identifying those of humans when exploring trust in HRI. This is a gap in the current HRI literature, which serves as the motivation for the presentation of the following work.

Already now, I like to clarify that my attempt to identify *the possible vulnerabilities of robots* is motivated by my interest in the conceptual scope of the trust-vulnerability relation in the specific context of HRI by also considering two different types of investigation. While I have already shown in the theoretical part of my dissertation (see e.g., chapter 2 and 3) that the conceptualization of vulnerability as a precondition for trust enables an understanding and analysis of trust in HRI as the result of the situated interaction between humans and robots (i.e., the event of trust rather than a property), this does not mean that the identification of vulnerability in such situated interactions are identical when taking a robot-centered perspective. To gain such insight, empirical work must be carried out to specify how the trust-vulnerability relation occurs during the interaction while at the same time ensure a sensitivity to the fundamental way human and robots differ: vulnerability from the perspectives of humans can manifest itself as something that they can attach their feelings or experiences, whereas it is only something for robots that can be ascribed to them from the perspective of those people designing or interacting with these robots. Consequently, I would characterize my investigation of robot vulnerabilities merely as an analytical exercise into its multi-dimensional aspects to challenge current understanding and analysis of trust in HRI.

6.2 Methodology

To explore vulnerability as a precondition of trust in HRI where the aim is to understand *in which ways robots could be considered vulnerable*, I decided to conduct expert interviews to elicit expert knowledge about what they might be. Different from lay knowledge that could be gathered from people who have no previous or only little experience interacting with robots, I find it beneficial to consider expert knowledge because it might generate an exploration into the vulnerabilities of robots that is guided mainly from what the technological systems are in fact capable of rather than what can be imagined or projected onto them (see e.g., the distinction between robots as technological systems vs. cultural object presented in chapter 1).

On a broader methodological level, the use of expert interviews is also important because of the ontological status of robots. First, given that robots do not have an inner life that connects their vulnerability to feelings or experiences, their particular vulnerabilities can only be studied from a third-person perspective. To paraphrase Latour (1993), whose views on scientific facts are equally relevant to this discussions, expert interviews are required because robots cannot "speak for themselves" (p. 29). Thus, it is advantageous to use the specialized knowledge of roboticists as a vehicle for giving expression to the specific vulnerabilities of robots. Secondly, and as mentioned in chapter 1, the method of conducting expert interviews is uniquely suited for gaining a more systematic overview of knowledge within a certain domain (Meuser and Nagel, 2009), the mastering of which requires many years of experience. For my purpose, expert interviews with experienced and leading roboticists are useful in the initial stage of identifying the possible vulnerabilities of robots. Not only do these experts have extensive knowledge about the technical challenges of developing robots, they can also provide insights into what types of vulnerability are common across various domains of application.

6.3 Expert Interviews

Over the period of nine months (December 2019 – June 2020), I conducted a total of eight semi-structured expert interviews. In the following, I briefly describe the interview process.

6.3.1 Sampling

I used purposeful sampling (Patton, 2015) as a method to recruit the experts with the following selection criteria (see e.g., Table 6.1 for a quick overview of how the different expertise was divided among the different experts):

1. Disciplinary background in robotics.

2. Work experience in HRI or social robotics.

3. Research interest on the topic of trust.

For the purpose of addressing the research question, it was enough for an expert to fulfill only one of the three criteria, although ideally they would cover all of them. To provide an overview of the selected experts, I asked them to introduce themselves briefly. The following description is based on their answers at the time when the interviews were conducted. Only one of the experts wished to remain anonymous; I have given this expert the code "Exp_XX". The table 6.1 provides an overview of experts and indicates the expert ID I assigned them, which will be used later for the analysis of results.

Expert	**ID**	**Affiliation**	**Expertise**
Justus Piater	Exp_JP	Department of Computer Science, University of Innsbruck (AT)	computer vision, ML, robotics
Allan Wagner	Exp_AW	Department of Aerospace Engineering, Penn State University (USA)	AI, robotics, HRI, robot ethics, trust
Marc Hanheide	Exp_MH	School of Computer Science, University of Lincoln (UK)	AI, robotics, HRI, social robotics
–	Exp_XX	–	social robotics, HRI, AI, trust
Birgit Graf	Exp_BG	Institute for Manufacturing Engineering and Automation, Fraunhofer IPA (DE)	HRI, service robotics, applications
Kristin Schaefer-Lay	Exp_KS	The Combat Capabilities Development Command, U.S. Army Research Laboratory (USA)	robotics, HRI, teams, trust
Michael Zillich	Exp_MZ	Research Development, Blue Danube Robotics (AT)	computer vision, robotics, HRI
Paul Robinette	Exp_PR	Department of Electrical & Computer Engineering, University of Massachusetts Lowell (USA)	robotics, HRI, trust

Table 6.1: An overview of the experts and the selection criteria for their inclusion.

6.3.2 Procedure

I contacted all experts via email with an invitation to participate, which also contained more background information and explained the purpose of the interview. After indicating 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 around 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. 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 system 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 ask questions and receive further clarifications, and were again informed about their rights as participants. Figure C.3 in appendix C provides a more detailed account of how the interviews proceeded considering the guidelines I used for interviewing the experts.

We intentionally left out a definition of vulnerability in the questions we asked the experts because we wanted to explore if they would be able to make sense and use of the vulnerability concept to speak about robots in the first place. Considering that some of the expert in fact asked about the intended meaning of robot vulnerabilities, we also brought this up as an interesting discussion point, which we will present later on in section 6.5.1.

6.4 Analysis of Results

After collecting all the expert interviews, the audio recordings were transcribed verbatim, with the spoken word as the only focus (McLellan et al., 2003). I solely coded the interviews using in-vivo coding[1] and from several cycles of coding, I collected the emerging codes into 13 different categories based on similarity of content and meaning (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 (Honig and Oron-Gilad, 2018; Ragni et al., 2016) and cybersecurity in robotics (Clark et al., 2017; Miller et al., 2018). I used 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 the software MAXQDA[2].

From the data analysis of the expert interviews, I was able to identify in total 13 categories of potential robot vulnerabilities since they cover all those weaknesses of robots that the experts considered would leave them most exposed to unsuccessful task completion or failure in smooth interaction or collaboration with humans. These categories were themselves grouped into four different themes (see e.g., Table 6.2 for an overview).

In the following, I will provide a short description of each theme and offer some examples of how they were supported by the different experts by drawing on their own wording, terminology, and formulations to summarize their main points.

[1] In-vivo is a term associated with the qualitative research methodology of Grounded Theory. In-vivo coding uses the specific words or formulations of those being interviewed as codes to label a section in the interview transcript (Miles et al., 2020).

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

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 6.2: A list of the different categories and themes identified during the coding and analysis of the expert interviews.

6.4.1 (T1) Embodiment

Because a tangible manifestation is required for robots to navigate among and interact with people *in the real world*, they have what Exp_JP and Exp_KS referred to as “physical vulnerability/vulnerabilities”. Under this theme of *embodiment*, I categorized all the various types of robot vulnerabilities that relate to their physical hardware or components.

(C1) Mechanical

Given that robots are made of various mechanical parts, Exp_JP explained that they can have “mechanical vulnerabilities” because some of their most basic components can easily be damaged. As he noted, it is often the case that “sensors and actuators, some of which are developed in the lab, [...] are quite fragile or surprisingly fragile”. Exp_JP also mentioned that sometimes there are “loose cables hanging off everywhere”. Some of the other experts stated that robots also need simple maintenance, e.g., “replace some tires that were worn down” (Exp_BG), and things like “batteries don’t stay charged the same as they did at the very beginning” (Exp_KS) need to be taken into account. For this reason, Exp_KS stress that when working with robots in real life application, “you need to understand the life cycle of those parts and pieces that make up the mechanics”.

(C2) Sensory

According to Exp_JP, there is also vulnerability related to how well robots can access the world around them in terms of their sensors being deficient, which he explained makes them less resilient because this “sensing is way underdeveloped compared to humans”. Exp_BG also stresses this point, saying that a robot “needs a lot of good perception abilities to constantly monitor the environment and changes in that”. Elaborating on how poor sensing can leave robots rather vulnerable, Exp_MH mentioned that they can easily get into serious accidents where they become physically damaged. He offers the example of how one of their robots being tested in a museum once “drove straight into the hole” that constructions workers had made to access some wiring in the floor, and describes how another robot tested in a eldercare home also broke simply because “the stairs would not be seen” by the robot in time before falling down them.

(C3) Functional

According to Exp_AW, one of the most intuitive robot vulnerabilities that comes to mind is that a “robot could break down”. Exp_MZ explains that this could be e.g., because a “cable is malfunctioning”. Thus, another area of vulnerability is technical problems hindering or interrupting robots in carrying out their tasks. Yet, this does not always have to be a bad thing according to Exp_MZ, who reflects on how their development of a skin for robots in fact leverages this kind of vulnerability to ensure better safety because “the skin is soft, and if you damage the skin, like if its punctured, it’s damaged and the robot can no longer move”. He further notes that this ability of the robot to be “hurt” is also very beneficial because such a “safety concept” helps protect the often expensive equipment that is mounted on robots.

(C4) Cybersecurity

According to Exp_MZ, robots are vulnerable to intrusions from outside in the sense that a robot “arm goes berserk because of a cyberhack”. Exp_KS also mentions that hacking is a serious problem when done with the ill intention of compromising the robot’s reliability, because “we need the information that is coming through our systems to be secure”. Exp_MH adds his concern about the kinds of dangers that come from underdeveloped measures against people interfering, because "robots are very badly implemented when it comes to cybersecurity". Exp_PR also points out that the challenge of safeguarding robots might be even greater in light of recent technological developments in related areas, since "a lot of the internet of things isn't very secure, and robots are being added into this sort of internet of things to create a greater system".

6.4.2 (T2) Processing

On a slightly more abstract level, but still related to how the robot is built, the second theme of robot vulnerability revolves around their ability to handle and use the information they collect from their surroundings, which Exp_JP expressed when he mentioned that "software[...] are also vulnerable". I have categorized the various types of vulnerabilities associated with the reasoning abilities of robots for interaction under the theme of *processing*.

(C5) Understanding

According to Exp_JP, one of the main challenges with robots today is that they "lack a conceptual framework that allows them to understand what is going on in the world". Elaborating on this point, he explains that a robot with "zero awareness" about its environment on a conceptual level is left vulnerable to some extent because "as soon as something doesn't quite work as expected then things will go wrong". Returning to the example by Exp_MH, the accident with the museum robot happened not only because it could not properly detect the visible signs of ongoing construction work, but also because the robot was not able to "understand the cues that we put into our environments to protect humans". Working on providing a robot with a better conceptual understanding of the world, Exp_JP adds, is important because the consequences can be quite serious given that "once you start filling the input that it's not prepared to handle the results will be unpredictable or even worse, you can provide software with adversarial input designed to throw it off".

(C6) Learning

Another kind of vulnerability that some of the experts identified is connected to the way in which robots acquire knowledge and use it to guide their behavior. While the development of "self-learning systems" is a desirable goal for making robots useful, Exp_MH also states that there is sometimes a concern about "robots learning the wrong thing". Despite robots being able to learn fast from the multitude of data they are given by people or they collect from interacting with their environment, he continues, robots are still missing the important ability to "judge if this is worth learning or whether it is a completely stupid thing". In a similar vein, Exp_XX mentions there is a potential for vulnerability of robots when they learn from developers, because they are not able to notice if the acquired knowledge is in fact correct or beneficial, and might in fact "learn bad behavior". Exp_XX further explains that the challenge of robots learning bad behavior relates to their lack of "a moral clause", which mean that they are not able to reason about what is morally right or wrong.

(C7) Decision-making

According to Exp_XX, another vulnerability that is important to mention concerns instances where "robots make decisions when they do not have all of the information". Because the decision-making by robots in such cases is based on incomplete prior knowledge, Exp_XX continues, they have a hard time dealing with new situations as they could be "making mistakes when there's more uncertainty in their dataset, in their algorithms". Reflecting on successful human-robot teams, Exp_KS emphasizes that robots could be considered vulnerable when they have difficulties determining whether or not to hand over the control to humans because they are restricted by "algorithms and this hard math in order to make the decisions". In cases of disagreement between humans and robots, this challenge could even lead to ethical issues, she explains, because robots could cause much damage when they are wrong in their decision to "automatically override" their own reasoning in favor of human judgment.

6.4.3 (T3) People

Moving on to aspects that are external to the robot but within its more immediate distance of reach, the next theme relates specifically to the actions or behavior of the people interacting with robots, and how this has a direct effect on their level of exposure. As Exp_MH has been observing during numerous field-trials with robots, there are many examples of how humans can cause much difficulty simply because “people do not understand how the robot sees the world”, or they too effectively “attribute sort of intentions and certain abilities to a robot”.

(C8) Obstacle

According to Exp_BG, people can cause vulnerabilities for robots when they unintentionally put up small hindrances in their paths, e.g., around their home because “people are moving in it, moving stuff around, moving chairs or whatever”. In this sense, she continues, robots have problems adapting to this sudden change and might risk not being able to complete their task since they are “not [...] able to drive to the target”. Using an example from a study where human-robot teams had to play the outdoor game Capture the Flag[3] on open water, Exp_PR explains that people sometimes intentionally use themselves as an obstacle to strategically outsmart robots because “human put themselves in a slightly unsafe situation that they know the robot will avoid in a certain way and thus force the robot into a different direction”. In his view, robots can be considered vulnerable from their interaction with cunning people as “most robots are at least at some level programmed for safety, and humans can take advantage of that”.

(C9) Perspective-taking

Another vulnerability that arises when humans interact with robots is according to Exp_MH that they get the robots into dangerous situations because of the limited “understanding in humans about how robots see the world”. He gives the example of how putting up striped bands around construction work is only taking into account the human perspective, because it is insufficient as a form of “risk assessment for robots”. Exp_MH then describes how his research team has been using a pair of virtual reality glasses showing the world from the perspective of the robot to educate people: this method makes it possible to “put yourself into the robot’s head”. Based on different evaluation studies in eldercare homes where Exp_BG observed that the staff or the elderly sometimes use robots wrongly when they are not intuitive in their design, she suggests making robots less exposed by making them “kind of self-explainable to ensure that people can deal with them”.

[3] In this game, Exp_PR explains, each team are competing in first getting the flag of the rival team to their own camp without being tagged.

(C10) Malicious

Exp_MH mentions that there are also more critical ways in which robots are vulnerable in their encounters with humans, because people are sometimes downright malicious and may engage in "abusive aggressive behavior towards robots". When introducing robots to the workplace, Exp_MH had observed that robots were sometimes regarded as competition to such a degree that "human workers are trying to sabotage the robotic systems so that it does not outperform them". Exp_PR also observed that people do not always have good intentions when they interact with robots, and sometimes "they are just trying to stop the robot from whatever its assigned task is". Exp_MH explains that the issue of malicious humans is also a problem in the development of robots that learn from interactions with people, because they "like to be very provocative, explore the edge of what is acceptable and eventually this thing learned something completely inappropriate". In his view, these vulnerabilities of robots are some of the most important challenges to address in order to ensure the safety of people, because "they are not only having a virtual embodiment, they have an actual physical embodiment, so they can do physical harm if they learn the wrong thing potentially."

6.4.4 (T4) Setting

The physical surrounding is often left unnoticed as a source of potential robot vulnerabilities, but Exp_JP points out it should be taken into account because "robots can do their jobs in lab settings but then [...] you change the setting a little bit and things won't work anymore". Thus with the last theme, I collect the expert assessments of the vulnerabilities of robots that relate to the specific application context or terrain in which the interaction between humans and robots unfolds or is made possible.

(C11) Infrastructure

The early stages of testing a robot in a real-world application scenario where frequent interactions with humans is to be expected, Exp_BG explains, really show how it can be considered vulnerable, because a robot in that state of development often requires "ten engineers standing around and making sure that it works". Robots become exposed easily on a very technical level when there is no adequate supervision, and there are also bureaucratic issues to consider when the vulnerability of robots relates to their potential use in industrial settings: in Exp_MZ's experience, "getting safety certified was a big procedure".

(C12) Environment

In the example by Exp_PR regarding the development of a robot to operate and be used on open water, robots are vulnerable to the forces of nature, as there is high chance of "the environment breaking the actual robot". According to Exp_MH, moving robots from the simple and confined laboratory setting into more dynamic and unstructured environments is problematic, and he has come to learn that "in the human habitats they are very vulnerable in general because this is not environment that is made for them". With his interest in longitudinal studies on HRI and from his own experience, he describes that people become aware of how vulnerable robots are from sharing living spaces with robots that are not sensitive to their limitations. He recalls a "good friend who has, more or less, unconsciously redesigned her place because she had discover[ed] that this robot gets trapped in this sort of thing". Exp_BG also mentions that robots are easily rendered vulnerable by the spacial design of eldercare homes or hospitals, because "sometimes the corridor was simply too narrow, the robot couldn't pass by".

(C13) Time

Exp_MH points out that robots are much more vulnerable in the early stages of testing or deployment as initially, people have a much higher interest in picking on them merely for amusement and as a way to explore their limitations, which he sees as problematic because "it's a novelty effect also which can have negative implications". Regarding the use of robots in a military context, Exp_KS explains that more long-term interaction could prevent people from overestimating robot capabilities and thereby placing them in more dangerous situations than they are designed to handle because if "you have more experience you're going to understand those vulnerabilities". In the more everyday context of the personal home, Exp_MH mentions that robots will become less exposed as people will try to compensate for their shortcomings through more extensive and repeated interactions, and that this willingness by people to step in shows that “adaptation will make a difference [over] a longer period of time”.

6.5 Discussion

Beside systematically mapping the potential vulnerabilities of robots from the data analysis, I came to reflect on some more general insights after interviewing the experts that I also consider relevant as two points for discussion when exploring the trust-vulnerability relation in the specific context of HRI. I will in the following sections present and connect these discussions points to relevant literature on trust in HRI, the philosophy of trust, and robot ethics.

6.5.1 Interpretations of Vulnerability

As expected, some of the experts commented on how to interpret the notion of vulnerability in relation to robots. For example, Exp_PR considered understanding robot vulnerability in light of how robots 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 closely relate to debates in HRI about how best to manage public expectations regarding the capabilities of robots. Known by now as the “expectation gap” (de Graaf et al., 2016; Kwon et al., 2016) it is also highly relevant to recent discussions about trust in HRI, as this gap could result in unwanted disappointment and even instigate fear (Malle et al., 2020).

More concerned with conceptual challenges, Exp_AW expressed difficulties with speaking about robot vulnerabilities, stating 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 that this issue made him hesitate to use the common definition of trust by Mayer et al. (1995) and instead turn towards a “definition that involved risk”, which is more practical and widespread in robotics because it is easier to operationalize. A 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”. Drawing on her more technical perspective, she then suggests to reformulate the relevant aspect of considering 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, p. 19-21), who write that it is still an open question “whether human-robot trust necessarily comes with a feeling of vulnerability that is characteristic of human trust”.

According to Exp_KS, it is necessary in these discussions to take into account that speaking about robot vulnerabilities also contains a normative dimension, because people in different contexts might need to ask themselves critically “how vulnerable do I 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 should almost always be avoided, while in a healthcare, it might in fact be useful 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 influences human group dynamics to the better (Traeger et al., 2020).

6.5.2 Ethical Dimensions

Coded 59 times„ it turned out that the theme of *People* (T3) ranked as the second most frequently mentioned robot vulnerability despite different domains of application . Especially the issue of malicious humans was mentioned by several experts, who noted that people would intentionally “kick”,“push”, “hit”, and “attack” robots, which adds to previous HRI literature reports on both adults and children engaging in such behavior (Brscić et al., 2015; Nomura et al., 2016; Scheeff et al., 2002). Assuming that this abusive behavior towards robots will happen more frequently with their increasing application in public spaces, which according to Exp_MH 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 not only to ask and study if 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 towards 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 understand 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. It is, from her perspective, 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 both power and capabilities. From her careful observation of interpersonal relationships of all kinds in which cooperation and care is cardinal, she recognized that some relationships are fundamentally unequal and sometimes not even voluntary, which severely challenges the liberal ideal of the conditions of trust. From this insight, Baier proposed instead to take trust to be a form of reliance on other people to act out of good will towards oneself. This so-called "goodwill" account of trust by Baier (1986) was not only important in stressing the close connection between interpersonal trust and moral obligation, but also was one of the first views on trust that went beyond reliance.

However, debates about mutual trust that are rooted in a liberal tradition become challenging for HRI because they presume that the two involved parties stand in an equal moral and power relation to each other (Faulkner and Simpson, 2017). The acknowledgment of robot vulnerability in relation to the human counterpart is then ethically problematic, as they can at most be considered "moral patients" (Coeckelbergh, 2018) who 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, I 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, which results 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 currently considered a rather controversial suggestion (Tavani, 2018).

6.5.3 Limitations and Concluding Remarks

Using the method of expert interviews for my investigation of robot vulnerabilities is not without limitations, which I will briefly present here.

First, deciding on a satisfactory sample size for the use of expert interviews can be difficult.. While statistics-based rules are typically used in quantitative research to set the sample size very precisely, the problem of determining and assessing the appropriate sample size for qualitative methods like expert interviews is an ongoing topic that usually ends in the conclusion that it depends on both theoretical, methodological and practical considerations (Sandelowski, 1995). Although various guiding principles for the decision on what is *enough* have been proposed, smaller sample sizes that are information-rich are often preferred in qualitative research, as they allow for more in-depth analysis (Patton, 2015). With my eight expert interviews, each of a duration of ca. 30 minutes, I have a rather small sample size. Given more time and resources, as well as better access to more experts, I would have liked to collect an additional 5-10 expert interviews to have more confidence in the categories and themes I have distilled. However, despite the fact that the categories and themes I have decided upon could be challenged in terms of generalizability and validity, they are nevertheless valuable for the aim of exploring robot vulnerabilities. Not only do they align well with the categories and themes already proposed in previous work on robot failures and cybersecurity issues in robotics, but category and theme saturation was also reached after all the expert interviews were analyzed[4]. Moreover, I agree with Sim et al. (2018) that qualitative sample size cannot be determined *a priori* because "what constitutes an adequate sample size to meet a study's aims is one that is necessarily a process of ongoing interpretation by the researcher" (p. 630).

Secondly, the way in which the cultural background of the selected experts might influence their perspective on a given topic is also important to consider.

[4] The criterion of saturation to determine and assess sample size in qualitative research was first introduced by Glaser and Strauss (1967) in their groundbreaking methodological approach of Grounded Theory. In their work, saturation refers to the point in the data collection where no further insights emerge from additional analysis and where all relevant categories and themes have been identified, explored, and exhausted. In this sense, the saturation criterion is more about adequacy than the actual sample size, and is today used widely in qualitative research beyond the approach of Grounded Theory. For a comprehensive overview of how the criterion of saturation has been discussed for sample size determination in qualitative research, please see e.g., Sebele-Mpofu (2020).

There has been a growing focus on the embeddedness of expert knowledge in milieus and socio-cultural settings (Maurtua et al., 2017). Moreover, studies on HRI have also suggested that the cultural background of people shapes how they come to view robots and their potential role in society (Lim et al., 2021). Especially comparisons between countries belonging to either the so-called "Western" or "Eastern" part of the world have gained much interest[5], as different expectations towards robots and their acceptance by the population have been explained with reference to underlying religious or philosophical world-views (Coeckelbergh, 2013; Robertson, 2018). There is no reason to believe that experts are free from such bias. Considering that the development of robots is spread out across various research laboratories, independent institutions, and commercial companies around the world, the pool of experts is not homogenous in terms of their cultural background. My eight experts were all from or working in the so-called Western part of the world (i.e., Europe and North America), which means that no cross-cultural perspectives were included in this investigation. Even though such perspectives would be interesting to also include for a more global view on how experts view robots, to do so would be beyond the scope of my dissertation.

My aim in this chapter was to map out in a systematic manner, the vulnerabilities of robots to explore the trust-vulnerability relation in the context of HRI specifically. By interviewing experts in robotics, I was surprised to find that one of main weaknesses or imperfections of robots that could potentially leave them exposed was less to their own capabilities but rather to the malicious intentions of people as either bystanders or interaction partners. Since interpersonal trust between people often require normative or moral concerns to be included to be beneficial for the relationship or interaction, it is important to also ask if this would also be important to strengthening trust between humans and robots. If people intentionally act with abusive behavior towards robots, trust from a robot-centered view cannot rest on the expectation of mutual recognition of the others vulnerability. Whether the HRI community would benefit from taking such perspective into account when studying trust in HRI is a question I leave open for discussion. Yet, I think it is a dimension of trust that is yet to be considered more seriously given that robots in the future will be expected to interact and collabo-

[5] Cross-cultural studies on the perception of robots between the US and Japan has been at the center of many discussions, with arguments both for and against the view that Japan is a particularly "robot-loving society" (MacDorman et al., 2009, p. 507)

rate with people in situations that require trust. The mutual consideration of the others vulnerability, from a more practical consideration, would be important to the development of robots capable of acting out of trust just as much as it would be to humans where it is used as a strategy to deal with their vulnerability. In this relation, we also have to consider if people would even consider the shortcoming and weaknesses of robots as their special kind of vulnerabilities. The conceptual work needed to determine if the concept of vulnerability would be useful for such purpose is for future work. For now, I have taken only initial steps to explore how the concept of vulnerability can be used as a conceptual tool to open up the discussion about trust in HRI in a new light as the imperfection or weaknesses of robots emerge as a result from their interaction with people.

CHAPTER 7

Meta Discussion

> "That is the nature of explorations: you don't know what you will learn, but you will learn something."
>
> Hawkins (2021, p. 237)

The content presented in section 7.1 is based on discussions with Nicholas Rabb (Tufts University), Theresa Law (Tufts University), and Patrícia Alves-Oliveira (University of Washington) for our THEORIA workshop held in conjunction with HRI'22. The description of the workshop has already been published as Hannibal et al. (2022b).

Section 7.2 contain reflections and insights drawn from my discussions with Patrícia Alves-Oliveira (University of Washington) about the basic principles and usage of design thinking.

Up to this point, I have with the chapters 1-6 presented the motivation and approach of my PhD project and provided both the theoretical perspectives and empirical work needed to explore the trust-vulnerability relation for research on trust in HRI. My overall aim was to show that *an emphasis on vulnerability for studying trust in HRI is important for deepening our understanding and analysis of* *interpersonal trust between people and robots intentionally designed to have apparent agency*. To quickly revisit, I intended with the chapters 2 and 3 to account for the concepts of trust and vulnerability to make them applicable to research on trust in the specific context of HRI. In the chapters 4 and 5, I explored how human experience of vulnerability and their trust perception relate through two online HRI studies where a robot assisting them in the everyday life situation of clothes shopping instigated a subtle trust violation instance. Based on interviews with HRI and robotics experts , I presented in chapter 6 an overview of the possible vulnerabilities of robots, and discussed how the case of malicious people creates a challenge for trust in HRI because robots are left exposed from this kind of encounters. Since my intention with this discussion chapter is to present how my contributions relates to the existing and interdisciplinary research landscape of HRI and supports the new movement towards experimental philosophy, I have moved the answers of my overall research questions and sub-questions into the concluding chapter 8, where I also mention some of the limitations of my PhD project .

What I present in the following chapter is a broader discussion about how my work contributes to current research on trust in HRI by considering the ways in which the knowledge I gained can most beneficially be transferred and integrated into the HRI community. To this end, I take a more *reflexive perspective* on how I view myself as a theory-driven and multidisciplinary researcher. In this sense, my discussion chapter is atypical for the field of HRI, but it is my aim to show that these kinds of meta-perspectives are important to bring up because they center questions about how to take part in and contribute to HRI research that is constantly moving across disciplinary borders – philosophy being among them. Consequently, and supported by my interest in the philosophy of science, I will focus this meta discussion on the benefits and challenges of using a theory-driven and multidisciplinary approach to the study of trust in HRI. This discussion is also the outcome of some more general reflections related to two workshops I led and conducted in collaboration with other HRI researchers to stimulate my own thoughts on this matter. The first workshop concerned the transfer of knowledge with regard to the role of theory and theorizing in HRI, while the other focused on the challenge of integrating theoretical knowledge about the trust topic into engineering practice aiming to develop trustworthy robots. At the end of this meta discussion , I also take the opportunity to reflect on how philosophical insights in particular can be considered valuable to research on HRI, and how my own work reflects a current trend towards experimental philosophy.

7.1 Knowledge Transfer

In chapter 2 and 3 I gave much attention to the theoretical foundation of the trust concept, how it is linked to the precondition of vulnerability, and what kind of vulnerability notion is relevant to the study of trust in HRI. This work was not only important for clarifying my own thinking about what is meant by trust in HRI and why vulnerability is an essential constitutive element, but also for ensuring that my empirical work was targeting the few and specific conceptual relationships that would be possible to study with current humanistic and scientific methods. My studies on trust in HRI and the results were reported in the chapters 4, 5, and 6, which mainly focused on the human vulnerability experience, though some room was also left for considering how robots could also be considered vulnerable . The intention was to show how the theoretical work in terms of conceptual analysis enabled the identification and formation of those conceptual relations that would be possible to investigate empirically considering the sparse knowledge and few suggestions in current HRI research about the connection between the notion of vulnerability and the concept of trust.

7.1.1 Conversation Starter

While working on my PhD project and writing on this dissertation, I became more and more aware of the lack of theoretical knowledge considered and used to guide current research on trust in HRI. This was all at the same time as an increasing interest in the trust topic began to spread like ripples in water within the HRI community. Especially at the 2021 ACM/IEEE international conferences on Human-Robot Interaction (HRI'21) and IEEE international conference on Robot & Human Interactive Communication (RO-MAN'21), there was an extensive number of contributions that focused on the topic of trust and its relation to other important HRI themes. At HRI'21, I found myself quickly connecting to several members of the community who also wanted to discuss the topic of trust, mainly in relation to its measurement, but to some extent also in terms of the more theoretical underpinnings. Few of my colleagues were as distinctly theory-driven in their approach as myself, and it seemed that the rest did not consider deeper theoretical discussions very important for their studies on trust in HRI. While I acknowledge that it sometimes beneficial to use a common and well established definition of trust to derive a hypothesis for an empirical study, this approach to studying trust in HRI was considered by the HRI community to be not only typical, but also sufficient. I consider this problematic, as there are many metaphysical and epistemological assumptions embedded in the various trust definitions that are oftentimes not explicated or taken into consideration when applying the trust concept to the HRI context. Consequently, much of the previous work on trust in HRI was sometime vague or unsure about what they actually measured in their studies, and what the results could really contribute back to discussions about trust as a phenomenon that is now extended to interactions between humans and robots.

My own impression was that most researchers were in fact measuring trust as reliance, which is neither problematic nor very surprising. What was an issue, in my my views, is that the results from such studies were then use to foster the research agenda of trust being promoted as essential for iteraction and collaboration with robots and their acceptance in society in general. Considering my work on both the theretical perspective and empirical work for studying trust in HRI„ things are not that simple. As I have showed in chapter 2, not only does the interest in trust beyond mere reliance seems to be of most relevance to cases where robots are designed with apparent agency, it also turns out that the interpersonal trust notion that I have been concerned with in this dissertation cannot be applied directly to the context of HRI without some serious metaphysical modifications: it requires us to think about trust in HRI as an event located in the interaction between humans and robots rather than being a property of either parties involved. Moreover, and as I discuss in chapter 5 where I present my follow-up study it seems that the potential added value of trust in HRI considering the everyday life situation of clothes shopping is less depend on normative or moral concerns but rather on those of mere utility or entertainment. Thus, while a modified understanding of interpersonal trust could be applicable to the HRI context and that this understanding can lead to interesting and new hypotheses that can be tested empirically, there is still much work left to be done to really understand what this new understanding and analysis of trust in HRI means in terms of improving interaction, acceptance and collaboration with robots. We are only at the beginning of exploring and understanding what implications follow from considering the trust-vulnerability relation for studying trust in HRI, and how such work might challenge current theoretical knowledge on the topic as a phenomenon normally reserved for the interactions and relationships between people.

As a response to my observations and concerns about how theory-driven approaches to research on HRI have been neglected in the HRI community, I planned and organized with three colleagues a workshop on theory and theorizing in HRI for the HRI'22 conference – the "Theory-Grounded Human-Robot Interaction" (THEORIA) workshop[1]. Our aim with this workshop was not only to create and foster an HRI community around theory and theorizing through collaboration, knowledge exchange, and networking; we also strived to highlight and discuss with people in the community the importance of theory and theorizing for the development of HRI into an established discipline with its own standards for the scientific and engineering practices (Hannibal et al., 2022b). As I argued in the introductory talk at our workshop, the ongoing discussion about whether to start with theoretical knowledge or quickly proceed to empirical work to ensure successful HRI research reflects the old dispute in the philosophy of science between Plato and Aristotle about how we best gain knowledge about the world. In the initial shaping of HRI into an independent field of research, in my view, it seemed that researchers were driven by a strong interest in applying the theoretical insights from social psychology and communication studies to form a paradigm shift in the design of robots so that the primary focus would be on making them socially capable. In this context, the specific design of these robots would be tested and validated through HRI studies mainly drawing on methods from the methodological tradition of positivism. But now, the scope of interest also includes HRI studies that are purely human-centered, as they seek to determine how people perceive and relate to robots through the application and testing of interpersonal phenomena (Lee et al., 2022). This new focus also led to the methodological debates about the integration of methods belonging to the interpretive tradition (Seibt et al., 2021). In this sense, the advancement of HRI into an established discipline has mainly been driven by an interest in the empirical work, which by nature relied mainly on practical knowledge (i.e., understanding through hands-on and personal experience), rather than theoret-

[1] Link to the website of our THEORIA workshop: https://theoriahri.weebly.com

ical knowledge (i.e., understanding through reasoning about the fundamental principles behind a phenomenon). This space left for theoretical knowledge to help push the HRI research field further has gradually been shrinking. Thus, the questions we aimed to raise or address during our workshop revolved around how the community could strike a better balance and encourage a dialogue between theoretical knowledge and empirical work for future generation of researchers (see e.g., Figure 7.1).

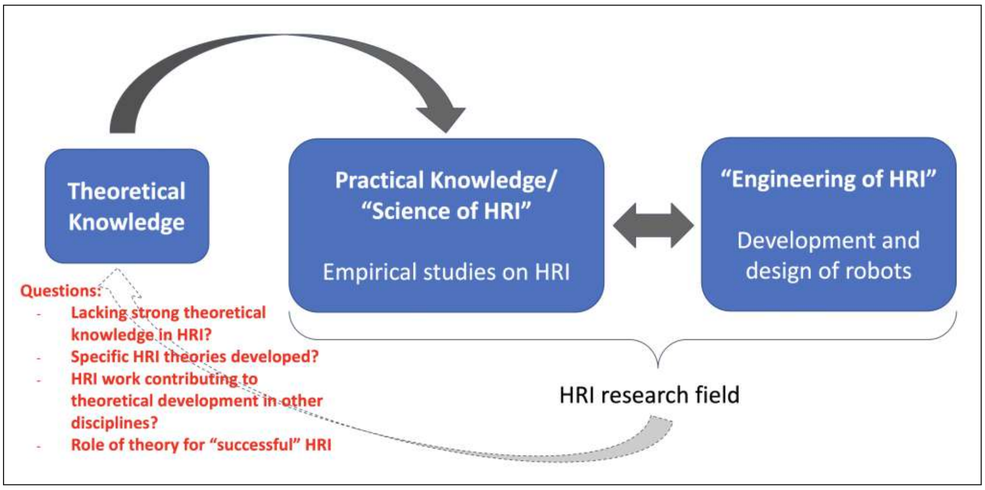

Theoretical Knowledge
Practical Knowledge/ "Science of HRI"
Empirical studies on HRI
"Engineering of HRI"
Development and design of robots
Questions:
- Lacking strong theoretical knowledge in HRI?
- Specific HRI theories developed?
- HRI work contributing to theoretical development in other disciplines?
- Role of theory for "successful" HRI
HRI research field

Figure 7.1: Visualization of the motivation for our THEORIA workshop.

7.1.2 Exactness and Dialogue

During our THEORIA workshop described above, we covered many different challenges and opportunities that people in HRI face when applying theories or engaging in theorizing. While it is not possible to present the discussion in full, nor to dive into all the important nuances of the discussion, there were two important reasons for advocating more theory-driven research in HRI that were mentioned by several of the workshop participants.

First, there seemed to be agreement about the significance of increasing theoretical knowledge to support HRI studies, as this would help ensure more rigorous results. As pointed out by our workshop participants, and well explained by Wacker (1998), a strong theoretical foundation is vital to any area of research because it "defines the variables, specifies the domain, builds internally consistent relationships, and makes specific predictions" (p. 361). Given that theorizing delivers definitions, domain, relationships, explanations, and predictions, it is necessary for the research field of HRI to acknowledge and nurture theory-driven approaches to support methodological discussions central to the vast empirical work done already. Connected to discussion about whether HRI would also risk suffering a crisis of replication similar to the "replication crisis" in psychology[2], it was also stressed that the research field of HRI would be strengthened if more effort was put into providing a robust theoretical foundation, especially considering the many diverging results from empirical work. Emerging from these workshop discussions, and recently presented as a longer argument by Leichtmann et al. (2022), the development and application of solid theoretical frameworks is among the factors needed to steer the research field of HRI away from a replication crisis, because the conceptualization and definition of constructs will be better specified, conceptual relationships will be clarified, and boundary conditions and auxiliary assumptions will be spotted. What the workshop discussion shed light on was the underlying need for researchers in HRI to be familiar with and use established theories from other disciplines, and to engage with theory-building or link their results to theoretical discussions. Considering the work I have carried out as part of my PhD project and written up in this dissertation, a theory-driven approach has enabled me to bring in new perspective from the philosophy of trust to substantiate current understanding and analysis of trust that aims to study trust in HRI specifically – insights that might eventually contribute to our body of knowledge about what make interactions between people and robots successful.

Secondly, it was also pointed out by some of our workshop participants that the kind of skill set acquired and refined when working with or creating theoretical knowledge is very valuable to the research field of HRI. HRI researchers from more theory-heavy disciplines have much to offer in terms of providing clarity to the core concepts in HRI, but the careful dissecting of the conceptual relationships studied in experimental study designs tends to be neglected because few know how to carry out this work in a way that allows this knowledge to also be well integrated. Because theory and theorizing in HRI requires much sensitivity to the nuances needed to interpret the accumulated knowledge gained from both abstract and empirical work, such focus is often dismissed as either

[2] Leichtmann et al. (2022) present a short historical account of how the crisis of replication resulted from a series of events in the study of psychology, and how it came to influence its confidence as a rigorous research field.

over-complicating things or being too speculative. I think the main problem is that such theoretical and often fundamental knowledge is challenging to feed back directly into the concrete development of robots, which I came to learn for myself while developing the discussion chapter for my dissertation. Coming from the humanities or social sciences, where theory and theorizing account for much of the work, the struggle in getting these kinds of competences and interests acknowledged as valuable for the HRI community can sometimes lead to a sense of exclusion or rejection, as I have experienced myself. Securing more cohesion and constructive dialogue among the various kinds of researchers coming together to advance HRI could in this sense begin with encouraging and giving space to theory-driven HRI. This point was close to my heart – while Dyson (2015) spoke about the field of mathematics, I believe his great insight into the different kinds of perspectives in a field is also helpful in understanding the conflicts that sometimes surface in HRI[3]. Describing the various types of researchers working in mathematics, Dyson (2015) writes that:

> “*Some mathematicians are birds, others are frogs. Birds fly high in the air and survey broad vistas of mathematics out to the far horizon. They delight in concepts that unify our thinking and bring together diverse problems from different parts of the landscape. Frogs live in the mud below and see only the flowers that grow nearby. They delight in the details of particular objects, and they solve problems one at a time*” (p. 37).

Using this metaphorical language, the research field of HRI is very much a pond too, with both birds and frogs visiting. I hope one day for myself to be a “bird-frog” – though I do not think it is a requirement for people doing HRI research to be such hybrids in order to bring the field forward. As I will elaborate on in the following section, I believe that there is great value in having philosophers in an HRI research team that are can combine their theoretical perspectives and training in conceptual analysis with a good understanding of how to develop and conduct empirical work to support the growth of knowledge. Even if they do not want to carry out such empirical work, the important contributions that a philosopher can bring to research in HRI are not lost as long as they are willing and

[3] I am only using this pond metaphor by Dryson as a narrative device because I do not want to imply with this use any value judgment by suggesting that a “bird” is any better than a “frog”.

able to explain to others in the field of HRI how their insights can be transferred and integrated. As such and as Dyson (2015) continues, it would be enough for the research field of HRI to cherish that both birds and frogs find this pond a suitable habitat:

> *"[...] Mathematics needs both birds and frogs. Mathematics is rich and beautiful because birds give it broad visions and frogs give it intricate details. Mathematics is both great art and important science, because it combines generality of concepts with depth of structures. It is stupid to claim that birds are better than frogs because they see farther, or that frogs are better than birds because they see deeper. The world of mathematics is both broad and deep, and we need birds and frogs working together to explore it"* (p. 37).

With my theory-driven approach to the study of trust in HRI, through the conceptual analysis of the trust concept and its link to the vulnerability notion, my work contributes to the research field of HRI by increasing thoroughness and by mediating a better communication between those frogs who have a hard time taking the perspective of the bird, and those birds who struggle to fly a bit lower for a different view.

7.2 Knowledge Integration

Having a disciplinary background in philosophy and some training in sociology (mainly from the interpretive tradition), I came into the research field of HRI with a perspective and skill set characteristic to the humanities. I had learned about empirical methodology and methods through course work prior to this PhD project, which also gave me the opportunity to jump straight into the cold water that data collection and analysis can sometimes feel like. As such, I consider myself mainly a philosopher who is also becoming something of a scientist (as I am interested in conducting empirical work to inform my thinking and research) given the empirical work that I have presented in this dissertation. The knowledge, skills, and practice of engineers and designers was something I had only touched upon very briefly during my short stay in the MA Philosophy of Science, Technology and Society (PSTS) program at the University of Twente. Throughout my time as a PhD student in a Computer Science department, I have therefore thought about how my work contributes to the research field of HRI that stands on the two legs of science and engineering. While I was very confident in how I both learned and became more familiar with the role of being a scientist, understanding and finding a space for myself among engineers was very challenging. How I have come to think about it and the way I position myself might not be satisfying to everyone, but I find it a useful first step towards a bigger discussion about how empirically informed philosophers could support the work of engineers for fruitful collaborations.

7.2.1 Design Method

The basic question I was asking myself for quite some time while working on the topic of trust in HRI was how I could contribute to the areas of HRI that are related to the engineering practice, without having myself been building any technological solutions to improve the development or design of robots. For quite some time, I thought I would end up with a discussion chapter where I would translate the finding of my empirical work into some kind of design matrix of the vulnerability-trust relation, that engineers could then use in their efforts to develop robots that are trustworthy. Along the way, I slowly realized that this would not be the outcome of my work, because I was very much using a theory-driven approach that set out to test conceptual relationships, rather than a bottom-up and explorative approach where the implicit and explicit needs or desires of people using robots would be made visible[4].

To identify and gain a better idea about how my work could support the development of robots to improve trust in HRI, I decided to team up with Dr. Patrícia Alves-Oliveira (University of Washington), who is experienced in design thinking and also has an interdisciplinary background as an HRI researcher. After several online meetings discussing how to provide a form of *knowledge transfer* for the insights I obtained from my conceptual analysis and the understanding I acquired of the trust phenomenon through the empirical work, we wanted to use the common human-computer interaction (HCI) method of a *card-based design tool* to prompt reflection and discussion among engineers about how to support

[4] I later came to learn that this bottom-up and explorative approach to understanding and analyzing technological solutions belongs to the aspect of design thinking and practice, which most engineers are concerned with at some point in the process of technology development. However, design and engineering is not the same, though they have many overlaps.

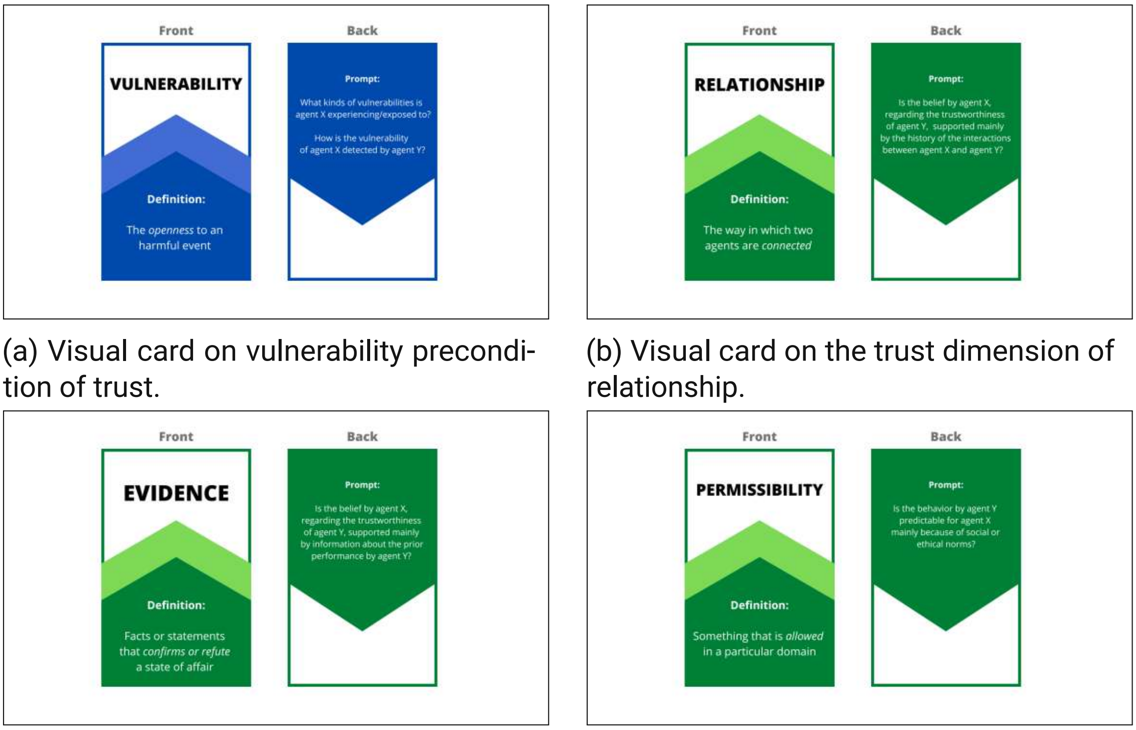
their design process for developing trustworthy robots. Card-based design tools consist of physical cards that serve many different purposes in facilitating e.g., creative thinking, knowledge uptake, problem framing, critical discussion, shared understanding, fact checking, and collective task definition (Roy and Warren, 2019; Wölfel and Merritt, 2013). In HRI, this method has mainly been used for the process of designing robots or the interactions with them through participatory design (PD) or co-design with children or the elderly (Schwaninger et al., 2021). Targeting engineers with this method for exploring the development and design of trustworthy robots had to our knowledge so far not yet been considered in the current HRI literature, even though cards could offer to support knowledge and integration of trust by extending the knowledge engineers might already have, as well as by bringing about new reflections.

We combined the method of a card-based design tool with *scenario-based design methodology*: according to Carrol (1999), “scenarios are stories” (p. 2) that contain the elements of a setting, a plot, agents, and objectives to guide imagined situations of future use. Scenarios are also used as a method in HRI, where they focus on the construction of a narrative about the situation, activity, goal(s) and action(s) of the people intended to interact with a robot, using different kinds of media for presentation (e.g., text, pictures, video) Xu et al. (2015). Scenario-based design methodology is considered a useful approach to the development of robots because it enables a better understanding of the possible interplay of the human and the robot for real-world application. While Carrol (1999) presents five different reasons for the advantages of the scenario-based design methodology, our motivation was to use scenarios mainly to support engineers in identifying and reflecting upon the different ways robots can be used by those people intended to interact with them, while also considering the constrains given by the conceptual knowledge on trust currently available. As he explains, since people using computers do not focus on low-level technical aspects (e.g., mouse click, typing, giving commands) but rather on what they can achieve with them (e.g., get work done, play games, communicate), it is important to focus on the experience people have about their own intentions and activities as the target level of design consideration. Especially for the case of trust in HRI, the development of trustworthy robots must take into account how people view the interaction in terms of their own understanding and experience of trust. Based on these considerations, we did some brainstorming and designed the specific visual cards we wanted to use to explore the design opportunities for trust in HRI. These cards were a mix of conceptual elements of trust related to either the situation (blue color) or the human and the robot (green color). On the front of the card, we provide a definition of the concept, and on the back, two questions to prompt reflection (see e.g., Figure 7.2 for an example of the visual cards). We also wrote two fictional scenarios, which revolved around a situation where a person in a hurry needed to find their way to a specific location. With the option of using a flow-chart and/or the the method of pseudo code for representation, we planned to ask the engineers to write one algorithm (or more) for programming the robot's behavior in such a way that the person in this fictional scenario would trust the robot to bring them to the desired place as quickly as possible. Additionally, we also provided a specific definition of trust and a conceptual model that could be used for reference in reflections and discussions.

To test our set-up, we organized in the beginning of August 2021 an online pilot workshop with five engineers experienced with both robotics and HRI, and who were also part of the Trust Robots Doctoral College (TRDC) at TU Wien. For the workshop, they were divided into two groups. They were first not given any cards or additional information about trust before trying to design a trustworthy robot using the previously described scenario. Afterwards, in the same groups, they were handed the cards to help them solve the task given the other, similar scenario.

7.2.2 Framing and Understanding

After using our set-up of visual cards, fictional scenarios, and the conceptual model and definition of trust to help reflect on how to design for trust in HRI, we talked to participants after the workshop and asked them where they saw benefits or challenges in using the card to stimulate their approach to the development of trustworthy robots. We received feedback from participants about the usability of the cards, and they shared their overall perspective on how the level of information provided on the topic of trust was useful or not. As such, we took initial steps to obtain a first impression of how our card-based design tool and scenario-based design methodology could support practitioners such as engineers in facilitating considerations of possible constraints when designing for trust in HRI. While there are many different suggestions in the current software engineering literature


Front
Back
VULNERABILITY
Definition:
The *openness* to an harmful event
Prompt:
What kinds of vulnerabilities is agent X experiencing/exposed to?
How is the vulnerability of agent X detected by agent Y?
Front
Back
RELATIONSHIP
Definition:
The way in which two agents are *connected*
Prompt:
Is the belief by agent X, regarding the trustworthiness of agent Y, supported mainly by the history of the interactions between agent X and agent Y?
Front
Back
EVIDENCE
Definition:
Facts or statements that *confirms or refute* a state of affair
Prompt:
Is the belief by agent X, regarding the trustworthiness of agent Y, supported mainly by information about the prior performance by agent Y?
Front
Back
PERMISSIBILITY
Definition:
Something that is *allowed* in a particular domain
Prompt:
Is the behavior by agent Y predictable for agent X mainly because of social or ethical norms?

(a) Visual card on vulnerability precondition of trust.

(b) Visual card on the trust dimension of relationship.

(c) Visual card on the trust dimension of evidence.

(d) Visual card on the trust dimension of permissible.

Figure 7.2: Example of the visual cards we developed for the online pilot workshop.

about how to view engineering practice, with various "software development life cycle" (SDLC) models proposed to capture the processes of development and design (Ruparelia, 2010)[5], I agree with Sheppard et al. (2006) that engineering practice is essentially about *constraint-based problem solving*. As Sheppard et al. (2006) also explain, the kinds of problems that engineers work with are motivated either by the identification of a need or by posing a question where the applications of technologies constitutes the solution. As such, I consider the recursive stages relevant to robotics and HRI as examples of engineering to be:

1. Problem framing (asking about the nature of the problem).
2. Knowledge gathering (learning about the relevant dimensions of the problem).

[5] Ruparelia (2010) provides a good overview of the different SDLC models such as e.g., the waterfall model, V-model, incremental model, spiral model, rapid application development, agile, and scrum.

3. Problem solving (imagining the possible technological solutions to the problem).
4. Candidate solution (plan in detail what kinds of technological and scientific requirements are needed to solve the problem).
5. Prototype/system design (implementing a functioning version of the purposed technological solution).
6. Evaluation activity (testing if the technological solution succeed or fail in addressing the problem).
7. Solution optimization (improve the technological solution based on the experience gained from the evaluation).

Overall, participants expressed that their participation in the online pilot workshop had helped them reflect on their integration and strategies for developing trustworthy robots, which became more focused when using the visual cards. As they mentioned, the visual cards added an extra layer to their knowledge about the different dimensions of trust in HRI that could and should be taken into account in the initial stages of the design process. Three relevant points were raised that significantly improved my own understanding of how our attempt of knowledge transfer maps onto various aspects of the engineering practice. First, the visual cards enabled participants to discuss whether their assumptions about the technological requirements in fact addressed the issue of trust in the fictional scenarios. The conceptual analysis of trust, which supported the provided content of the visual cards, helped participants understand and think about what kind of problems needed to be addressed when developing robots meant to behave in a trustworthy way. Secondly, participants discussed how the visual cards provided some clarification about how the concept of trust should be understood in this specific context of HRI. While they were all somewhat familiar with the topic of trust in HRI, many of them did not work explicitly with this concept in their own work, or did not have any extended understanding of its conceptual elements. Having the visual cards to support their individual understanding or provide input to the group discussion gave some of the participants more confidence to link knowledge about trust with the imagination of a possible technological solution. These two points raised during the discussion show that visual cards could be one way of transferring more theoretical knowledge into a format that is helpful in the first two stages of engineering practice, which focus on the problem framing and knowledge gathering about trust central to the problem they needed to address with their technological solution. Thirdly, some of the participants reflected on how the visual cards facilitated knowledge integration in a way that was different to a mere a check-list that they could use to ensure that trust has been properly addressed in the way the behaviors of robots were imagined. There seemed to be an underlying wish that the visual cards be used beyond abstract group discussions about possible constraints to instead guide the concrete implementation of the knowledge provided. Since the visual cards were not intended to restrict or provide a single "correct" way for discussion about how to design for trust in HRI, but rather were meant to leave room for exploration and discussion, this was a fundamental limitation to our approach of knowledge transfer through design thinking.

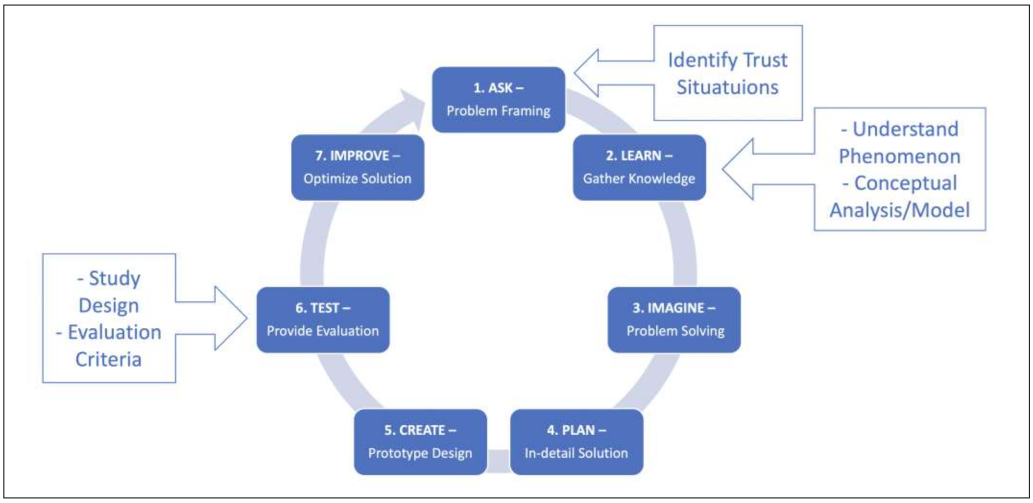

Identify Trust Situatuions
1. ASK – Problem Framing
2. LEARN – Gather Knowledge
- Understand Phenomenon - Conceptual Analysis/Model
3. IMAGINE – Problem Solving
4. PLAN – In-detail Solution
5. CREATE – Prototype Design
6. TEST – Provide Evaluation
- Study Design - Evaluation Criteria
7. IMPROVE – Optimize Solution

Figure 7.3: Visualization of how my theoretical knowledge and empirical work contribute and match the requirements engineering practice.

7.2.3 Evaluation with Scenarios

The realization that our online pilot workshop limited the knowledge transfer only to the very initial stages of the design practice made me wonder whether my theoretical knowledge and empirical work resulting from the PhD project could also support engineers in the stages closer to the concrete development of a technological solution. Since the knowledge resulting from the conceptual analysis and the online HRI studies were not on the level of technological and scientific detail needed for concrete implementation (e.g., through either formal or mathematical formulae), I recognized that additional support for the engineering practice could be placed at the stage of the evaluation activity. Through the conceptual shift towards trust no longer being only about the properties of either the human or robot, some space is left to consider instead the evaluation of trust in HRI as an event occurring between people and robots, which becomes the situation of trust. From this perspective, I was able to bring into focus vulnerability as an overlooked criterion, and to show through my online HRI studies that it is not only relevant in the way people experience their interactions with robots, but can also be considered relevant for robots in terms of their limited ability to protect themselves from malicious humans. Understood in relation to design practice, I believe that the knowledge I gained not only helps identify what kinds of situations are actually instances of trust (i.e., those that include the preconditions of risk, uncertainty and vulnerability), but also can be used to construct specific scenarios (i.e., the situations in which there are opportunities for being vulnerable) in which the developed behaviors of the robots can be evaluated as either fostering or hindering trust (i.e., the robot is perceived by the person interacting with it as trustworthy (or not). Because the outcome of trust in HRI is to ensure that robots are accepted by people to interact and engage with, it is very important to be very clear under which conditions we can say that the desired aim aligns with our understanding of trust in a way that is meaningful in this context. The knowledge transferred in this sense is about the extent to which we have adequate knowledge and understanding of when we can rightfully say that the development of behaviors that makes robots trustworthy are satisfactory.

With the online pilot workshop that aimed to use design thinking the knowledge transfer to the engineering practice, and with my own considerations of how my theoretical knowledge and empirical work could provide the evaluation criteria for testing whether robots are behaving in a way that is in fact trustworthy given the specific situation, my contribution to the field of HRI in terms of engineering can be associated with the stages that are less about concrete creation and implementation but rather those leading up to there, as well as in the later stages of experimentation for determining whether the solution was "successful" or not. I have gained insight into the ways interdisciplinary collaboration could potentially work between very different perspectives, aims and skill-sets within a team: even though engineering and its sister discipline of design seem to do well without much help from outside, I believe that the kind of challenges that arise in the development of robots that are perceived as trustworthy stand to benefit greatly from the insights and competences offered by people coming from other disciplines. Especially the analytic skills that many philosophers acquire and the mastery of empirical work that sociologists bring to the discussion and evaluation of proposed technological solutions is valuable to the collaborative effort of realizing trust in HRI.

7.3 Philosophers in HRI

Even though, as Colburn (2000) have observed, it might seem strange to combine philosophy and computer science, there are in fact many overlaps of interest and discussion between these two disciplines. In fact, it is hard to imagine the headway made in computer science without also diving into very deep philosophical debate – vice versa, the challenging problems being dealt with in contemporary philosophy have been raised by the advancements in computer science (Colburn, 2000). When I was given the opportunity to combine my disciplinary background in philosophy with my long-standing interest in science and technology through the interdisciplinary Trust Robots Doctoral College, I was eager to share my excitement and bring into play the ways in which philosophical perspectives and methods can lend weight to our understanding and analysis of trust in HRI. However, the challenge of clarifying to myself and others what exactly philosophers could offer in such exploration was great. This was not only because I found it hard to find my own identity as a philosopher working in the area of computer science, but also because many assumptions would be made about my approach to the topic based on people's understanding (and even stereotypical views) of what a philosopher can or will do. As such, one of the contributions of my PhD project is also to position myself with confidence in the discussion about the advantages and limitations of philosophy, and to encourage other philosophers eager to enter the field of HRI that is by now a sub-field of computer science. For this last part of the discussion, I therefore address the specific relationship between philosophy and computer science given the joint interest in AI and robotics.

7.3.1 From AI to Robotics

In my view, there are at least two different but connected waves in which philosophy and computer science have managed to engage in conversation with each other in a fruitful manner. I also believe this growing interrelation is the result of increasing interest among philosophers to extend their engagement from discussions about AI to those about about robots.

There was a first wave of heated discussion among philosophers and computer scientists about the fundamental metaphysical and epistemological challenges raised by achieving genuine artificial intelligence (AI)[6], which I refer to as the *classical debate*. Whether coming from philosophy or computer science, the significant work by Turing (1950) on the principles of the universal computing machine is without doubt the clearest example of the way in which both disciplines come into dialogue. Not only did Turing ask very philosophical questions about the possibility of computation as a way to solve problems in mathematics, he also used his skills as a mathematician with interest in computational thinking to address fundamental problems in philosophy (French, 2000). Proposed by Turing (1950) as the *imitation game* (known today as the Turing Test), he presented an operational definition of intelligence by replacing the attempt to answer the question "Can machines think?" with a pragmatic evaluation of any potential thinking machine based on how well it performs relative to the rules set by the game. As Turing concluded, a machine that would manage to pass as human because it was not possible to differentiate it from a human based on a conversation had to be considered capable of human-level intelligence. While dispute about whether or not the imitation game is a valid and valuable way of framing the fundamental goal and assessment of AI still remains, there is at least an agreement between philosophers and computer scientists about the significant impact of his work to the extent that Turing has by now earned the title of the founding father of the AI program (French, 2000). Adding to discussions about the plausibility of not only weak but also strong AI (Searle, 1980), philosophers relied on many years of insight from various philosophical sub-disciplines studying human nature to understand both the content and structure of the human mind and consciousness (e.g., philosophy of mind, phenomenology, logic, philosophy of language).

[6] Psychologists and cognitive scientists were also part of these early discussions, as their work aimed to help understand and model human cognition in areas of e.g., learning, memory, emotions, action, language, reasoning, and perception.

During this time, the philosophical method of *thought experiments* was commonly used among philosophers who aimed to demonstrate why the goal of strong AI would not suffice to achieve general human-level intelligence. Famous in this context is the "Chinese Room" thought experiment by Searle (1980) that was proposed to show how intelligence taken as the manipulation of symbols in a computational system could not in itself be proof of genuine intelligence, because it did not necessarily imply understanding[7]. In this sense, symbolic AI was narrow in scope, and the systems that were developed excelled only in rule-based knowledge domains. Targeting the different rationalist assumptions upon which the symbolic AI program relied, Dreyfus (see e.g., Dreyfus, 1992; Dreyfus and Dreyfus, 1987) was another philosopher who challenged the idea of genuine AI as a form of disembodied cognitive agent. Motivated by the discussion in computer science about the *logical frame problem*, he saw this problem as a symptom of failure to understand that the basic everyday (or common sense) knowledge applied in ordinary and ever-changing situations cannot be solved through abstract representation supported by computational models or logical inferences. Understood as a more fundamental problem of how to establish any facts given that everything is constantly changing, the *philosophical frame problem* requires that AI systems are able to differentiate relevant change from that which does not change in a factual situation (Dreyfus and Dreyfus, 1987). Even though the event calculus[8] has been accepted as a workaround to the logical variation of the frame problem, solutions are not yet in sights for how to address its philosophical variation first introduced with the proposal of so-called "Heideggerian AI" (Dreyfus, 2009). Overall, the classical debate about the feasibility of strong AI through symbolic representation was to a large extent not only related to the technical aspects of computation, but also to the many philosophical questions about the possibility of genuine human-level AI in the first place.

There has been a recent and second wave that revolves around the identification of the ethical consequences of AI pertaining to its application in robotics. While many computer scientists and philosophers are still very interested and engaged in the various metaphysical and epistemological problems in the search

[7] The argument by Searle (1980) against strong AI drew on the very important syntax/semantic (also known as form/content) distinction that brought into discussion a clearer understanding of the conceptual relations between the world, the mind and language through *reference*.

[8] Originally developed by Kowalski and Sergot (1986) and extended by the work of Shanahan (1995).

for genuine AI (brought to life by symbolic AI), the first wave of discussions made clear that computational thinking was limited. Most people will remember from discussions in the early days the arrival of the two AI winters in the late 1970's and 1980's (Kim, 2022) – the exponential growth of computer power and the development of machine learning (ML) algorithms in the 1990's, however, combined with the the availability and use of big data that began in the 2000's, kick-started new hope and increasing interest in more widespread application of AI in society. The useful applications of AI-powered tools to improve various services and activities for both public and commercial purposes is today visible in areas such as transportation, marketing, health care, finance and insurance, security and the military, science, art, education, office and industrial work, diagnosis and therapy, and personal assistance and companionship. Especially the vision of using AI-powered tools (e.g, planning, speech, face recognition, and decision-making) to support the development of socially capable robots to help assist elderly people in need of care, as well as families with daily activities in their homes, has caught the interest of many philosophers. Focusing on the potential social end ethical issues of developing, testing and introducing socially capable robots into the daily lives of people, philosophers and computer scientists are once again coming together in discussion under themes of *machine ethics* (see e.g., Anderson and Anderson (2011); Moor (2006); Tolmeijer et al. (2021); Wallach and Allen (2009)) and *AI/robot ethics* (see e.g., Coeckelbergh (2020); Gunkel (2012); Lin et al. (2012,1)). Roughly speaking, the promotion of machine ethics, which has mainly been of interest to computer scientists, focuses on how to formalize various ethical theories and principles to ensure that the behaviors of robots live up to acceptable ethical and moral standards[9]. They now must rise to the challenge of adding an ethical dimension to AI and robotics by developing artificial moral agents (Cervantes et al., 2020). In contrast, philosophers tend to take a step back in such debates, starting to question instead some of the more fundamental challenges of how to understand and integrate robots into our ethical and moral reasoning. Critical voices have been eager to discuss whether robots are in fact capable of ethical and moral reasoning in terms of fulfilling the necessary and sufficient criteria for moral agency proposed by various theories

[9] Inspired by Kantian duty-based ethics, the application of deontic logic in robotics is a good example of how computer scientists aim to formalize the logical relations between facts, obligations, prohibitions and permissions to guide normative reasoning in autonomous robots (Scheutz and Malle, 2014)

in normative ethics, while more moderate approaches have argued that ascribing moral status is problematic because it by default excludes robots as deserving any ethical consideration (Coeckelbergh, 2012). Further, concerned with the impact that the use of robots might have on the safety and well-being of more vulnerable groups in society, philosophers have also been discussing the extent to which there is a need for more regulation and legal actions to prevent any potential harm to humans, and whether it makes sense to also consider rights for robots (Mamak, 2021). In general, there is a common interest today between computer scientists and philosophers about the feasibility of developing artificial moral agents ensuring accountability when their actions have ethical and moral consequences[10].

Whether in the first or second wave of discussion, the typical role philosophers come to take is either joining forces with computer scientists by participating in the hands-on work of formalizing abstract concepts to support computational models, or conversely, taking a stance as a critical opponent who takes a step back to observe and point out the many problems in developing and using robots. I experience this general view on the philosopher working in a computer science department myself, when people assume that I am either a logician or an ethicist. Most of the time fellow researchers are surprised when I explain that I am neither, and the work I present in this dissertation is an example of what I do instead. A philosophical approach beyond logic and ethics that integrates very well with the work done in a computer science department is the new movement of experimental philosophy, which I will outline in more detail in the following section.

7.3.2 Experimental Philosophy

Continuing from the main point made in the previous sections, I consider the outcome of my PhD project as presented in this dissertation the fruitful combination of what (analytic) philosophy can offer in terms of conceptual clarity with the diligent work of a (computer) scientist exploring how people interact with robots. The aim is to broaden our understanding and analysis of trust in HRI through emphasizing the link between the concept of interpersonal trust and the notion of vulnerability. I began by relying heavily on theoretical perspectives on trust from

[10] See e.g., Malle (2016) as well as Wallach and Asaro (2017) for a more detailed account of the many differences and similarities between machine ethics and robot ethics.

philosophical discussions. Base on this, I then used various scientific methods to carry out empirical work to support my arguments for why the trust-vulnerability relation needs to be incorporated into future work in HRI. I see myself mainly as a philosopher committed to including empirical work in efforts to further the understanding and study of relevant or core concepts in HRI. In relation to more contemporary themes in analytic philosophy, I would identify my PhD project to join to some degree – in spirit – the recent movement of *experimental philosophy*.

Although experimental philosophy is a relatively new sub-discipline in philosophy that surfaced in the literature of the 2000's, there has already been much discussion about what it aims to accomplish or ought to be. Plakias (2015) explains that the first proponents of experimental philosophy tended to argue that the work that falls under this label can be characterized by requiring that (1) experimental studies are conducted (2) by philosophers (3) to test claims about the intuitions of people. This somewhat narrow view on what experimental philosophy is all about was presented in the work of Knobe and Nichols (2008), who write that "experimental philosophers proceed by conducting experimental investigations of the psychological processes underlying people's intuitions about central philosophical issues" (p. 3). Controversially, they argue that philosophers from the analytic tradition can no longer rely solely on their own experience, introspection, and intuition to investigate various philosophical problems. Instead, they have to get out of their comfortable armchairs and conduct rigorous experiments to ensure a more representative sample including ordinary people in the pursuit of possible answers (Knobe and Nichols, 2008). Needless to say, the call for experimental philosophy was met with much resistance by those who thought empirical evidence is irrelevant to philosophical thinking, and there is currently much discussion among analytic philosophers about how to actually demarcate the scope and basic commitments in experimental philosophy[11]. In a sense, my own PhD project takes part in this discussion by providing an example of how experimental philosophy could look like in practice. It is in my view important, however, to reconsider current conceptualization of what experimental is because the previous criteria can be challenged. As Plakias (2015) explain, more recent work in experimental philosophy is also done in collaboration with researchers from multiple disciplines other than cognitive sciences or psychology (e.g., so-

[11] For a more detailed overview and account of the many disagreements about experimental philosophy, the writing by e.g., Alexander (2010); Alexander and Weinberg (2007); Horvath and Grundmann (2013); Nadelhoffer and Nahmias (2007); Williamson (2016) are very helpful.

ciologists, historians, physicists, and linguists), is using very different methods beside experiments to gain empirical evidence (e.g., case study, ethnography, interview, observation, and document analysis), and range broader than studying the intuition of people (e.g., technology usage, health concerns, and political engagement). For my PhD project alone, I have collaborated with sociologists and computer scientists, have used both experimental and non-experimental methods, and have examined how we can understand and study the (interpersonal) trust notion in the context of human robot interaction.

While many analytic philosophers inform their philosophical thinking based on empirical evidence provided by scientists in both the natural and social sciences, my own motivation and commitment to carry out empirical work during my PhD project to better understand how interpersonal trust becomes meaningful and useful in the context of HRI stemmed from a more general perspective on the relationship between philosophy and the sciences. As a trained analytic philosopher, I see myself as most valuable to the HRI community in the role of quasi-scientist (as I join in on the development, implementation, and evaluation of HRI studies), as my work contributes to the fundamental knowledge needed to ensure that people want to interact, collaborate, and engage with robots. With my argument for focusing on the vulnerability-trust (conceptual) relation to guide empirical studies in HRI around the topic of trust, I basically blur the distinction between (analytic) philosophy and (computer) science. Thus, I position myself in the HRI community not in the role of a logician or ethicsist (though I still find such topics interesting to discuss), but as an experimental philosopher using my skills to provide conceptual clarity that guides both the development of precise HRI studies, and practical scenarios for the evaluation of interpersonal trust in HRI. I take my role as an experimental philosopher as the third and alternative role of philosophers in the fields of HRI and computer science. Not only is this a useful way of explaining the aim of my PhD project as presented in this dissertation, it also ensures that people outside philosophy can gain a better feeling of what philosophers can do when engaged in multi-disciplinary collaborations or projects. Philosophers can offer analytic skills useful to the clarification of very abstract concepts and the interpretation of various data, and they can apply these skills across multiple topics and fields of research. It would be wasteful to restrict philosophers in HRI and computer science to addressing only questions of logic or ethics. And still, when I look around for other experimental philosophers in HRI or computer science, I often come to find that nobody else is there. Most philosophers are happy in their role as the logician or ethicist, which of course is a fulfilling and purposeful role. But this is precisely the reason why it was necessary for me to take the path less walked, which of course also made it necessary for me to argue well for why my path is worthwhile. This has frequently occupied my thoughts during this PhD project and while writing up this dissertation. But the long and intertwined history of philosophy and the sciences (whether natural or social) proves that they can – and should – support each other in the pursuit of knowledge. As Gare (2018) argue, it is only a recent trend to separate philosophical thinking from scientific investigations, because many of the greatest philosophers also took up scientific work, and many renowned scientists were engaged in philosophical debates. I hope that my contextualization of my work in this dissertation by discussing what an experimental philosopher might enable some people to understand how my PhD project connects to, although it does not fit, the mainstream – from the perspective of philosophy and from that of HRI and computer science more broadly. I want to encourage all philosophers interested in the field of HRI or any topic within computer science that needs conceptual clarity, precise studies, and practical scenarios, to get involved as an experimental philosopher. Although it is very challenging to understand and communicate to others the advantages and disadvantages of philosophical thinking, this understanding and communication ensures that such work is more easily transferred and integrated in the knowledge pool of HRI. From the play with words used by Pradeu et al. (2021) to discuss whether philosophers can contribute substantially to scientific investigations, and also very relevant to my discussion here, I think that there is a space for philosophers as part of the HRI community to not only contribute to the philosophy of HRI, but rather to develop philosophy *in* HRI.

CHAPTER 8

Conclusion

> "It is not even the beginning of the end. But it is, perhaps, the end of the beginning."
>
> Williamson (2007, p. 292)

In many ways, reaching the end always comes with an ambivalent feeling. There is for me a joy in looking back and trying to draw a conclusion about what I have learned from all the intellectual labor put into my PhD project. Yet, it is hard to accept that it is time to leave things as they are and provide the last concluding remarks before the lights go out and I leave this stage. As I see it, it is the nature of research to be ongoing, and our attempts to share it with colleagues through writing and publication can always only capture a snap-shot of the ongoing work and discussions that bring it all to life. From before writing this dissertation until well after the documentation process was over, my thoughts and conversations with people about its content and contributions will keep me busy for some time still.

So while the conclusion for my dissertation will provide some closure for now, it is not the end of the interesting reflections and discussions that it sparked. The last comments from my side are remarks to ensure that the work can be evaluated. I will start with a section that provides short and precise answers to the overall research question and the following sub-questions. I will not only show that it is possible to learn a multitude of new things when diving into the fascinating topic of trust in HRI, but also how I achieved to contribute to the discussion by delivering a response to the open questions steering my PhD project. Afterwards, I will present some of the main limitations of my PhD project, which relate to both the theoretical perspectives and empirical work. I also provide a last remark to end this dissertation.

8.1 Answers

I will begin with my replies to the sub-questions, since these are more specific and support the overall research question. After they have been addressed, the knowledge that has been gained to this point will be satisfactory to also clearly present how this answers the main aim of this dissertation – to understand how vulnerability as a precondition for trust in HRI can and does play a role in our understanding and analysis of trust between people and robots. Intended as short replies to my sub-questions and overall research question, in a way that allows people from various disciplinary backgrounds to understand the main points of my PhD project, I will simply proceed by responding to them one by one:

(1) *How has trust been conceptualized in current HRI research?*

Conceptualizing trust as reliance has dominated (and indeed still dominates) current HRI studies on trust. *Reliance trust* stresses the predictive belief or assumption held in relation to the performance of robots when investigating whether people trust in robots or find them trustworthy. However, the design and development of robots with apparent agency has motivated a discussion about whether a conceptualization of trust as interpersonal might be needed. Taking into account whether assumptions of underlying motives or intentions play a role when people trust in robots or find them trustworthy has enabled a focus on *interpersonal trust* for studies on trust in HRI. Whether the conceptualization of trust as interpersonal can be applied directly to the context of HRI is subject to discussion because it is usually used to capture the social and ethical complexity of trust between humans. As I have argued in **chapter 2**, the interpersonal trust conceptualization requires substantial modification for application in HRI, since the identification and evaluation of trust in HRI must be situated in the interaction between robots and humans. With my proposal of an event approach to trust in HRI, I aimed to bring the conceptualization and analysis of trust towards the inclusion of its preconditions, from where I also provided a working definition of the trust-vulnerability relation.

(2) *What is the conceptual relationship between trust and vulnerability?*

Vulnerability stands in conceptual relation to trust as one of its preconditions, which I defined as *the openness to harm*. While the preconditions of risk and uncertainty are also required for the establishment of a trust situation, vulnerability is uniquely linked to the trust concept. In **chapter 3**, I argued that vulnerability describes that which people can decide to offer in situations where there is no guarantee that others will either honor or betray their openness. I used the art performance *Rest Energy* (1980) by Marina Abramović and Ulay as an example to clearly illustrate the trust-vulnerability conceptual relationship in a powerful and intuitive manner. Moreover, I have provided a working definition of trust that includes this conceptual relationship between trust and vulnerability to guide my own and future studies on trust in HRI.

(3) *What are the relevant differences between the vulnerability of humans and robots?*

Although vulnerability is one of the preconditions of trust (whether conceptualized as reliance or interpersonal), it does not carry the same meaning when applied to humans and robots: humans have a rich inner life that includes both intentional states and various emotions, they have the capacity to experience their vulnerability from a first-person perspective. This experience of vulnerability can manifest in their lives and experiences in many different ways (e.g., in relation to their body, self-identity, interpersonal relationships, and social belonging). Devoid of such inner life (as far as we can tell), robots cannot be said to be vulnerable in any equivalent manner. How the vulnerability of robots needs to be interpreted must depend solely on a third-person perspective, and have to be associated with the many ways in which they are unable to properly execute their set goal or task (e.g., given inherent deficiency or technical problems with hardware and software, issues resulting from navigating among or interacting with humans, and limitations to operation caused by environmental factors). This important distinction is visible from the requirement of applying the different methodologies and methods that I used to study either human experience of vulnerability as documented in **chapter 4** and **chapter 5** or the potential vulnerabilities of robots that I presented in **chapter 6**.

> **(4)** *How can vulnerability be studied empirically in relation to trust in HRI?*

Given that the vulnerabilities of humans and robots rest on different methodological assumptions, we decided on two separate study set-ups. For the online HRI studies aiming to explore human experience of vulnerability as presented in **chapter 4** and **chapter 5**, we first had to operationalize the notion of vulnerability, as it is abstract in nature. As such, we recognized and categorized any sign of discomfort in terms of negative emotions (e.g., disappointment, anger, sadness, fear, distress, helplessness, and frustration) as expressions of vulnerability. We then used self-report (subjective) measures through questionnaires and open-ended questions, as well as semi-structured interviews, which allowed us to capture the first-person perspective[1]. For my work presented in **chapter 6**, intended to explore the vulnerabilities of robots, examining how the notion of vulnerability could be interpreted for the case of robots was also part of the study itself. As suggested by one of the experts, thoroughly understanding the vulnerabilities of robots is crucial to the identification of all the things that could go wrong for robots in a given situation. With their extensive knowledge and experience with the various (technical) challenges that arise from developing and deploying robots, interviewing experts proved a good method to anchor the third-person perspective required to get an overview of the ways in which robots can be considered vulnerable.

> **(5)** *Assuming the design of trustworthy robots to be desirable, how can a focus on vulnerability for trust in HRI contribute to this aim?*

Because the notion of vulnerability and its conceptual relation to the concept of trust testify for the occurrence of a trust situation emerging from the interactions between humans and robots, the insights gained from this focus cannot be translated directly into design guidelines for trustworthy robots. With a colleague, I thus set out to test the method of a card-based design tool in a pilot workshop,

[1] Though a third-person perspective can also be considered using more direct measures (e.g., observation and physiologic markers), this was not possible to consider, as everything had to be conducted completely online due to the outbreak of COVID-19.

that I discussed in **chapter 7**, from which we gleaned that taking into account a more abstract level, the potential vulnerabilities of robots when attempting to develop trustworthy robots can be an alternative to a design strategy focusing only on the specific design features. The potential application in engineering practice specifically concerns mainly the *transfer* and *integration* of the fundamental and rich knowledge about trust that is available. Its conceptual relation to vulnerability to support the development process is relevant either the initial stages of problem framing and knowledge gathering, or the later stage of evaluating whether the robots developed are in fact able to take part in and process instances of trust, through scenario-based testing.

> **RQ:** *How can an investigation into the relation between trust and vulnerability advance current understanding and analysis of trust in human-robot interaction?*

To address my overall research question, I worked my way through several challenges with regards to both theoretical perspectives and empirical work, which have now accumulated to several smaller as well as greater insights. Some of the contributions of my work in this PhD project have already been accounted for when relating my results to the broader discussion on trust in HRI (see e.g., chapter 4, 5, and 6), and when replying specifically to my sub-questions above. Yet, there are two interrelated points of discussion resulting from my investigation of the trust-vulnerability relation for trust in HRI that I want to stress as important "take-home messages". Understanding vulnerability as a precondition of trust enables the inclusion of uncomfortable experiences, emotions, encounters, and dimensions of trust in HRI in a way that does not take vulnerability to be only problematic or something to explain away. On the contrary: bringing to attention and investigating in detail the vulnerabilities of humans and robots, we learn that both parties play an active role in making trust between them successful in terms of the value it has or could have. Through their interaction, collaboration, and engagement with each other, humans and robots create trust between them, depending on how capable they are in identifying, acknowledging, and responding to the vulnerability of the other party. Consequently, *trust in HRI is something contingent, situated, and emerging that is not only to be considered as a strategy to minimize risks or deal with uncertainty – it is also about how people and robots can support each other in situations where they are left vulnerable*. This more nuanced view on trust in HRI is relevant for any visions of society and the everyday life that embrace co-existence between humans and robots. As such, the insights gained from my PhD project add another argument for why the overall goal of strengthening trust in HRI is needed to secure and advance the acceptance and use of robots. Unfortunately, upon closer inspection, it turns out that things are more complicated than they seem, and that much work still needs to be done to realize the overall goal of establishing trust in HRI that carries the same value as trust between humans. In addition to our empirical work suggesting an observable relation between the trust ratings people provide in different clothes shopping scenarios and their perceived vulnerability, we also found that people were motivated to trust an imperfect robot for reasons of utility or entertainment despite experienced vulnerability from a mild privacy breach. As we discussed, the reasons for people to trust in robots when considering repeated or future interactions with them, as a response to their felt vulnerability, seems , from our follow-up study to be less related to the social and ethical concerns directly that are more commonly highlighted in the literature on trust between humans (e.g., out of goodwill or personal commitment). Accordingly, the added value that could potentially follow from focused attempts to ensure people trust in robots, or find them trustworthy, does not seem to achieve the intended and desired goal of strengthening trust in HRI . In this sense, the expected added value of utility and entertainment that we found to be the case with our focus on a subtle trust violation through a mild privacy breach, suggests that people have a lower opinion of the potential interaction benefit, which might after all not be very surprising when considering any technologies that are unable to meet their emotional, social, and ethical needs (see e.g., discussions in Sharkey and Sharkey, 2010; Sparrow and Sparrow, 2006; Turkle, 2011). As much as trust is regarded as the holy grail for understanding and explaining how society and human relationships work, we should not mistake it as serving the same purpose in the context of HRI. What will enable more widespread acceptance and use of robots in society and everyday life might in fact have little to do with trust. Unless, I would argue, the recognition of human or robot vulnerability is rooted in both social and ethical concerns, as this is needed in order to speak about interpersonal trust for HRI in a way that is similar to interactions and relationships between humans.

8.2 Future Work

What I have presented in this dissertation is only a selected and small fraction of all the work I have put into my PhD project over the last years. Many thoughts, ideas, discussion points, and potential improvements have been left out, many of which I would have gladly continued to work on if given more time and resources. My dissertation is a snapshot of how I am currently thinking about the topic of trust in HRI, and a reflection on the work I have already done to bring my own unique perspective into the current discussion. Overall, I am happy with what I have to show for myself, but like with many other things we do to learn, there is also space for improvement and more work still to be done for refinement. The following sections aims to point out the many things I would like to keep working on, or hope that others will find interesting to take up, if the project did not have to end at this time, and why I think it would be valuable to further our understanding and analysis of interpersonal trust in HRI.

Although I did my best to adapt and carry out my PhD research during the unpredictable time of the global COVID-19 outbreak, the first thing I would do is carry out my two interactive online survey studies as in-person HRI experiments. Despite showing that our development and use of the interactive online survey is a possible way to explore interpersonal trust in HRI, and that the choice of stimuli (i.e., still images vs. videos) does not make a big difference, it is necessary to compare our results to a HRI study set-up, where people get to meet the PEPPER robot either in a laboratory setting or even in a more unrestricted everyday life setting . In terms of ecological validity, it is necessary to see if we would find results with a in-person HRI study to be similar or different to those we found with the interactive online survey. Given the literature on how the embodied nature of robots tends to have an effect of how people perceive and interact with them, we expect that there will be a difference in the results. The question is: how it would they be different, and would the differences be significant? While less crucial, I would also have liked to conduct the semi-structured interviews from the follow-up interactive online study and the expert interviews in person. While digital platforms are frequently used to support online meetings, and often provide good quality in terms of audio-video recordings, much is lost in carrying out the interviews online. Not only were there sometimes technical problems that would interrupt or make it harder for me to hear what people said, but much of the intuitive feeling of how well the interview was going went missing, making it a very different experience. I believe that conducting the interviews in-person would have also contributed better to my training as an interviewer, though I also recognize that the online format also allowed me to get people to participate in my study that would otherwise not have been able to. Overall, I think that way that I and many other PhD students were forced to suddenly find ways to study HRI online is something that also challenged the field to consider to what extent these methods could have a place in future work, especially in relation to the topic of trust, which already faces a multitude of methodological challenges. As I have shown with my PhD project, the online format might be something to advance further, as it seems that the results we provided were both meaningful and gave rise to some serious discussion about how we study interpersonal trust in HRI.

Still related to the empirical work I presented in this dissertation, there are two things I would have liked to study further. First, it would have been very interesting to extend the work I did on the vulnerabilities of robots with a consideration of how the views of the experts compared to those of non-experts. With the work I already did for this PhD project, I was able to align my findings of the possible robot vulnerabilities with issues of both hardware and software, as well as the way robots handle the environment they are intended to operate in, which are already highlighted in the literature on robot failure and cybersecurity in robotics. Additionally, I found that the robots also become vulnerable during interaction with humans, as some people treat robots with bad intent. In this sense, my work added new knowledge to discussions about interpersonal trust in HRI by delivering an initial overview of how robots can be considered vulnerable that is rooted in the current HRI literature and the specialized knowledge of experts in robotics. This could lay the foundation for extending this line of thinking towards examining how lay people think about these issues. The idea would be to present the overview of robot vulnerabilities as items for people to rank and reflect upon through the use of an online survey. This survey could then be used to explore how the possible vulnerabilities of robots are considered by people intended to integrate these robots into both their professional and private life in the near future. I believe a study like this could be done fairly easily, and studying the differences in how experts in robotics and laypersons understand the various robot vulnerabilities is relevant to the study of interpersonal trust, since each group would have different reasons to then trust in robots or find them trustworthy.

Based on current literature in HRI, we can expect that non-experts would reflect very differently on the vulnerabilities of robots, because they cannot draw on the same pool of knowledge and insights as the experts. Secondly, to gain a better understanding of how exactly the conceptual relation between vulnerability and trust is influenced in the context of HRI, I would have liked to have another round of iteration on the follow-up study. As we operationalized vulnerability in our online interactive survey as instances of negative emotions (identified in the data analysis through mentions of e.g., discomfort, stress, fear), we were able to investigate empirically that trust in robots did also relate to the experience of vulnerability. Still, we also found that people who felt vulnerable also gave lower trust ratings, which might be surprising at first glance when normally considering the conceptualization of trust to be associated with something positive because it enables people to develop supportive relationships to others. Based on the current conceptual knowledge on interpersonal trust between people, would the expected prediction not be that the more vulnerable people feel, the more likely they are also to trust in robots? In my understanding and analysis of trust, the problem is that the vulnerability-trust relation is not bidirectional. While people have to be vulnerable in order for trust to be relevant, people might experience their vulnerability but not consider trust as a strategy to handle the negative emotions. I would like to explore how this problem in the conceptual relation between vulnerability and trust plays out in the HRI context. As already touched upon in the discussion of our follow-up study with the interactive online survey (see e.g., chapter 5), I suspect that when robots are put in the role of a human counterpart, the application of simply using the conceptual knowledge we have of trust between people will not lead to the same results in the context of HRI. However, this is something that would require a completely new study to be conducted. I thereby also show the limitation of what we are currently able to say about the vulnerability-trust relation at this stage. All we have done is to show that the conceptual relation is present in the HRI context. Further investigation is needed to clarify what happens with the vulnerability-trust relation in different scenarios and when using different kinds of robots, which would be valuable knowledge to the HRI community because it would shed light on whether the motivation of people to trust in robots or consider them trustworthy is in fact a response to their felt vulnerability or not.

In chapter 2, I presented my event approach to trust in HRI as a way to omit the tendency to view trust as a form of property of either the human or robot by reconsidering the metaphysical assumptions. As an alternative, I proposed to consider the interaction or exchange between humans and robots as an *event*, which allows us to take into account the preconditions of trust, because the question would now revolve around the identification of whether the interaction or exchange is something we would recognize as an instance of trust, or not. Serving merely as an a argumentative step to bringing the notion of vulnerability into our understanding and analysis of interpersonal trust in HRI, I had to leave out a more detailed account and discussion of what an event approach would consist of. However, with my longstanding interest in the metaphysical challenges prompted by the development of robots with apparent agency, I think it would be beneficial to further elaborate on my proposed event approach to trust in HRI. The reason why a more event-based view on trust in HRI is important is that it opens up questions about what kind of a phenomenon trust *is*, and what kind of implications follow in terms of the way it can be studied empirically. My suggestion of viewing trust as an event for application in the context of HRI is motivated by current debates in analytic philosophy to take processes (e.g., temporality, emergence, activity, occurrence) to be the metaphysical basis of the real and being. As such, my proposed event approach to trust in HRI is also my attempt to introduce into the HRI community some of the fundamental assumptions of process philosophy – that reality and being is fundamentally *dynamic* (Rescher, 1996, 2000; Seibt, 2017a). I think that starting our understanding and analysis of trust in HRI as something dynamic (rather than static, where the properties of humans or robots are treated as truth-makers for trust in HRI) provides a promising starting point for future work. Moreover, this underlying theoretical perspective would also align well with trends in the HRI community to also include more qualitative methods to study trust in HRI, as such approaches rest on the methodological commitment that there is an element of becoming or emergence, which requires the use of methods that can take such dynamics into account. Because the basic metaphysical challenge that my event approach to trust in HRI aims to address is of more general interest to the interdisciplinary research field of HRI, I am convinced that developing such approach is not only stimulating to the philosophers in the community, but also to everyone else exploring HRI with their various backgrounds. For this reason, I hope that researchers working in HRI on the topic of trust will be motivated to pick up my event approach, or are even already working on something similarly to advance these perspectives for future work.

8.3 Last Remark

Although my PhD project and its documentation in the form of this dissertation in a sense mark the end of my time as a student with the highest qualification, I consider this moment the beginning of my journey towards becoming an independent researcher searching for ways to gain more knowledge to push the field of HRI further. I am excited about all the things I have learned along the way, and I am more confident in using the many research skills I have acquired in realizing my PhD project. Looking back at what I have been working on since I entered the four walls of the university, my PhD project can also be seen as a continuation of a longer quest to better understand how philosophy can contribute to extending our knowledge on social robotics and HRI. With my BA project, I wanted to understand and explore why the development of social robotics is of philosophical interest, and I found the challenges of socially capable robots related not only to their descriptive but also normative dimensions. For my MA project, I decided to carry out some of the conceptual work needed to be understand in which sense socially capable robots could be considered social at all, and realized that philosophical thinking on this topic would need to be accompanied by empirical work to update our conceptual knowledge. With my PhD project, I did exactly that. I used my theoretical perspectives and methods mainly from philosophy to deepen our understanding, analysis, and study of interpersonal trust in HRI through empirical work. Identifying now as an experimental philosopher, I am ready to work my way through other questions relevant to the field of HRI, with the aim to share my passion for philosophy with the people constituting this community. I would therefore like to end this dissertation with a quote from Colburn (2000):

> *"While thinkers come and go, philosophy itself endures, even as the objects of its inquiry change from the forms of Plato, to the human understanding, to layers of abstractions in virtual worlds. Despite the inroads that empirical results in artificial intelligence may make onto philosophical turf, the universal value and appeal of philosophizing remain."* (p. 209).

In my view, Colburn summarizes very well why I think philosophy is (and always has been) a good friend of computer science, and vice versa, despite the incredible challenges they often pose to each other. This quote also helped me reflect on how this PhD project has not necessarily turned me into an computer scientist, but rather into a better philosopher. I am grateful for this realization.

Bibliography

Abate, A. F., Barra, P., Bisogni, C., Cascone, L., and Passero, I. (2020). Contextual Trust Model With a Humanoid Robot Defense for Attacks to Smart Eco-Systems. *IEEE Access*, 8:207404–207414.

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

Adler, J. E. (1994). Testimony, Trust, Knowing. *Journal of Philosophy*, 91(5):264–275.

Al-Youssel, M. (2017). Merkur setzt menschlichen Roboter „Pepper“ in Filialen ein | futurezone.at.

Alexander, J. (2010). Is experimental philosophy philosophically significant? *Philosophical Psychology*, 23(3):377–389.

Alexander, J. and Weinberg, J. M. (2007). Analytic Epistemology and Experimental Philosophy. *Philosophy Compass*, 2(1):56–80.

Allen, C., Wallach, W., and Smit, I. (2006). Why Machine Ethics? *IEEE Intelligent Systems*, 21(4):12–17.

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

Anderson, M. and Anderson, S. L. (2011). *Machine ethics*. Cambridge University Press, New York, USA.

Arkin, R. C., Borenstein, J., and Wagner, A. R. (2019). Competing Ethical Frameworks Mediated by Moral Emotions in HRI: Motivations, Background, and Approach. In *Procceedings of the International Conference on Robot Ethics and Standards.*, pages 29–30.

Aroyo, A. M., Rea, F., Sandini, G., and Sciutti, A. (2018). Trust and Social Engineering in Human Robot Interaction: Will a Robot Make You Disclose Sensitive Information, Conform to Its Recommendations or Gamble? *IEEE Robotics and Automation Letters*, 3(4):3701–3708.

Asimov, I. (1995). *The Complete Robot: The Definitive Collection of Robot Stories*. Voyager/HarperCollins, London, UK.

Atkinson, D., Hancock, P., Hoffman, R. R., Lee, J. D., Rovira, E., Stokes, C., and Wagner, A. R. (2012). Trust in Computers and Robots: The Uses and Boundaries of the Analogy to Interpersonal Trust. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 56, pages 303–307, Boston (MA), USA. SAGE Publications.

Atkinson, D. J. and Clark, M. H. (2013). Autonomous agents and human interpersonal trust: Can we engineer a human-machine social interface for trust? *AAAI Spring Symposium - Technical Report*, SS-13-07:2–7.

Babin, B. J., Darden, W. R., and Griffin, M. (1994). Work and/or Fun: Measuring Hedonic and Utilitarian Shopping Value. *Journal of Consumer Research*, 20(4):644.

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

Bartneck, C., Belpaeme, T., Eyssel, F., Kanda, T., Keijsers, M., and Šabanović, S. (2020). *Human-robot interaction : an introduction*. Cambridge University Press, Cambridge, UK.

Beck, U. (1992). *Risk Society : Towards a New Modernity*. Sage Publications, London, UK.

Belpaeme, T. (2020). Advice to New Human-Robot Interaction Researchers. In Jost, C., Pévédic, B. L., Belpaeme, T., Bethel, C. L., Chrysostomou, D., Crook, N., Grandgeorge, M., and Mirning, N., editors, *Human-Robot Interaction: Evaluation Methods and Their Standardization*, chapter 14, pages 355–369. Springer, Cham, Cham, Switzerland.

Bennett, J. (1988). *Events and Their Names*, volume 28. Clarendon Press, Oxford, UK.

Bennett, K. (2017). Conceptual Analysis and its Limits. *Philosophic Exchange*, 46(1):1–12.

Bertel, L. B. and Hannibal, G. (2015). The NAO robot as a Persuasive Educational and Entertainment Robot (PEER) – a case study on children's articulation, categorization and interaction with a social robot for learning. *Læring og Medier (LOM)*, 8(14).

Bethel, C. L. and Murphy, R. R. (2010). Review of Human Studies Methods in HRI and Recommendations. *International Journal of Social Robotics*, 2(4):347–359.

Biernacki, P. and Waldorf, D. (1981). Snowball Sampling: Problems and Techniques of Chain Referral Sampling. *Sociological Methods & Research*, 10(2):141–163.

Bishop, L., van Maris, A., Dogramadzi, S., and Zook, N. (2019). Social robots: The influence of human and robot characteristics on acceptance. *Paladyn, Journal of Behavioral Robotics*, 10(1):346–358.

Bogner, A., Littig, B., and Menz, W. (2009). Introduction: Expert Interviews – An Introduction to a New Methodological Debate. In Alexander Bogner, Beate Littig, and Wolfgang Menz, editors, *Interviewing Experts*, page 281. Palgrave Macmillan, Basingstoke, England.

Bogner, A. and Menz, W. (2009). The Theory-Generating Expert Interview: Epistemological Interest, Forms of Knowledge, Interaction. In Bogner, A., Littig, B., and Menz, W., editors, *Interviewing Experts*, chapter 2, page 281. Palgrave Macmillan.

Booth, S., Tompkin, J., Pfister, H., Waldo, J., Gajos, K., and Nagpal, R. (2017). Piggybacking Robots: Human-Robot Overtrust in University Dormitory Security. In *Proceedings of the 12th ACM/IEEE International Conference on Human-Robot Interaction*, pages 426–434, Vienna, Austria. ACM Press.

Bostrom, N. (2005). Transhumanist Values. *Journal of Philosophical Research*, 30(9999):3–14.

Braun, V., Clarke, V., Hayfield, N., and Terry, G. (2019). Thematic Analysis. In Liamputtong, P., editor, *Handbook of Research Methods in Health Social Sciences*, chapter 48, pages 843–860. Springer Nature Singapore Pte Ltd., Singapore.

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

Breazeal, C. (2003). Toward Sociable Robots. *Robotics and Autonomous Systems*, 42:167–175.

Breazeal, C. and Scassellati, B. (2002). How to build robots that make friends and influence people. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) - Human and Environment Friendly Robots with High Intelligence and Emotional Quotients (Cat. No.99CH36289)*, volume 2, pages 858–863, Kyongju, South Korea. IEEE.

Bridgwater, T., Giuliani, M., van Maris, A., Baker, G., Winfield, A., and Pipe, T. (2020).

Examining Profiles for Robotic Risk Assessment. In *15th ACM/IEEE International Conference on Human-Robot Interaction*, pages 23–31, Cambridge, UK. ACM.

Brooks, A. G., Gray, J., and Hoffman, G. (2004). Robot's play: interactive games with sociable machines. In *Proceedings of the 2004 ACM SIGCHI International Conference on Advances in computer entertainment technology - ACE '04*, pages 74–83, New York, NY. ACM Press.

Brooks, R. A. (1999). *CAMBRIAN INTELLIGENCE: the early history of the new AI*. MIT Press, Cambridge, MA.

Brown, B. (2012). *Daring greatly: How the courage to be vulnerable transforms the way we live, love, parent, and lead*. Gotham Books, New York.

Brscić, D., Kidokoro, H., Suehiro, Y., and Kanda, T. (2015). Escaping from Children's Abuse of Social Robots. In *Proceedings of the 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI'15)*, pages 59–66, Portland (OR), USA. ACM.

Bruckenberger, U., Weiss, A., Mirnig, N., Strasser, E., Stadler, S., and Tscheligi, M. (2013). The Good, The Bad, The Weird: Audience Evaluation of a "Real" Robot in Relation to Science Fiction and Mass Media. In *International Conference on Social Robotics*, pages 301–310, Bristol, UK. Springer, Cham.

Bryant, D., Borenstein, J., and Howard, A. (2020). Why Should We Gender? In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, pages 13–21, Cambridge, UK. ACM.

Brynjolfsson, E. and McAfee, A. (2011). *Race Against the Machine: How the Digital Revolution is Accelerating Innovation, Driving Productivity, and Irreversibly Transforming Employment and the Economy*. Digital Frontier Press, Lexington, MA.

Butler, J. (2006). *Precarious Life: The Powers of Mourning and Violence*. Verso, London, UK.

Cai, H. and Lin, Y. (2010). Tuning Trust Using Cognitive Cues for Better Human-Machine Collaboration. In *Proceeding of the 54th Human Factors and Ergonomics Society annual meeting (HFES 2010)*, volume 54, pages 2437–2441, San Francisco, CA. SAGE Publications.

Calhoun, C. S., Bobko, P., Gallimore, J. J., and Lyons, J. B. (2019). Linking precursors of interpersonal trust to human-automation trust: An expanded typology and exploratory experiment. *Journal of Trust Research*, 9(1):28–46.

Cameron, D., Aitken, J. M., Collins, E. C., Boorman, L., Chua, A., Fernando, S., McAree, O., Martinez-Hernandez, U., and Law, J. (2015). Framing Factors: The Importance of Context and the Individual in Understanding Trust in Human-Robot Interaction. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1–6, Hamburg, Germany. White Rose.

Campbell, D. (2018). HSBC bank branch on Fifth Avenue has Pepper the robot - Business Insider Deutschland.

Čapek, K. (2004). *R.U.R. (Rossum's Universal Robots)*. Penguin Publishing Group, New York, NY, trans. cla edition.

Carrol, J. (1999). Five reasons for scenario-based design. In *Proceedings of the 32nd Annual Hawaii International Conference on Systems Sciences.*, page pp. 11, Maui, HI. IEEE.

Castelfranchi, C. and Falcone, R. (2010). *Trust Theory - A socio-cognitiveand and Computational Model*, volume 3. John Wiley & Sons Ltd. (Wiley Series in Agent Technology), Chichester.

Cervantes, J.-A., López, S., Rodríguez, L.-F., Cervantes, S., Cervantes, F., and Ramos, F. (2020). Artificial Moral Agents: A Survey of the Current Status. *Science and Engineering Ethics*, 26(2):501–532.

Cha, E., Dragan, A. D., and Srinivasa, S. S. (2015). Perceived Robot Capability. In *24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 541–548, Kobe, Japan. IEEE.

Chatzimichali, A. and Chrysostomou, D. (2019). Human-data interaction and user rights at the personal robot era. In *4th International Conference on Robot Ethics and Standards*, pages 117–124, London, UK. CLAWAR Association Ltd.

Chen, C. C., Saparito, P., and Belkin, L. (2011). Responding to trust breaches: The domain specificity of trust and the role of affect. *Journal of Trust Research*, 1(1):85–106.

Chen, M., Nikolaidis, S., Soh, H., Hsu, D., and Srinivasa, S. (2018). Planning with Trust for Human-Robot Collaboration. In *Proceedings of the 13th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 307–315, Chicago, USA. ACM.

Christoforakos, L., Gallucci, A., Surmava-Große, T., Ullrich, D., and Diefenbach, S. (2021). Can Robots Earn Our Trust the Same Way Humans Do? A Systematic Exploration of Competence, Warmth, and Anthropomorphism as Determinants of Trust Development in HRI. *Frontiers in robotics and AI*, 8(640444):1–15.

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

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

Clarke, V., Braun, V., and Hayfield, N. (2015). Thematic analysis. In Smith, J. A., editor, *Qualitative psychology: A practical guide to research methods*, chapter 10, pages 222–248. SAGE Publications Ltd., London, UK, 3 edition.

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

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

Coeckelbergh, M. (2011). Humans, animals, and robots: A phenomenological approach to human-robot relations. *International Journal of Social Robotics*, 3(2):197–204.

Coeckelbergh, M. (2012). *Growing Moral Relations - Critique of Moral Status Ascription*. Palgrave Macmillan, Hampshire, UK.

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

Coeckelbergh, M. (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):339–348.

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

Coeckelbergh, M. (2020). *AI Ethics*. MIT Press, Cambridge, MA.

Colburn, T. (2000). *Philosophy and Computer Science*. Routledge, New York, NY.

Coleman, J. S. (1990). *Foundations of social theory*. Belknap Press of Harvard University Press.

Correia, F., Alves-Oliveira, P., Maia, N., Ribeiro, T., Petisca, S., Melo, F. S., and Paiva, A. (2016). Just follow the suit! Trust in human-robot interactions during card game playing. In *25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 507–512, New York, USA. IEEE.

Correia, F., Guerra, C., Mascarenhas, S., Melo, F. S., and Paiva, A. (2018). Exploring the Impact of Fault Justification in Human-Robot Trust. In *Proc. of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, page 7, Stockholm, Sweden. ACM.

Creswell, J. W. and Creswell, J. D. (2018). *Research Design – Qualitative, Quantitative, and Mixed Methods Approaches*. SAGE Publications, Inc., Thousand Oaks, CA, 5th edition.

Dagan, E., Márquez Segura, E., Altarriba Bertran, F., Flores, M., and Isbister, K. (2019). Designing 'True Colors': A So-cial Wearable that Affords Vulnerability. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–14, Glasgow, Scotland. ACM.

Danaher, J. (2019). The Philosophical Case for Robot Friendship. *Journal of Posthuman Studies*, 3(1):5–24.

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

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

Dautenhahn, K. (2007). Socially Intelligent Robots: Dimensions of Human-Robot Interaction. *Philosophical transactions of the Royal Society of London (Series B, Biological sciences)*, 362(1480):679–704.

Dautenhahn, K., Nehaniv, C. L., Walters, M. L., Robins, B., Kose-Bagci, H., Mirza, N. A., Blow, M., and Blow, M. (2009). KASPAR – a minimally expressive humanoid robot for human–robot interaction research. *Applied Bionics and Biomechanics*, 6(3-4):369–397.

de Graaf, M. M. A. (2016). An Ethical Evaluation of Human–Robot Relationships. *International Journal of Social Robotics*, 8(4):589–598.

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

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

Deligianis, C., Stanton, C. J., McGarty, C., and Stevens, C. J. (2017). The Impact of Intergroup Bias on Trust and Approach Behaviour Towards a Humanoid Robot. *Journal of Human-Robot Interaction*, 6(3):4.

Deng, E., Mutlu, B., and Mataric, M. J. (2019). Embodiment in Socially Interactive Robots. *Foundations and Trends® in Robotics*, 7(4):251–356.

Dennett, D. C. (1994). The practical requirements for making a conscious robot. *Philosophical Transactions of the Royal. Series A: Physical and Engineering Sciences*, 349(1689):133–146.

Desai, M., Kaniarasu, P., Medvedev, M., Steinfeld, A., and Yanco, H. (2013). Impact of robot failures and feedback on real-time trust. In *8th ACM/IEEE International Conference on Human-Robot Interaction*, pages 251–258, Tokyo, Japan. IEEE.

Desai, M., Medvedev, M., Vázquez, M., McSheehy, S., Gadea-Omelchenko, S., Bruggeman, C., Steinfeld, A., and Yanco, H. (2012). Effects of changing reliability on trust of robot systems. In *7th ACM/IEEE international conference on Human-Robot Interaction*, page 73, Boston (MA), USA. ACM Press.

Desai, M., Stubbs, K., Steinfeld, A., and Yanco, H. (2009). Creating Trustworthy Robots: Lessons and Inspirations from Automated Systems. In *Annual Convention of The Society for the Study of Artificial Intelligence and Simulation for Behaviour*, Edinburgh, UK.

Deutsch, M. (1960). The Effect of Motivational Orientation upon Trust and Suspicion. *Human Relations*, 13(2):123–139.

Dobrosovestnova, A. and Hannibal, G. (2020a). Teachers' Disappointment: Theoretical Perspective on the Inclusion of Ambivalent Emotions in Human-Robot Interactions in Education. In *Proceedings of the 215th ACM/IEEE International Conference on Human-Robot Interaction*, pages 471–480, Cambridge, UK. ACM.

Dobrosovestnova, A. and Hannibal, G. (2020b). Working Alongside Service Robots: Challenges to Workplace Identity Performance. In *Proceedings of Robophilosophy 2020 – Culturally Sustainable Social Robotics*, pages 148–157, Aarhus, Denmark. IOS Press.

Dobrosovestnova, A., Hannibal, G., and Reinboth, T. (2022). Service robots for affective labor: a sociology of labor perspective. *AI & SOCIETY*, 37(2):487–499.

Domenicucci, J. and Holton, R. (2016). Trust as a Two-Place Relation. In Faulkner, P. and Simpson, T., editors, *The Philosopphy of Trust*, chapter 9, pages 149–160. Oxford University Press.

Dreyfus, H. (1992). *What Computers Still Can't Do : a critique of artificial reason*. MIT Press, New York, USA, 3rd edition.

Dreyfus, H. and Dreyfus, S. (1987). How to stop worrying about the frame problem even though it's computationally insoluble. In Pylyshyn, Z. W., editor, *The Robot's Dilemma: The Frame Problem in Artificial Intelligence*, chapter 5, pages 95–112. ABLEX Publishing Corporation, Norwood, NJ.

Dreyfus, H. L. (2009). How Representational Cognitivism Failed and is being replaced by Body/World Coupling. In *After Cognitivism*, pages 39–73. Springer Netherlands, Dordrecht.

Duffy, B. R. (2006). Fundamental Issues in Social Robotics. *The International Review of Information Ethics*, 6:31–36.

Dyson, F. J. (2015). Birds and Frogs. In *Birds and Frogs: Selected Papers of Freeman Dyson, 1990–2014*, chapter 2.3, pages 37–57. World Scientific, Singapore, Republic of Singapore.

Edmonds, M., Gao, F., Liu, H., Xie, X., Qi, S., Rothrock, B., Zhu, Y., Wu, Y. N., Lu, H., and Zhu, S.-C. (2019). A tale of two explanations: Enhancing human trust by explaining robot behavior. *Science Robotics*, 4(37):1–13.

Ellul, J. (1980). *The Technological System*. Continuum, New York, NY.

Fast-Berglund, Å. and Romero, D. (2019). Strategies for Implementing Collaborative Robot Applications for the Operator 4.0. In *Advances in Production Management Systems. Production Management for the Factory of the Future*, pages 682–689, Austin, TX. Springer, Cham.

Faulkner, P. (2007). On Telling and Trusting. *Mind*, 116(464):875–902.

Faulkner, P. and Simpson, T. (2017). *The Philosophy of Trust*. Oxford University Press, Oxford, UK.

Feil-Seifer, D., Haring, K. S., Rossi, S., Wagner, A. R., and Williams, T. (2021). Where to Next? The Impact of COVID-19 on Human-Robot Interaction Research. *ACM Transactions on Human-Robot Interaction*, 10(1):1–7.

Fineman, M. A. (2008). The Vulnerable Subject: Anchoring Equality in the Human

Condition. *Yale Journal of Law & Feminism*, 20(1):8–40.

Fischer, K., Weigelin, H. M., and Bodenhagen, L. (2018). Increasing trust in human–robot medical interactions: effects of transparency and adaptability. *Paladyn, Journal of Behavioral Robotics*, 9(1):95–109.

Flick, U. (2009). *An Introduction to Qualitative Research*. SAGE Publications Ltd., London, 4th edition.

Flook, R., Shrinah, A., Wijnen, L., Eder, K., Melhuish, C., and Lemaignan, S. (2019). On the impact of different types of errors on trust in human-robot interaction. *Interaction Studies. Social Behaviour and Communication in Biological and Artificial Systems*, 20(3):455–486.

Floyd, M. W., Drinkwater, M., and Aha, D. W. (2015). Trust-Guided Behavior Adaptation Using Case-Based Reasoning. In *24th International Conference on Artificial Intelligence*, page 4261–4267, Buenos Aires, Argentina. AAAI Press.

Fooladi Mahani, M., Jiang, L., and Wang, Y. (2020). A Bayesian Trust Inference Model for Human-Multi-Robot Teams. *International Journal of Social Robotics*, pages 1–15.

Fosch-Villaronga, E., Felzmann, H., Pierce, R. L., de Conca, S., de Groot, A., del Castillo, A. P., and Robbins, S. (2018). 'Nothing Comes between My Robot and Me': Privacy and Human-Robot Interaction in Robotised Healthcare. In Leenes, R., van Brakel, R., Gutwirth, S., and De Hert, P., editors, *Data Protection and Privacy: The Internet of Bodies*, chapter 4, pages 93–123. Hart Publishing, Oxford, UK.

Fosch-Villaronga, E. and Mahler, T. (2021). Cybersecurity, safety and robots: Strengthening the link between cybersecurity and safety in the context of care robots. *Computer Law & Security Review*, 41:105528.

Fraune, M. R., Leite, I., Karatas, N., Amirova, A., Legeleux, A., Sandygulova, A., Neerincx, A., Dilip Tikas, G., Gunes, H., Mohan, M., Abbasi, N. I., Shenoy, S., Scassellati, B., de Visser, E. J., and Komatsu, T. (2022). Lessons Learned About Designing and Conducting Studies From HRI Experts. *Frontiers in Robotics and AI*, 8:401.

Freedy, A., DeVisser, E., Weltman, G., and Coeyman, N. (2007). Measurement of trust in human-robot collaboration. In *International Symposium on Collaborative Technologies and Systems*, pages 106–114, Orlando. IEEE.

French, R. M. (2000). The Turing Test: the first 50 years. *Trends in Cognitive Sciences*, 4(3):115–122.

Frost-Arnold, K. (2014). The cognitive attitude of rational trust. *Synthese*, 191:1957–1974.

Gambetta, D. (1988). Can We Trust Trust? In Gambetta, D., editor, *Trust: Making and Breaking Cooperative Relations*, chapter 13, pages 213–237. Basil Blackwell, Oxford, UK.

Gare, A. (2018). Natural Philosophy and the Sciences: Challenging Science's Tunnel Vision. *Philosophies*, 3(4):33.

Gasparetto, A. and Scalera, L. (2019). A Brief History of Industrial Robotics in the 20th Century. *Advances in Historical Studies*, 8(1):24–35.

Giddens, A. (1990). *The Consequences of Modernity*. Stanford University Press, Stanford, CA, USA.

Glaser, B. G. and Strauss, A. L. (1967). *The discovery of grounded the- ory: Strategies for qualitative research*. Aldine Publishing Company, Chicago, IL.

Glock, H. J. (2008). *What is analytic philosophy?* Cambridge University Press, Cambridge, UK.

Gompei, T. and Umemuro, H. (2015). A robot's slip of the tongue: Effect of speech error on the familiarity of a humanoid robot. In *24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 331–336, Kobe, Japan. IEEE.

Goudey, A. and Bonnin, G. (2016). Must smart objects look human? Study of the impact of anthropomorphism on the acceptance of companion robots. *Recherche et Applications en Marketing (English Edition)*, 31(2):2–20.

Grodzinsky, F. S., Miller, K. W., and Wolf, M. J. (2015). Developing Automated Deceptions and the Impact on Trust. *Philosophy & Technology*, 28(1):91–105.

Gunkel, D. J. (2012). *The Machine Question: Critical Perspectives on AI, Robots, and Ethics*. MIT Press, Cambridge, USA.

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

Guo, Y., Zhang, C., and Yang, X. J. (2020). Modeling Trust Dynamics in Human-robot Teaming: A Bayesian Inference Approach. In *ACM CHI Conference on Human Factors in Computing Systems*, pages 1–7, Honolulu (HI), USA. ACM.

Hamacher, A., Bianchi-Berthouze, N., Pipe, A. G., and Eder, K. (2016). Believing in BERT: Using expressive communication to enhance trust and counteract operational error in physical Human-robot interaction. In *25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 493–500, New York (NY), USA. IEEE.

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

Hancock, P. A., Kessler, T. T., Kaplan, A. D., Brill, J. C., and Szalma, J. L. (2020). Evolving Trust in Robots: Specification Through Sequential and Comparative Meta-Analyses. *Human factors*, pages 1–34.

Handcock, M. S. and Gile, K. J. (2011). Comment: On the concept of snowball sampling. *Sociological Methodology*, 41(1):367–371.

Hannibal, G. (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*, pages 288–293, Boulder, CO. ACM.

Hannibal, G., Dobrosovestnova, A., and Weiss, A. (2022a). Tolerating Untrustworthy Robots: Studying Human Vulnerability Experience within a Privacy Scenario for Trust in Robots. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 821–828. IEEE.

Hannibal, G., Rabb, N., Law, T., and Alves-Oliveira, P. (2022b). Towards a Common Understanding and Vision for Theory-Grounded Human-Robot Interaction (THEORIA). In *Proceedings of the 17th ACM/IEEE International Conference on Human-Robot Interaction*, page 1254–1257, Sapporo, Japan. IEEE.

Hannibal, G., Weiss, A., and Charisi, V. (2021). "The robot may not notice my discomfort"- Examining the experience of vulnerability for trust in human-robot interaction. In *2021 30th IEEE International Conference on Robot and Human Interactive Communication, RO-MAN 2021*, pages 704–711, Vancouver, Canada. IEEE.

Hardin, R. (1992). The Street-Level Epistemology of Trust. *Analyse & Kritik*, 12(2):152–176.

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

Haring, K. S., Matsumoto, Y., and Watanabe, K. (2013). How Do People Perceive and Trust a Lifelike Robot. In *Proceedings of the World Congress on Engi-*

neering and Computer Science (WCECS), pages 23–25, San Francisco, USA. IAENG.

Haring, K. S., Silvera-Tawil, D., Matsumoto, Y., Velonaki, M., and Watanabe, K. (2014). Perception of an Android Robot in Japan and Australia: A Cross-Cultural Comparison. In Beetz, M., Johnston, B., and Williams, M.-A., editors, *International Conference on Social Robotics (ICSR). Lecture Notes in Computer Science book series, Vol. 8755*, pages 166–175, Sydney, Australia. Springer International Publishing.

Hauskeller, M. (2019). Ephemeroi - Human Vulnerability, Transhumanism, and the Meaning of Life. *Scientia et Fides*, 7(2):9–21.

Hawkins, J. (2021). *A THOUSAND BRAINS : A new theory of intelligence*. Basic Books, New York, NY.

Hawley, K. (2014). Trust, Distrust and Commitment. *Source: Noûs*, 48(1):1–20.

Hawley, K. (2019). *How to Be Trustworthy*. Oxford University Press, Oxford, UK.

Henschel, A., Laban, G., and Cross, E. S. (2021). What Makes a Robot Social? A Review of Social Robots from Science Fiction to a Home or Hospital Near You. *Current Robotics Reports*, 2(9):1–19.

Hesse-Biber, S. N. (2010). *Mixed methods research: Merging theory with practice*. Guilford Press, New York.

Hieronymi, P. (2008). The reasons of trust. *Australasian Journal of Philosophy*, 86(2):213–236.

Hirose, M. and Ogawa, K. (2007). Honda humanoid robots development. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 365(1850):11–19.

Hoffman, R. R., Johnson, M., Bradshaw, J. M., and Underbrink, A. (2013). Trust in Automation. *IEEE Intelligent Systems*, 28(1):84–88.

Holton, R. (1994). Deciding to Trust, Coming to Believe. *Australasian Journal of Philosophy*, 72(1):63–76.

Honig, S. and Oron-Gilad, T. (2018). Understanding and Resolving Failures in Human-Robot Interaction: Literature Review and Model Development. *Frontiers in Psychology*, 9(861).

Horkheimer, M. and Adorno, T. W. (1972). *Dialectic of Enlightenment*. Seabury Press, New York, NY, trans. by edition.

Horvath, J. and Grundmann, T., editors (2013). *Experimental Philosophy and its Critics*. Routledge, London, UK, 1 edition.

Howard, A. and Borenstein, J. (2018). The Ugly Truth About Ourselves and Our Robot Creations: The Problem of Bias and Social Inequity. *Science and Engineering Ethics*, 24(5):1521–1536.

Huang, S. H., Held, D., Abbeel, P., and Dragan, A. D. (2019). Enabling robots to communicate their objectives. *Autonomous Robots*, 43(2):309–326.

Ihde, D. (1990). *Technology and the Lifeworld: From Garden to Earth*. Indiana University Press, Bloomington, IN.

Joinson, A., Reips, U.-D., Buchanan, T., and Schofield, C. B. P. (2010). Privacy, Trust, and Self-Disclosure Online. *Human-Computer Interaction*, 25(1):1–24.

Jones, K. (1996). Trust as an Affective Attitude. *Ethics*, 107(1):4–25.

Jones, K. (2004). Trust and Terror. In Desautels, P. and Walker, M., editors, *Moral Psychology; feminist ethics and social theory*, chapter 1, pages 3–18. Rowman & Littlefield Publishers, Oxford, UK.

Jung, M., Lazaro, M. J. S., and Yun, M. H. (2021). Evaluation of Methodologies and Measures on the Usability of Social Robots: A Systematic Review. *Applied Sciences*, 11(4):1388.

Kahn, P., Reichert, A., Gary, H., Kanda, T., Ishiguro, H., Shen, S., Ruckert, J., and Gill, B. (2011). The new ontological category hypothesis in human-robot interaction. In *Proceedings of the 6th ACM/IEEE International Conference on Human-Robot Interaction*, pages 159–160, Lausanne, Switzerland. IEEE.

Kahn, P. H., Kanda, T., Ishiguro, H., Freier, N. G., Severson, R. L., Gill, B. T., Ruckert, J. H., and Shen, S. (2012). "Robovie, you'll have to go into the closet now": Children's social and moral relationships with a humanoid robot. *Developmental Psychology*, 48(2):303–314.

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

Kaplan, A., Kessler, T., Sanders, T., Cruit, J., Brill, J., and Hancock, P. (2021). A time to trust: Trust as a function of time in human-robot interaction. In Nam, C. S. and Lyons, J. B., editors, *Trust in Human-Robot Interaction*, pages 143–157. Academic Press, London, UK.

Keren, A. (2014). Trust and belief: a preemptive reasons account. *Synthese*, 191(12):2593–2615.

Keren, A. (2020). Trust and Belief. In *The Routledge Handbook of Trust and Philosophy*, pages 109–120. Routledge.

Kessler, T. T., Larios, C., Walker, T., Yerdon, V., and Hancock, P. A. (2017). A Comparison of Trust Measures in Human–Robot Interaction Scenarios. In Savage-Knepshield, P. and Chen, J., editors, *Advances in Human Factors in Robots and Unmanned Systems (Advances in Intelligent Systems and Computing book series, Vol. 499)*, pages 353–364. Springer, Cham.

Khalid, H., Liew, W. S., Voong, B. S., and Helander, M. (2019). Creativity in Measuring Trust in Human-Robot Interaction Using Interactive Dialogs. In *Proceedings of the 20th Congress of the International Ergonomics Association (IEA'18)*, pages 1175–1190, Florence, Italy. Springer, Cham.

Kim, H. (2022). Historical Sketch of Artificial Intelligence. In *Artificial Intelligence for 6G*, chapter 1, pages 3–14. Springer International Publishing, Cham, Switzerland.

Kirschgens, L. A., Ugarte, I. Z., Uriarte, G., Mu~, A. M., Rosas, M., and Vilches, M. (2019). ROBOT HAZARDS: From safety to security. Technical report, Alias Robotics S.L.

Kirton, A. (2020). Matters of Trust as Matters of Attachment Security. *International Journal of Philosophical Studies*, 28(5):583–602.

Knobe, J. and Nichols, S. (2008). An Experimental Philosophy Manifesto. In Knobe, J. and Nichols, S., editors, *Experimental Philosophy*, chapter 1, pages 3–14. Oxford University Press, New York, NY.

Kok, B. C. and Soh, H. (2020). Trust in Robots: Challenges and Opportunities. *Current Robotics Reports*, 1(4):297–309.

Kokolakis, S. (2017). Privacy attitudes and privacy behaviour: A review of current research on the privacy paradox phenomenon. *Computers & Security*, 64:122–134.

Komatsu, T. (2016). Japanese students apply same moral norms to humans and robot agents: Considering a moral HRI in terms of different cultural and academic backgrounds. In *Proceedings of the 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 457–458, Christchurch, New Zealand. ACM.

Komatsu, T., Malle, B. F., and Scheutz, M. (2021). Blaming the Reluctant Robot. In *Proceedings of the 16th ACM/IEEE International Conference on Human-Robot Interaction*, pages 63–72, Boulder. ACM.

Kosterec, M. (2016). Methods of conceptual analysis. *Filozofia*, 71(3):220–230.

Kowalski, R. and Sergot, M. (1986). A Logic-Based Calculus of Events. *New*

Generation Computing, 4(1):67–95.

Krupp, M. M., Rueben, M., Grimm, C. M., and Smart, W. D. (2017). A focus group study of privacy concerns about telepresence robots. In *26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 1451–1458, Lisbon, Portugal. IEEE.

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

Langer, A., Feingold-Polak, R., Mueller, O., Kellmeyer, P., and Levy-Tzedek, S. (2019). Trust in socially assistive robots: Considerations for use in rehabilitation. *Neuroscience & Biobehavioral Reviews*, 104:231–239.

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

Law, T. and Scheutz, M. (2021). Trust: Recent concepts and evaluations in human-robot interaction. In Nam, C. S. and Lyons, J. B., editors, *Trust in Human-Robot Interaction: Research and Application*, chapter 2, pages 27–57. Academic Press, London, UK.

Lee, H. R., Cheon, E., Lim, C., and Fischer, K. (2022). Configuring Humans: What Roles Humans Play in HRI Research. In *17th ACM/IEEE International Conference on Human-Robot Interaction*, page 478–492, Sapporo, Japan. IEEE Press.

Lee, J. D. and See, K. A. (2004). Trust in Automation: Designing for Appropriate Reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 46(1):50–80.

Lee, J.-E. R. and Nass, C. I. (2010). Trust in Computers: The Computers-Are-Social-Actors (CASA) Paradigm and Trustworthiness Perception in Human-Computer Communication. In Latusek, D. and Gerbasi, A., editors, *Trust and Technology in a Ubiquitous Modern Environment: Theoretical and Methodological Perspectives*, chapter 1, pages 1–15. Information Science Reference (an imprint of IGI Gobal), Hershey, PA: United States of America.

Lee, J. J., Knox, W. B., Wormwood, J. B., Breazeal, C., and DeSteno, D. (2013). Computationally modeling interpersonal trust. *Frontiers in Psychology*, 4:893.

Lee, M., Ruijten, P., Frank, L., de Kort, Y., and IJsselsteijn, W. (2021). People May

Punish, But Not Blame Robots. In *Proceedings of the Conference on Human Factors in Computing Systems*, pages 1–11, Yokohama, Japan. ACM.

Lee, M. K., Tang, K. P., Forlizzi, J., and Kiesler, S. (2011). Understanding users' perception of privacy in human-robot interaction. In *Proceedings of the 6th international conference on Human-robot interaction*, page 181, Lausanne, Switzerland. ACM Press.

Leichtmann, B., Nitsch, V., and Mara, M. (2022). Crisis Ahead? Why Human-Robot Interaction User Studies May Have Replicability Problems and Directions for Improvement. *Frontiers in Robotics and AI*, 9:60.

Lewis, M., Sycara, K., and Walker, P. (2018). The Role of Trust in Human-Robot Interaction. In Abbass, H. A., Reid, D. J., and Scholz, J., editors, *Foundations of Trusted Autonomy (Studies in Systems, Decision and Control, Vol. 117)*, volume 177, chapter 8, pages 135–159. Springer Open, Cham, Switzerland.

Liedo, B. and Rueda, J. (2021). In Defense of Posthuman Vulnerability. *Scientia et Fides*, 9(1):215–239.

Lim, V., Rooksby, M., and Cross, E. S. (2021). Social Robots on a Global Stage: Establishing a Role for Culture During Human–Robot Interaction. *International Journal of Social Robotics*, 13(6):1307–1333.

Lin, P., Abney, K., and Bekey, G. A. (2012). *Robot ethics : the ethical and social implications of robotics*. MIT Press, Cambridge, USA.

Lin, P., Abney, K., and Jenkins, R. (2017). *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*. Oxford University Press, New York, USA.

Lindblom, J. and Wang, W. (2018). Towards an Evaluation Framework of Safety, Trust, and Operator Experience in Different Demonstrators of Human-Robot Collaboration. In *Proceeding of the 16th International Conference on Manufacturing Research (incorporating the 33rd National Conference on Manufacturing Research)*, pages 145–150, Skövde, Sweden. IOS Press.

Lindner, F. and Bentzen, M. M. (2017). The Hybrid Ethical Reasoning Agent IMMANUEL. In *Proceedings of the 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 187–188, Vienna. ACM.

Livnat, Y. (2004). On the Nature of Benevolence. *Journal of Social Philosophy*, 35(2):304–317.

Loux, M. J. and Crisp, T. M. (2017). *Metaphysics: A contemporary introduction*. Routledge, New York (NY), USA, 4 edition.

Luhmann, N. (1979). *Trust and Power*. Polity Press, Cambridge, UK, 2017 edition.

Lutz, C., Schöttler, M., and Hoffmann, C. P. (2019). The privacy implications of social robots: Scoping review and expert interviews. *Mobile Media & Communication*, 7(3):412–434.

Lutz, C. and Tamó-Larrieux, A. (2020). The Robot Privacy Paradox: Understanding How Privacy Concerns Shape Intentions to Use Social Robots. *Human-Machine Communication*, 1:87–111.

Lutz, C. and Tamò-Larrieux, A. (2021). Do Privacy Concerns About Social Robots Affect Use Intentions? Evidence From an Experimental Vignette Study. *Frontiers in Robotics and AI*, 8:63.

Lyons, J. B. and Havig, P. R. (2014). Transparency in a Human-Machine Context: Approaches for Fostering Shared Awareness/Intent. In *International Conference on Virtual, Augmented and Mixed Reality (VAMR) - Designing and Developing Virtual and Augmented Environments*, pages 181–190, Heraklion, Greece. Springer, Cham.

MacDorman, K. F., Vasudevan, S. K., and Ho, C.-C. (2009). Does Japan really have robot mania? Comparing attitudes by implicit and explicit measures. *AI & SOCIETY*, 23(4):485–510.

Mackenzie, C., Rogers, W., and Dodds, S., editors (2014). *Vulnerability: New Essays in Ethics and Feminist Philosophy*. Oxford University Press, New York (NY), USA.

Malle, B. F. (2016). Integrating robot ethics and machine morality: the study and design of moral competence in robots. *Ethics and Information Technology*, 18(4):243–256.

Malle, B. F., Fischer, K., Young, J. E., Moon, A., and Collins, E. C. (2020). Trust and the discrepancy between expectations and actual capabilities of social robots. In Zhang, D. and Wei, B., editors, *Human-robot interaction: Control, analysis, and design*, chapter 1, pages 1–23. Cambridge Scholars Publishing, New York, NY, USA.

Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., and Cusimano, C. (2015). Sacrifice One For the Good of Many? In *Proceedings of the 10th ACM/IEEE International Conference on Human-Robot Interaction*, pages 117–124, Portland, USA. ACM.

Malle, B. F. and Ullman, D. (2021). A Multi-Dimensional Conception and Measure of Human-Robot Trust. In Nam, C. S. and Lyons, J. B., editors, *Trust in Human-Robot Interaction: Research and Applications*, chapter 1, pages 3–

25. Academic Press, London, UK.

Mamak, K. (2021). Whether to Save a Robot or a Human: On the Ethical and Legal Limits of Protections for Robots. *Frontiers in robotics and AI*, 8:712427.

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

Matarić, M. J., Eriksson, J., Feil-Seifer, D. J., and Winstein, C. J. (2007). Socially assistive robotics for post-stroke rehabilitation. *Journal of NeuroEngineering and Rehabilitation*, 4(1):5.

Mathur, M. B. and Reichling, D. B. (2009). An Uncanny Game of Trust: Social Trustworthiness of Robots Inferred from Subtle Anthropomorphic Facial Cues. In *4th ACM/IEEE International Conference on Human-Robot Interaction*, pages 11–13, La Jolla, USA. ACM.

Matthias, A. (2015). Robot Lies in Health Care: When Is Deception Morally Permissible? *Kennedy Institute of Ethics Journal*, 25(2):169–162.

Maurtua, I., Ibarguren, A., Kildal, J., Susperregi, L., and Sierra, B. (2017). Human–robot collaboration in industrial applications: Safety, interaction and trust. *International Journal of Advanced Robotic Systems*, 14(4):1–10.

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

McBride, N. and Hoffman, R. R. (2016). Bridging the Ethical Gap: From Human Principles to Robot Instructions. *IEEE Intelligent Systems*, 31(5):76–82.

McKnight, D. H., Choudhury, V., and Kacmar, C. (2002). The impact of initial consumer trust on intentions to transact with a web site: a trust building model. *The Journal of Strategic Information Systems*, 11(3-4):297–323.

McLellan, E., MacQueen, K. M., and Neidig, J. L. (2003). Beyond the Qualitative Interview: Data Preparation and Transcription. *Field Methods*, 15(1):63–84.

McLeod, C. (2021). Trust. *Stanford Encyclopedia of Philosophy*.

McMyler, B. (2011). *Testimony, Trust, and Authority*. Oxford University Press, New York, NY.

McMyler, B. (2017). Deciding to Trust. In Faulkner, P. and Simpson, T. W., editors, *The Philosophy of Trust*, chapter 10, pages 161–176. Oxford University Press, Oxford, UK.

Melson, G. F., Kahn, P. H., Beck, A., and Friedman, B. (2006). Toward Understanding

Children's and Adults' Encounters with Social Robots. In Metzler, T., editor, *Papers from the AAAI Workshop on Human Implications of Human-Robot Interaction (HRI)*, pages 36–43, Boston, USA. AAAI Press.

Meuser, M. and Nagel, U. (2009). The Expert Interview and Changes in Knowledge Production. In Bogner, A., Littig, B., and Menz, W., editors, *Interviewing Experts*, chapter 1, page 281. Palgrave Macmillan, Hampshire, UK, 1st edition.

Miles, M. B., Huberman, A. M., and Saldaña, J. (2020). *Qualitative Data Analysis: A Methods Sourcebook*. SAGE Publications Ltd., Thousand Oaks, CA, 4 edition.

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

Miller, R. L. and Brewer, J. D., editors (2003). *The A-Z of social research : a dictionary of key social science research concepts*. SAGE Publications Ltd., London, UK, 3rd edition.

Mirnig, N., Stollnberger, G., Miksch, M., Stadler, S., Giuliani, M., and Tscheligi, M. (2017). To Err Is Robot: How Humans Assess and Act toward an Erroneous Social Robot. *Frontiers in Robotics and AI*, 4:21.

Misztal, B. A. (1996). *Trust in Modern Societies: The Search for the Bases of Social Order*. Polity Press.

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

Möllering, G. (2006). *Trust: Reason, Routine, Reflexivity*. Emerald Group Publiching Limited, Bingley, UK.

Moor, J. (2006). The Nature, Importance, and Difficulty of Machine Ethics. *IEEE Intelligent Systems*, 21(4):18–21.

Mota, R. C. R., Rea, D. J., Le Tran, A., Young, J. E., Sharlin, E., and Sousa, M. C. (2016). Playing the 'trust game' with robots: Social strategies and experiences. In *Proceedings of the 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 519–524, New York, NY. IEEE.

Nadelhoffer, T. and Nahmias, E. (2007). The Past and Future of Experimental Philosophy. *Philosophical Explorations*, 10(2):123–149.

Natarajan, M. and Gombolay, M. (2020). Effects of Anthropomorphism and Accountability on Trust in Human Robot Interaction. In *Proceedings of the*

15th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pages 33–42, Cambridge, United Kingdom. ACM.

Neubauer, H. (2017). Merkur setzt Roboter als Marketing-Gag ein « DiePresse.com.

Nocks, L. (2008). *The Robot: the Life Story of a Technology*. Johns Hopkins University Press, Baltimore (MD), USA, illustrate edition.

Nomura, T., Kanda, T., Kidokoro, H., Suehiro, Y., and Yamada, S. (2016). Why do children abuse robots? *Interaction Studies*, 17(3):347–369.

Nomura, T., Suzuki, T., Kanda, T., and Kato, K. (2006). Measurement of negative attitudes toward robots. *Interaction Studies*, 7(3):437–454.

Norberg, P. A., Horne, D. R., and Horne, D. A. (2007). The Privacy Paradox: Personal Information Disclosure Intentions versus Behaviors. *Journal of Consumer Affairs*, 41(1):100–126.

Nordqvist, M. and Lindblom, J. (2018). Operators' Experience of Trust in Manual Assembly with a Collaborative Robot. In *Proceedings of the 6th International Conference on Human-Agent Interaction*, pages 341–343, Southampton, UK. ACM.

Nyholm, S. and Smids, J. (2020). Can a Robot Be a Good Colleague? *Science and Engineering Ethics*, 26(4):2169–2188.

Ogawa, R., Park, S., and Umemuro, H. (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)*, pages 606–607, Daegu, South Korea. IEEE.

O'Neill, O. (2002). *Autonomy and Trust in Bioethics*. Cambridge University Press, Cambridge, UK.

Ono, S., Obo, T., Kiong, L. C., and Kubota, N. (2015). Robot Communication Based on Relational Trust Model. In *41st Annual Conference of the IEEE Industrial Electronics Society*, pages 5335–5338, Yokohama, Japan. IEEE.

Pace, M. (2021). Trusting in order to inspire trustworthiness. *Synthese*, 198(12):11897–11923.

Parasuraman, R. and Riley, V. (1997). Humans and Automation: Use, Misuse, Disuse, Abuse. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 39(2):230–253.

Paridon, T. J., Carraher, S., and Carraher, S. C. (2006). The income effect in Personal Shopping Values, Consumer Self-confidence, and Information

Sharing (Word of Mounth Communication) Research. *Academy of Marketing Studies Journal*, 10(2):107–124.

Patton, M. Q. (2015). *Qualitative Research & Evaluation Methods : Integrating Theory and Practice*. SAGE Publications, Inc., Thousand Oaks, CA, 4 edition.

Plakias, A. (2015). Experimental Philosophy. In *Oxford Online Handbooks of Philosophy*, pages 1–18. Oxford University Press, Oxford, UK.

Pradeu, T., Lemoine, M., Khelfaoui, M., and Gingras, Y. (2021). Philosophy in Science: Can philosophers of science permeate through science and produce scientific knowledge? *The British Journal for the Philosophy of Science*, pages 1–80.

PytlikZillig, L. M. and Kimbrough, C. D. (2016). Consensus on Conceptualizations and Definitions of Trust: Are We There Yet? In Shockley, E., Neal, Tess, M. S., PytlikZillig, L. M., and Bornstein, B. H., editors, *Interdisciplinary Perspectives on Trust*, chapter 2, pages 17–48. Springer, Cham, Switzerland.

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

Reeves, B. and Nass, C. I. (1996). *The media equation : how people treat computers, television, and new media like real people and places*. CSLI Publications.

Reig, S., Forlizzi, J., and Steinfeld, A. (2019). Leveraging Robot Embodiment to Facilitate Trust and Smoothness. In *Proceedings 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 742–744, Daegu, South Korea. IEEE.

Rescher, N. (1996). *Process Metaphysics: An Introduction to Process Philosophy*. State University of New York Press, Albany, NY.

Rescher, N. (2000). *Process Philosophy: A Survey of Basic Issues*. University of Pittsburgh Press, Pittsburgh, PA.

Rezazadegan, F., Geng, J., Ghirardi, M., Menga, G., Murè, S., Camuncoli, G., and Demichela, M. (2015). Risked-based Design for the Physical Human-Robot Interaction (pHRI): an Overview. *Chemical Engineering Transactions*, 43:1249–1254.

Richardson, K. (2015). *An Anthropology of Robots and AI : Annihilation Anxiety and Machines*. Routledge, New York (NY), USA.

Rifkin, J. (1995). *The End of Work: The Decline of the Global Labor Force and the*

Dawn of the Post-Market Era. Putnam Publishing Group.

Robertson, J. (2018). *Robo Sapiens Japanicus: Robots, Gender, Family, and the Japanese Nation*. University of California Press, Oakland, USA.

Robinette, P., Li, W., Allen, R., Howards, A. M., and Wagner, A. R. (2016). Overtrust of Robots in Emergency Evacuation Scenarios. In *11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 101–108, Christchurch, New Zealand. IEEE Press.

Rosenthal-von der Pütten, A. and Abrams, A. M. H. (2020). Social Dynamics in Human-Robot Groups – Possible Consequences of Unequal Adaptation to Group Members Through Machine Learning in Human-Robot Groups. In *22nd International Conference on Human-Computer Interaction*, pages 396–411, Copenhagen, Denmark. Springer, Cham.

Rossi, A., Dautenhahn, K., Koay, K. L., and Walters, M. L. (2017). How the Timing and Magnitude of Robot Errors Influence Peoples' Trust of Robots in an Emergency Scenario. In *9th International Conference on Social Robotics*, pages 42–52, Tsukuba, Japan. Springer, Cham.

Rossi, A., Perugia, G., and Rossi, S. (2021). Investigating Customers' Perceived Sensitivity of Information Shared with a Robot Bartender. In Li, H., Ge, S. S., Wu, Y., Wykowska, A., He, H., Liu, X., Li, D., and Perez-Osorio, J., editors, *13th International Conference on Social Robotics (Lecture Notes in Computer Science, vol. 13086)*, pages 119–129. Springer, Cham, Singapore, Republic of Singapore.

Rousseau, D. M., Sitkin, S. B., Burt, R. S., and Camerer, C. (1998). Not So Different After All: A Cross-Discipline View Of Trust. *Academy of Management Review*, 23(3):393–404.

Roy, R. and Warren, J. P. (2019). Card-based design tools: a review and analysis of 155 card decks for designers and designing. *Design Studies*, 63:125–154.

Rueben, M., Aroyo, A. M., Lutz, C., Schmolz, J., Van Cleynenbreugel, P., Corti, A., Agrawal, S., and Smart, W. D. (2018). Themes and Research Directions in Privacy-Sensitive Robotics. In *2018 IEEE Workshop on Advanced Robotics and its Social Impacts (ARSO)*, pages 77–84, Genova, Italy. IEEE.

Ruparelia, N. B. (2010). Software development lifecycle models. *ACM SIGSOFT Software Engineering Notes*, 35(3):8–13.

Sadrfaridpour, B., Saeidi, H., Burke, J., Madathil, K., and Wang, Y. (2016). Modeling and Control of Trust in Human-Robot Collaborative Manufacturing. In

Ranjeev, M., Sofge, D., Wagner, A. R., and Lawless, W. F., editors, *Robust Intelligence and Trust in Autonomous Systems*, chapter 7, pages 115–141. Springer, Boston, MA.

Sætra, H. (2021). Social robot deception and the culture of trust. *Paladyn, Journal of Behavioral Robotics*, 12(1):276–286.

Salem, M., Eyssel, F., Rohlfing, K., Kopp, S., Joublin, F., Eyssel, F., Rohlfing, K., Kopp, S., and Joublin, F. (2013). To Err is Human(-like): Effects of Robot Gesture on Perceived Anthropomorphism and Likability. *Int J Soc Robot*, 5(3):313–323.

Salem, M., Lakatos, G., Amirabdollahian, F., and Dautenhahn, K. (2015a). Towards Safe and Trustworthy Social Robots: Ethical Challenges and Practical Issues. In *7th International Conference on Social Robotics. Lecture Notes in Artificial Intelligence (Vol. 9388)*, pages 584–593, Paris, France. Springer, Cham.

Salem, M., Lakatos, G., Amirabdollahian, F., and Dautenhahn, K. (2015b). 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)*, pages 141–148, Portland (OR), USA. ACM.

Salomons, N., Van Der Linden, M., Sebo, S. S., and Scassellati, B. (2018). Humans Conform to Robots: Disambiguating Trust, Truth, and Conformity. In *14 ACM/IEEE International Conference on Human-Robot Interaction*, pages 187–195, Chicago (IL), USA. ACM.

Sandelowski, M. (1995). Sample size in qualitative research. *Research in Nursing & Health*, 18(2):179–183.

Sanders, T., Kaplan, A., Koch, R., Schwartz, M., and Hancock, P. A. (2019). The Relationship Between Trust and Use Choice in Human-Robot Interaction. *Human factors*, 61(4):614–626.

Sanders, T. L., MacArthur, K., Volante, W., Hancock, G., MacGillivray, T., Shugars, W., and Hancock, P. A. (2017). Trust and Prior Experience in Human-Robot Interaction. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting (HFES)*, volume 61, pages 1809–1813, Austin, USA. SAGE Publications.

Sanders, T. L., Wixon, T., Schafer, K. E., Chen, J. Y. C., and Hancock, P. A. (2014). The influence of modality and transparency on trust in human-robot interaction. In *IEEE International Inter-Disciplinary Conference on Cognitive Methods in*

Situation Awareness and Decision Support (CogSIMA), pages 156–159, San Antonio, USA. IEEE.

Saunderson, S. and Nejat, G. (2022). Investigating Strategies for Robot Persuasion in Social Human-Robot Interaction. *IEEE Transactions on Cybernetics*, 52(1):641–653.

Schaefer, K. E., Sanders, T. L., Yordon, R. E., Billings, D. R., and Hancock, P. A. (2012). Classification of robot form: Factors predicting perceived trustworthiness. *Proceedings of the Human Factors and Ergonomics Society*, pages 1548–1552.

Scheeff, M., Pinto, J., Rahardja, K., Snibbe, S., and Tow, R. (2002). Experiences with Sparky, a Social Robot. In Dautenhahn, K., Bond, A., Cañamero, L., and Edmonds, B., editors, *Socially Intelligent Agents - Creating Relationships with Computers and Robots*, chapter 21, pages 173–180. Springer, Boston (MA), USA.

Scheman, N. (2020). Trust and Trustworthiness. In Simon, J., editor, *The Routledge Handbook of Trust and Philosophy*, chapter 2, pages 28–40. Routledge, New York, NY (USA).

Scheutz, M. and Malle, B. F. (2014). "Think and do the right thing" - A Plea for morally competent autonomous robots. In *2014 IEEE International Symposium on Ethics in Science, Technology and Engineering*, pages 1–4, Chicago, IL. IEEE.

Schwaninger, I., Güldenpfennig, F., Weiss, A., and Fitzpatrick, G. (2021). What Do You Mean by Trust? Establishing Shared Meaning in Interdisciplinary Design for Assistive Technology. *International Journal of Social Robotics*, 13(8):1879–1897.

Searle, J. R. (1980). Minds, Brains, and Programs. *Behavioral and Brain Sciences*, 3(3):417–424.

Sebele-Mpofu, F. Y. (2020). Saturation controversy in qualitative research: Complexities and underlying assumptions. A literature review. *Cogent Social Sciences*, 6(1):1838706.

Sebo, S., Traeger, M., Jung, M., and Scassellati, B. (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*, pages 178–186, Chicago (IL), USA. ACM.

Sebo, S. S., Krishnamurthi, P., and Scassellati, B. (2019). 'I Don't Believe You': Investigating the Effects of Robot Trust Violation and Repair. In *14th ACM/IEEE International Conference on Human-Robot Interaction*, pages 57–65, Daegu, South Korea. IEEE.

Seibt, J. (2017a). Process Philosophy.

Seibt, J. (2017b). Towards an Ontology of Simulated Social Interaction: Varieties of the "As If" for Robots and Humans. In Hakli, R. and Seibt, J., editors, *Sociality and Normativity for Robots*, pages 11–39. Springer International Publishing, Cham, Switzerland.

Seibt, J., Vestergaard, C., and Damholdt, M. F. (2021). The Complexity of Human Social Interactions Calls for Mixed Methods in HRI. *ACM Transactions on Human-Robot Interaction*, 10(1):1–4.

Seidl, D. (2004). Luhmann's theory of autopoietic social systems. *Munich Business Research*, 2:1–28.

Shanahan, M. (1995). A circumscriptive calculus of events. *Artificial Intelligence*, 77(2):249–284.

Sharkey, A. and Sharkey, N. (2010). Granny and the robots: ethical issues in robot care for the elderly. *Ethics and Information Technology*, 14(1):27–40.

Sharkey, A. and Sharkey, N. (2020). We need to talk about deception in social robotics! *Ethics and Information Technology*, pages 1–8.

Sharkey, A. and Sharkey, N. (2021). We need to talk about deception in social robotics! *Ethics and Information Technology*, 23(3):309–316.

Shaw, K. (2018). HSBC Bank Rolls Out Pepper Robot at Flagship U.S. Branch.

Shelley, M. (2012). *Frankenstein*. Penguin Publishing Group, London, UK.

Sheppard, S., Colby, A., Macatangay, K., and Sullivan, W. (2006). What is engineering practice? *International Journal of Engineering Education*, 22(3):429–438.

Siino, R. M., Chung, J., and Hinds, P. J. (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)*, pages 558–562, Munich, Germany. IEEE.

Sim, J., Saunders, B., Waterfield, J., and Kingstone, T. (2018). Can sample size in qualitative research be determined a priori? *International Journal of Social Research Methodology*, 21(5):619–634.

Simmel, G. (2004). *The Philosophy of Money*. Routledge, London, UK, 3 edition.

Smids, J., Nyholm, S., and Berkers, H. (2020). Robots in the Workplace: a

Threat to—or Opportunity for—Meaningful Work? *Philosophy & Technology*, 33(3):503–522.

Smith, M. N. (2010). Reliance. *Noûs*, 44(1):135–157.

Soh, H., Xie, Y., Chen, M., and Hsu, D. (2020). Multi-task trust transfer for human–robot interaction. *The International Journal of Robotics Research*, 39(2-3):233–249.

Solove, D. J. (2021). The Myth of the Privacy Paradox. *George Washington Law Review*, 89(1):1–51.

Sparrow, R. and Sparrow, L. (2006). In the hands of machines? The future of aged care. *Minds and Machines*, 16(2):141–161.

Spencer, D. A., Wang, Y., and Humphrey, L. R. (2016). Trust-based human-robot interaction for multi-robot symbolic motion planning. In *EEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1443–1449, Daejeon, South Korea. IEEE.

Stanton, C. J. and Stevens, C. J. (2017). Don't Stare at Me: The Impact of a Humanoid Robot's Gaze upon Trust During a Cooperative Human–Robot Visual Task. *International Journal of Social Robotics*, 9(5):745–753.

Stone, W. L. (2005). The History of Robotics. In Kurfess, T. R., editor, *Robotics and Automation Handbook*, chapter 1, page 12. CRC Press, Boca Raton, FL, 1 edition.

Strawson, P. F. (1962). Freedom and Resentment. *Proceedings of the British Academy*, 48:187–211.

Sturken, M., Thomas, D., and Ball-Rokeach, S. J., editors (2004). *Technological Visions: The Hopes and Fears that Shape New Technologies*. Temple University Press, Philadelphia, PA.

Subramanian, R. (2017). Emergent AI, Social Robots and the Law: Security, Privacy and Policy Issues. *Journal of International, Technology and Information Management*, 6(3):1–27.

Sullins, J. P. (2020). Trust in Robots. In Simon, J., editor, *The Routledge Handbook of Trust and Philosophy*, pages 313–325. Routledge.

Syrdal, D. S., Walters, M. L., Otero, N., Koay, K. L., and Dautenhahn, K. (2007). &quot;He Knows When You Are Sleeping&quot; : Privacy and the Personal Robot Companion. In *Proceedings from Workshop Human Implications of Human-Robot Interaction, Association for the Advancement of Artificial Intelligence (AAAI 07)*, pages 28–33, Vancouver, Canada. AAAI Press.

Takayama, L. (2012). Perspectives on Agency Interacting with and through Personal Robots. In Zacarias, M. and de Oliveira, J. V., editors, *Human-Computer Interaction: The Agency Perspective (Studies in Computational Intelligence, Vol. 396)*, pages 195–214. Springer, Berlin/Heidelberg, Germany.

Takayama, L., Nass, C., and Ju, W. (2008). Beyond dirty, dangerous and dull: what everyday people think robots should do. In *3rd ACM/IEEE International Conference on Human-Robot Interaction*, pages 25–32, Amsterdam, The Netherlands. ACM.

Tallant, J. (2019). You Can Trust the Ladder, But You Shouldn't. *Theoria*, 85(2):102–118.

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

Thomasson, A. L. (2015). *Ontology Made Easy*. Oxford University Press, New York, NY.

Thompson, C. (2017). Trust without Reliance. *Ethical Theory and Moral Practice*, 20(3):643–655.

Tolmeijer, S., Kneer, M., Sarasua, C., Christen, M., and Bernstein, A. (2021). Implementations in Machine Ethics. *ACM Computing Surveys*, 53(6):1–38.

Tolmeijer, S., Weiss, A., Hanheide, M., Lindner, F., Powers, T. M., Dixon, C., and Tielman, M. L. (2020). Taxonomy of Trust-Relevant Failures and Mitigation Strategies. In *15th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 3–12, Cambridge, United Kingdom. ACM.

Traeger, M. L., Strohkorb Sebo, S., Jung, M., Scassellati, B., and Christakis, N. A. (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):6370–6375.

Tuisku, O., Pekkarinen, S., Hennala, L., and Melkas, H. (2019). "Robots do not replace a nurse with a beating heart". *Information Technology & People*, 32(1):47–67.

Turing, A. (1950). Computing Machinery and Intelligence. *Mind*, 59(236):433–460.

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

Ullman, D. and Malle, B. F. (2017). Human-Robot Trust: Just a Button Press Away. In *Proceedings of the 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 309–310, Vienna, Austria. ACM.

Ullman, D. and Malle, B. F. (2018). What Does it Mean to Trust a Robot? Steps Toward a Multidimensional Measure of Trust. In *Proceedings of the 13th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 263–264, Chicago, USA. ACM.

Vallor, S. (2011). Carebots and Caregivers: Sustaining the Ethical Ideal of Care in the Twenty-First Century. *Philosophy & Technology*, 24(3):251–268.

van den Brule, R., Dotsch, R., Bijlstra, G., Wigboldus, D. H. J., and Haselager, P. (2014). Do Robot Performance and Behavioral Style affect Human Trust? *International Journal of Social Robotics*, 6(4):519–531.

Van der Hoorn, D. P., Neerincx, A., and de Graaf, M. M. (2021). "I think you are doing a bad job!": The Effect of Blame Attribution by a Robot in Human-Robot Collaboration. In *Proceedings of the 16th ACM/IEEE International Conference on Human-Robot Interaction*, pages 140–148, Boulder (CO), USA. ACM.

van Maris, A., Lehmann, H., Natale, L., and Grzyb, B. (2017). The Influence of a Robot's Embodiment on Trust. In *Proceedings of the 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 313–314, Vienna, Austria. ACM.

van Straten, C. L., Peter, J., Kühne, R., de Jong, C., and Barco, A. (2018). Technological and Interpersonal Trust in Child-Robot Interaction. In *Proceedings of the 6th International Conference on Human-Agent Interaction (HAI)*, pages 253–259, Southampton, UK. ACM.

Vanderelst, D. and Winfield, A. (2018). An architecture for ethical robots inspired by the simulation theory of cognition. *Cognitive Systems Research*, 48:56–66.

Vilches, V. M., Gil-Uriarte, E., Ugarte, I. Z., Mendia, G. O., Pisón, R. I., Kirschgens, L. A., Calvo, A. B., Cordero, A. H., Apa, L., and Cerrudo, C. (2018). Towards an open standard for assessing the severity of robot security vulnerabilities, the Robot Vulnerability Scoring System (RVSS). Technical report, Alias Robotics S.L.

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

Vincent, D. A. (1999). Changin Our World (guest forword). In Nof, S. Y., editor, *Handbook of Industrial Robotics*, page 1349. John Wiley & Sons, Inc., 2 edition.

Wacker, J. G. (1998). A definition of theory: research guidelines for different theory-building research methods in operations management. *Journal of Operations Management*, 16(4):361–385.

Wada, K., Shibata, T., Saito, T., Sakamoto, K., and Tanie, K. (2005). Psychological and Social Effects of One Year Robot Assisted Activity on Elderly People at a Health Service Facility for the Aged. In *Proceedings of the 2005 IEEE International Conference on Robotics and Automation*, pages 2785–2790, Barcelona, Spain. IEEE.

Wagner, A. R. and Robinette, P. (2021). An explanation is not an excuse: Trust calibration in an age of transparent robots. In Nam, C. S. and Lyons, J. B., editors, *Trust in Human-Robot Interaction*, chapter 9, pages 197–208. Academic Press, London, UK.

Wagner, A. R., Robinette, P., and Howard, A. (2018). Modeling the Human-Robot Trust Phenomenon: A Conceptual Framework based on Risk. *ACM Transactions on Interactive Intelligent Systems*, 8(4).

Walker, I. D., Mears, L., Mizanoor, R. S. M., Pak, R., Remy, S., and Wang, Y. (2015). Robot-Human Handovers Based on Trust. In *2nd International Conference on Mathematics and Computers in Sciences and in Industry*, pages 119–124, Sliema, Malta. IEEE.

Wallach, W. and Allen, C. (2009). *Moral Machines: Teaching Robots Right from Wrong*. Oxford University Press, New York, USA.

Wallach, W. and Asaro, P. (2017). *Machine ethics and robot ethics*. Routledge, New York, NY, 1 edition.

Wanderer, J. and Townsend, L. (2013). Is it Rational to Trust? *Philosophy Compass*, 8(1):1–14.

Wang, N., Pynadath, D. V., and Hill, S. G. (2016). Trust calibration within a human-robot team: Comparing automatically generated explanations. In *11th ACM/IEEE International Conference on Human-Robot Interaction (HRI'16)*, pages 109–116, Christchurch, New Zealand. IEEE.

Wang, Y., Shi, Z., Wang, C., and Zhang, F. (2014). Human-Robot Mutual Trust in (Semi)autonomous Underwater Robots. In Koubaa, A. and Khelil, A., editors, *Cooperative Robots and Sensor Networks*, chapter 6, pages 115–137. Springer, Berlin and Heidelberg, Germany.

Washburn, A., Adeleye, A., An, T., and Riek, L. D. (2019). Robot Errors in Proximate HRI: How Functionality Framing Affects Perceived Reliability and Trust. *ACM*

Transactions on Human-Robot Interaction, 9(3):1–21.

Weiss, A., Igelsböck, J., Wurhofer, D., and Tscheligi, M. (2011). Looking Forward to a “Robotic Society”? *International Journal of Social Robotics*, 3(2):111–123.

Weiss, A. and Spiel, K. (2022). Robots beyond Science Fiction: mutual learning in human–robot interaction on the way to participatory approaches. *AI & SOCIETY*, 37(2):501–515.

Weiss, A., Wortmeier, A.-K., and Kubicek, B. (2021). Cobots in Industry 4.0: A Roadmap for Future Practice Studies on Human–Robot Collaboration. *IEEE Transactions on Human-Machine Systems*, 51(4):335–345.

Westerlund, L. (2000). *The Extended Arm of Man: A History of the Industrial Robot*. Informationsförlaget, Stockholm, Sweden.

Whitby, B. (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):326–333.

Williamson, T. (2007). *The Philosophy of Philosophy*. Blackwell Publishing, Oxford, UK.

Williamson, T. (2016). Philosophical Criticisms of Experimental Philosophy. In *A Companion to Experimental Philosophy*, pages 22–36. John Wiley & Sons, Ltd, Chichester, UK.

Wölfel, C. and Merritt, T. (2013). Method Card Design Dimensions: A Survey of Card-Based Design Tools. In *Human-Computer Interaction – INTERACT 2013. Lecture Notes in Computer Science, vol 8117.*, pages 479–486, Berlin, Germany. Springer.

Wolfert, P., Deschuyteneer, J., Oetringer, D., Robinson, N., and Belpaeme, T. (2020). Security Risks of Social Robots Used to Persuade and Manipulate. In *Proceedings of the 15th ACM/IEEE International Conference on Human-Robot Interaction*, pages 523–525, Cambridge, UK. ACM.

Woods, S., Walters, M., Kheng Lee Koay, and Dautenhahn, K. (2006). Comparing human robot interaction scenarios using live and video based methods: towards a novel methodological approach. In *9th IEEE International Workshop on Advanced Motion Control (AMC)*, pages 750–755, Istanbul, Turkey. IEEE.

Wooldridge, M. and Jennings, N. R. (1995). Intelligent agents: theory and practice. *The Knowledge Engineering Review*, 10(2):115–152.

Xie, Y., Bodala, I. P., Ong, D. C., Hsu, D., and Soh, H. (2019). Robot Capability and Intention in Trust-Based Decisions Across Tasks. In *Proceedings of the 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 39–47, Daegu, South Korea. IEEE.

Xu, A. and Dudek, G. (2015). OPTIMo: Online Probabilistic Trust Inference Model for Asymmetric Human-Robot Collaborations. In *10th Annual ACM/IEEE International Conference on Human-Robot Interaction*, pages 221–228, Portland (OR), USA. ACM.

Xu, A. and Dudek, G. (2016). Maintaining efficient collaboration with trust-seeking robots. In *EEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3312–3319, Daejeon, South Korea. IEEE.

Xu, Q., Ng, J., Tan, O., Huang, Z., Tay, B., and Park, T. (2015). Methodological Issues in Scenario-Based Evaluation of Human–Robot Interaction. *International Journal of Social Robotics*, 7(2):279–291.

Zacharaki, A., Kostavelis, I., Gasteratos, A., and Dokas, I. (2020). Safety bounds in human robot interaction: A survey. *Safety Science*, 127:1–19.

Zagzebski, L. (2009). *On Epistemology*. Wadsworth, Belmont, CA.

Zhang, M., Gursoy, D., Zhu, Z., and Shi, S. (2021). Impact of anthropomorphic features of artificially intelligent service robots on consumer acceptance: moderating role of sense of humor. *International Journal of Contemporary Hospitality Management*, 33(11):3883–3905.

Zhu, L. and Williams, T. (2020). Effects of Proactive Explanations by Robots on Human-Robot Trust. In *12th International Conference on Social Robotics*, pages 85–95, Golden (CO), USA. Springer.

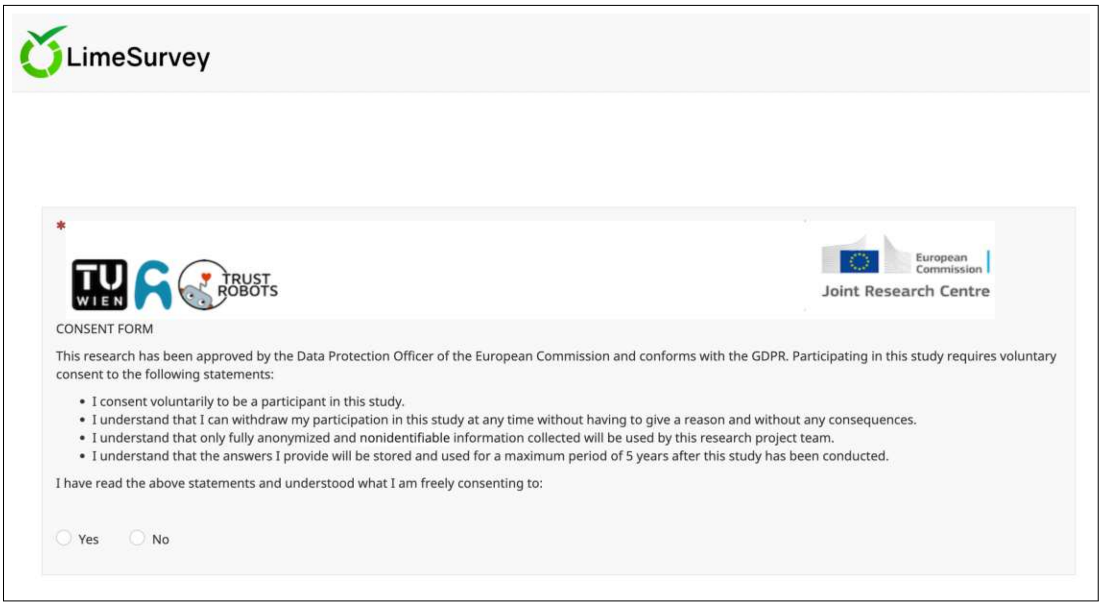
Złotowski, J., Proudfoot, D., Yogeeswaran, K., and Bartneck, C. (2015). Anthropomorphism: Opportunities and Challenges in Human–Robot Interaction. *International Journal of Social Robotics*, 7(3):347–360.

Złotowski, J., Sumioka, H., Nishio, S., Glas, D. F., Bartneck, C., and Ishiguro, H. (2016). Appearance of a robot affects the impact of its behaviour on perceived trustworthiness and empathy. *Paladyn*, 7(1):55–66.

APPENDIX A

Online HRI Study

A.1 Informed Consent Form



*

CONSENT FORM

This research has been approved by the Data Protection Officer of the European Commission and conforms with the GDPR. Participating in this study requires voluntary consent to the following statements:

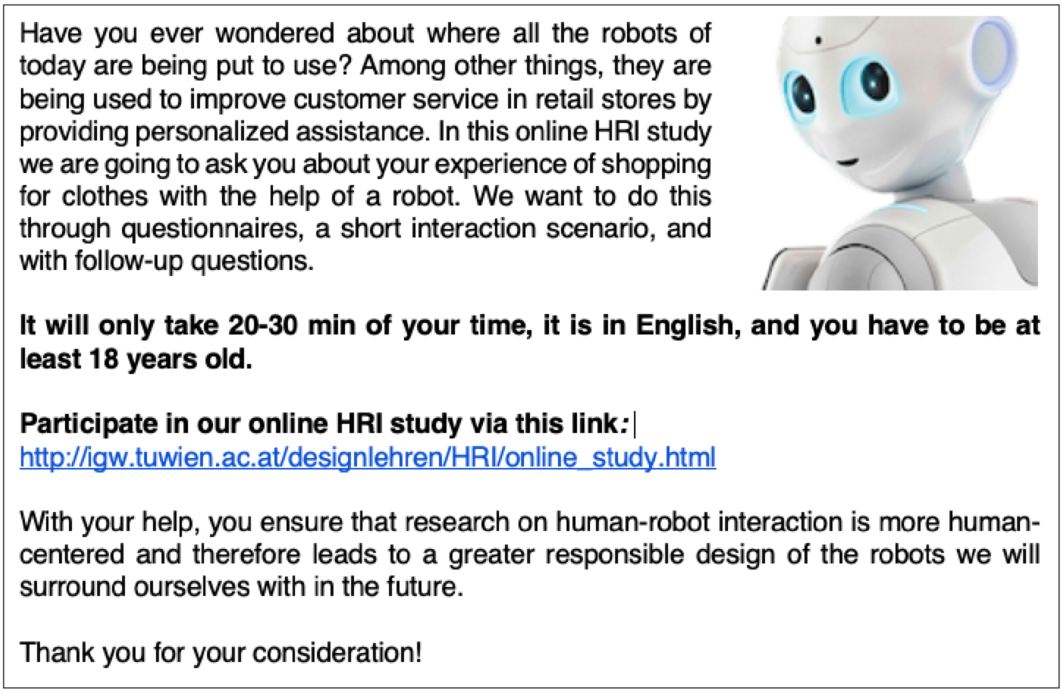
- I consent voluntarily to be a participant in this study.
- I understand that I can withdraw my participation in this study at any time without having to give a reason and without any consequences.
- I understand that only fully anonymized and nonidentifiable information collected will be used by this research project team.
- I understand that the answers I provide will be stored and used for a maximum period of 5 years after this study has been conducted.

I have read the above statements and understood what I am freely consenting to:

○ Yes ○ No

Figure A.1: The consent form used for the online study as displayed to the participants filling in the interactive online survey.

A.2 Participant Recruitment Material



Have you ever wondered about where all the robots of today are being put to use? Among other things, they are being used to improve customer service in retail stores by providing personalized assistance. In this online HRI study we are going to ask you about your experience of shopping for clothes with the help of a robot. We want to do this through questionnaires, a short interaction scenario, and with follow-up questions.

It will only take 20-30 min of your time, it is in English, and you have to be at least 18 years old.

Participate in our online HRI study via this link:
http://igw.tuwien.ac.at/designlehren/HRI/online_study.html

With your help, you ensure that research on human-robot interaction is more human-centered and therefore leads to a greater responsible design of the robots we will surround ourselves with in the future.

Thank you for your consideration!

Figure A.2: The flyer we used to recruit participants through mailing lists and posts on social media platforms.

A.3 Interactive Online Survey Links

Link to the **economy** scenario: `http://128.130.182.53/index.php?r=survey/index&sid=10&lang=en`

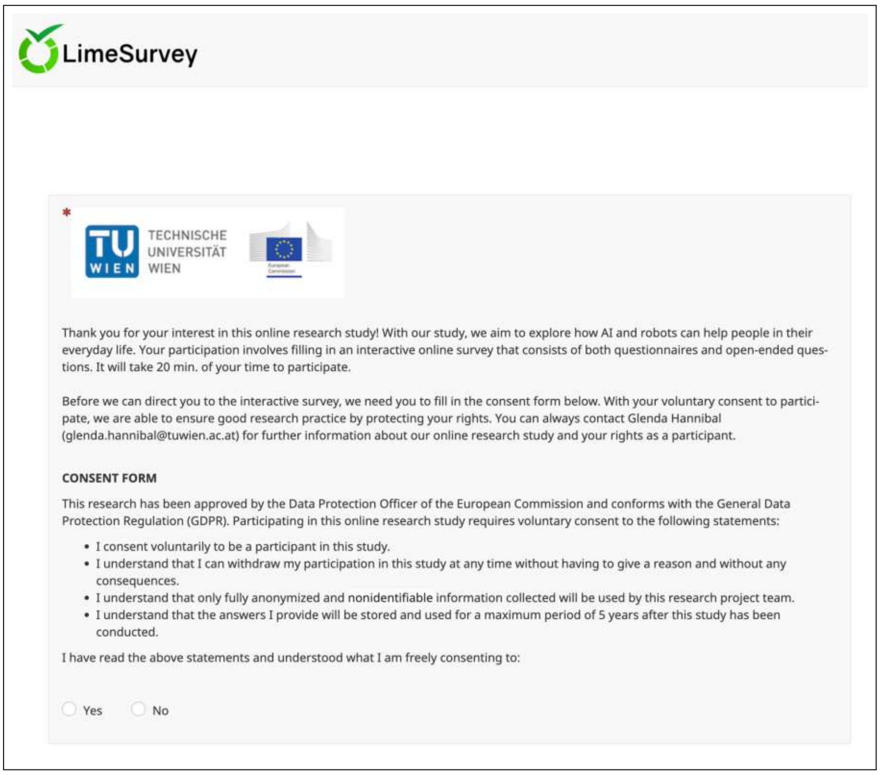
Link to the **privacy** scenario: `http://128.130.182.53/index.php?r=survey/index&sid=11&lang=en`

Link to the **transparency** scenario: `http://128.130.182.53/index.php?r=survey/index&sid=12&lang=en`

APPENDIX B

Follow-up Online HRI Study

B.1 Informed Consent Form



Thank you for your interest in this online research study! With our study, we aim to explore how AI and robots can help people in their everyday life. Your participation involves filling in an interactive online survey that consists of both questionnaires and open-ended questions. It will take 20 min. of your time to participate.

Before we can direct you to the interactive survey, we need you to fill in the consent form below. With your voluntary consent to participate, we are able to ensure good research practice by protecting your rights. You can always contact Glenda Hannibal (glenda.hannibal@tuwien.ac.at) for further information about our online research study and your rights as a participant.

CONSENT FORM

This research has been approved by the Data Protection Officer of the European Commission and conforms with the General Data Protection Regulation (GDPR). Participating in this online research study requires voluntary consent to the following statements:

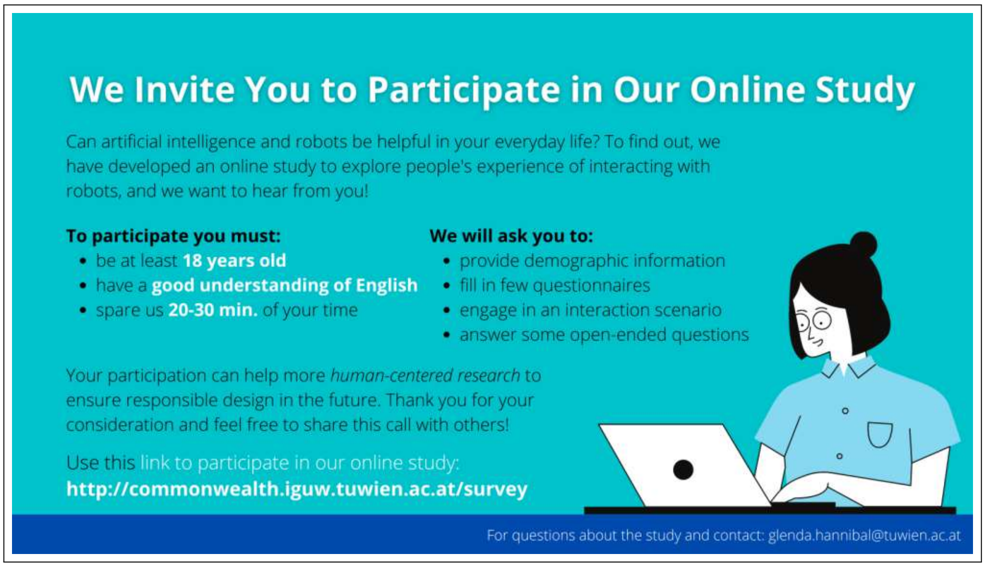
- I consent voluntarily to be a participant in this study.
- I understand that I can withdraw my participation in this study at any time without having to give a reason and without any consequences.
- I understand that only fully anonymized and nonidentifiable information collected will be used by this research project team.
- I understand that the answers I provide will be stored and used for a maximum period of 5 years after this study has been conducted.

I have read the above statements and understood what I am freely consenting to:

○ Yes ○ No

Figure B.1: The consent form used for the follow-up online study as displayed to the participants filling in the interactive online survey.

B.2 Participant Recruitment Material



We Invite You to Participate in Our Online Study

Can artificial intelligence and robots be helpful in your everyday life? To find out, we have developed an online study to explore people's experience of interacting with robots, and we want to hear from you!

To participate you must:

- be at least **18 years old**
- have a **good understanding of English**
- spare us **20-30 min.** of your time

We will ask you to:

- provide demographic information
- fill in few questionnaires
- engage in an interaction scenario
- answer some open-ended questions

Your participation can help more *human-centered research* to ensure responsible design in the future. Thank you for your consideration and feel free to share this call with others!

Use this link to participate in our online study:
http://commonwealth.iguw.tuwien.ac.at/survey

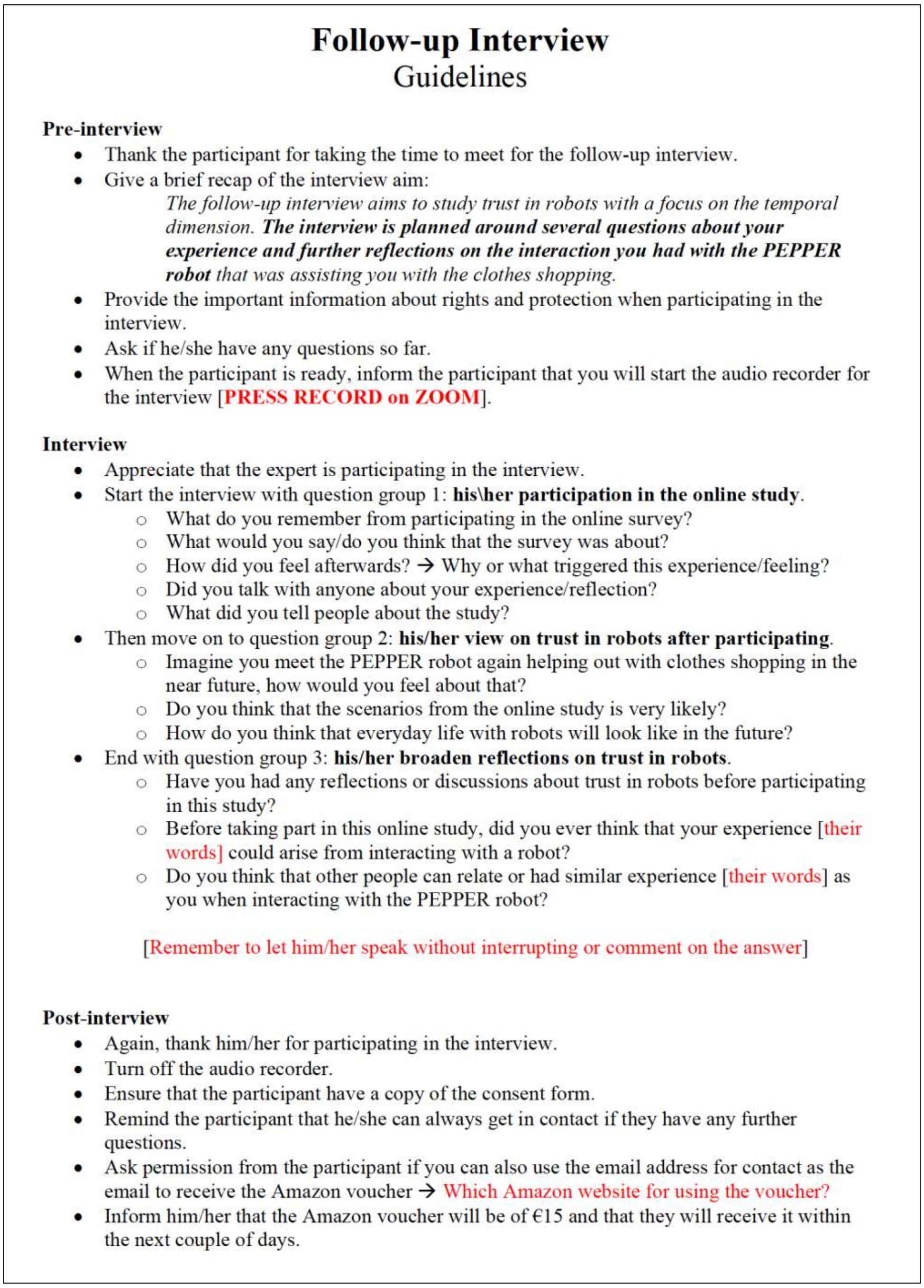
For questions about the study and contact: glenda.hannibal@tuwien.ac.at

Figure B.2: The flyer we used to recruit participant through mailing lists and posts on social media platforms.

B.3 Iterated Privacy Survey Link

Link to the privacy scenario: `http://128.130.182.53/index.php?r=survey/index&sid=24&lang=en`

B.4 Interview Guideline



Follow-up Interview
Guidelines

Pre-interview

- Thank the participant for taking the time to meet for the follow-up interview.
- Give a brief recap of the interview aim:

  The follow-up interview aims to study trust in robots with a focus on the temporal dimension. ***The interview is planned around several questions about your experience and further reflections on the interaction you had with the PEPPER robot*** *that was assisting you with the clothes shopping.*
- Provide the important information about rights and protection when participating in the interview.
- Ask if he/she have any questions so far.
- When the participant is ready, inform the participant that you will start the audio recorder for the interview [**PRESS RECORD on ZOOM**].

Interview

- Appreciate that the expert is participating in the interview.
- Start the interview with question group 1: **his\her participation in the online study**.
  - What do you remember from participating in the online survey?
  - What would you say/do you think that the survey was about?
  - How did you feel afterwards? → Why or what triggered this experience/feeling?
  - Did you talk with anyone about your experience/reflection?
  - What did you tell people about the study?
- Then move on to question group 2: **his/her view on trust in robots after participating**.
  - Imagine you meet the PEPPER robot again helping out with clothes shopping in the near future, how would you feel about that?
  - Do you think that the scenarios from the online study is very likely?
  - How do you think that everyday life with robots will look like in the future?
- End with question group 3: **his/her broaden reflections on trust in robots**.
  - Have you had any reflections or discussions about trust in robots before participating in this study?
  - Before taking part in this online study, did you ever think that your experience [their words] could arise from interacting with a robot?
  - Do you think that other people can relate or had similar experience [their words] as you when interacting with the PEPPER robot?

[Remember to let him/her speak without interrupting or comment on the answer]

Post-interview

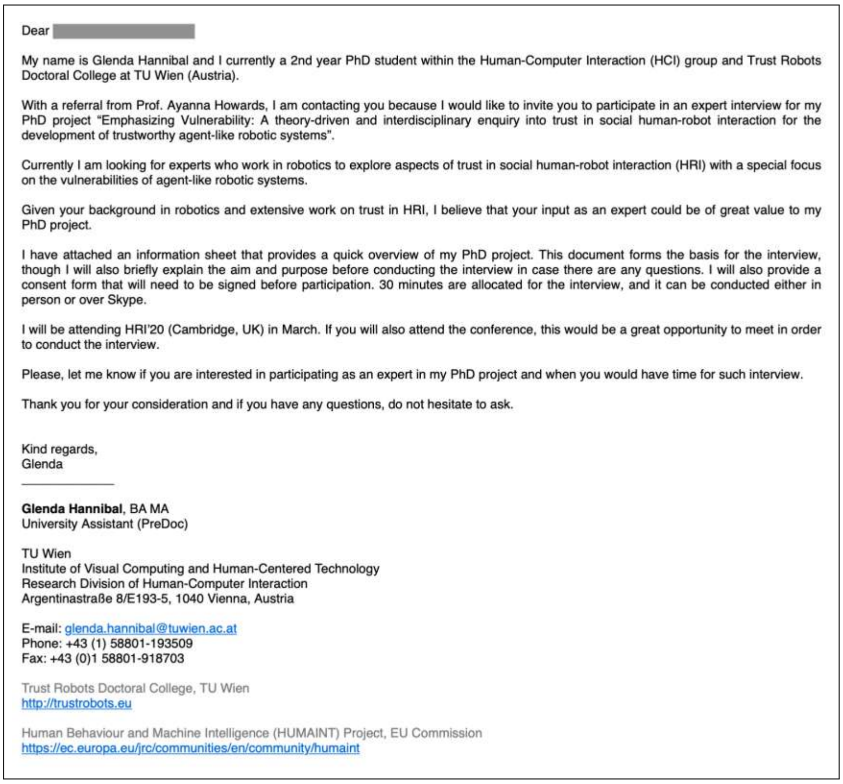
- Again, thank him/her for participating in the interview.
- Turn off the audio recorder.
- Ensure that the participant have a copy of the consent form.
- Remind the participant that he/she can always get in contact if they have any further questions.
- Ask permission from the participant if you can also use the email address for contact as the email to receive the Amazon voucher → Which Amazon website for using the voucher?
- Inform him/her that the Amazon voucher will be of €15 and that they will receive it within the next couple of days.

Figure B.3: The interview guidelines used for conducting the semi-structured interviews.

APPENDIX C

Expert Interviews

C.1 Participant Recruitment Material



Dear [redacted]

My name is Glenda Hannibal and I currently a 2nd year PhD student within the Human-Computer Interaction (HCI) group and Trust Robots Doctoral College at TU Wien (Austria).

With a referral from Prof. Ayanna Howards, I am contacting you because I would like to invite you to participate in an expert interview for my PhD project "Emphasizing Vulnerability: A theory-driven and interdisciplinary enquiry into trust in social human-robot interaction for the development of trustworthy agent-like robotic systems".

Currently I am looking for experts who work in robotics to explore aspects of trust in social human-robot interaction (HRI) with a special focus on the vulnerabilities of agent-like robotic systems.

Given your background in robotics and extensive work on trust in HRI, I believe that your input as an expert could be of great value to my PhD project.

I have attached an information sheet that provides a quick overview of my PhD project. This document forms the basis for the interview, though I will also briefly explain the aim and purpose before conducting the interview in case there are any questions. I will also provide a consent form that will need to be signed before participation. 30 minutes are allocated for the interview, and it can be conducted either in person or over Skype.

I will be attending HRI'20 (Cambridge, UK) in March. If you will also attend the conference, this would be a great opportunity to meet in order to conduct the interview.

Please, let me know if you are interested in participating as an expert in my PhD project and when you would have time for such interview.

Thank you for your consideration and if you have any questions, do not hesitate to ask.

Kind regards,
Glenda

Glenda Hannibal, BA MA
University Assistant (PreDoc)

TU Wien
Institute of Visual Computing and Human-Centered Technology
Research Division of Human-Computer Interaction
Argentinastraße 8/E193-5, 1040 Vienna, Austria

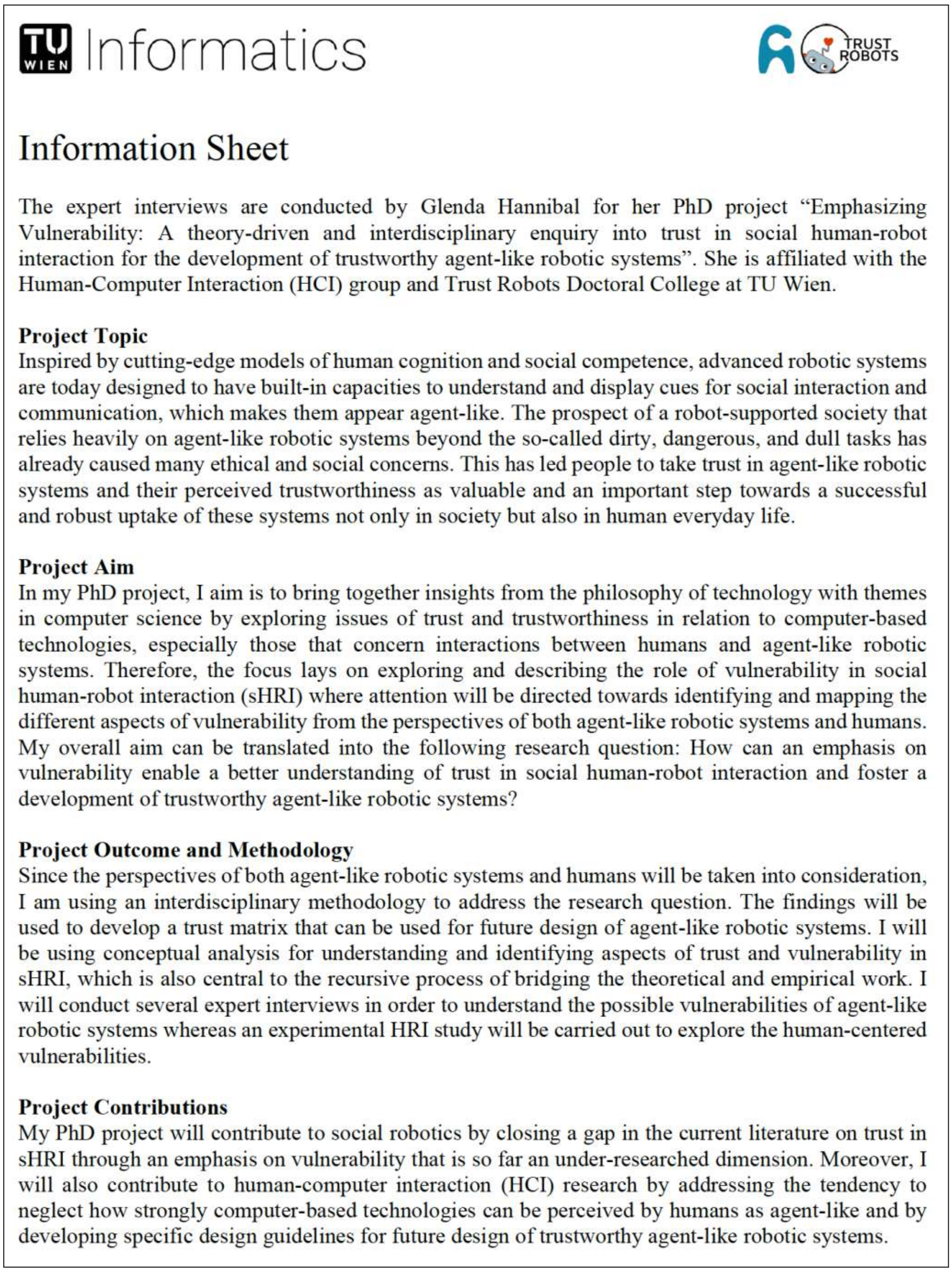
E-mail: glenda.hannibal@tuwien.ac.at
Phone: +43 (1) 58801-193509
Fax: +43 (0)1 58801-918703

Trust Robots Doctoral College, TU Wien
http://trustrobots.eu

Human Behaviour and Machine Intelligence (HUMAINT) Project, EU Commission
https://ec.europa.eu/jrc/communities/en/community/humaint

Figure C.1: The invitation via email that I sent to the experts for requesting an interview.

C.2 Project Description



TU WIEN Informatics

TRUST ROBOTS

Information Sheet

The expert interviews are conducted by Glenda Hannibal for her PhD project "Emphasizing Vulnerability: A theory-driven and interdisciplinary enquiry into trust in social human-robot interaction for the development of trustworthy agent-like robotic systems". She is affiliated with the Human-Computer Interaction (HCI) group and Trust Robots Doctoral College at TU Wien.

Project Topic
Inspired by cutting-edge models of human cognition and social competence, advanced robotic systems are today designed to have built-in capacities to understand and display cues for social interaction and communication, which makes them appear agent-like. The prospect of a robot-supported society that relies heavily on agent-like robotic systems beyond the so-called dirty, dangerous, and dull tasks has already caused many ethical and social concerns. This has led people to take trust in agent-like robotic systems and their perceived trustworthiness as valuable and an important step towards a successful and robust uptake of these systems not only in society but also in human everyday life.

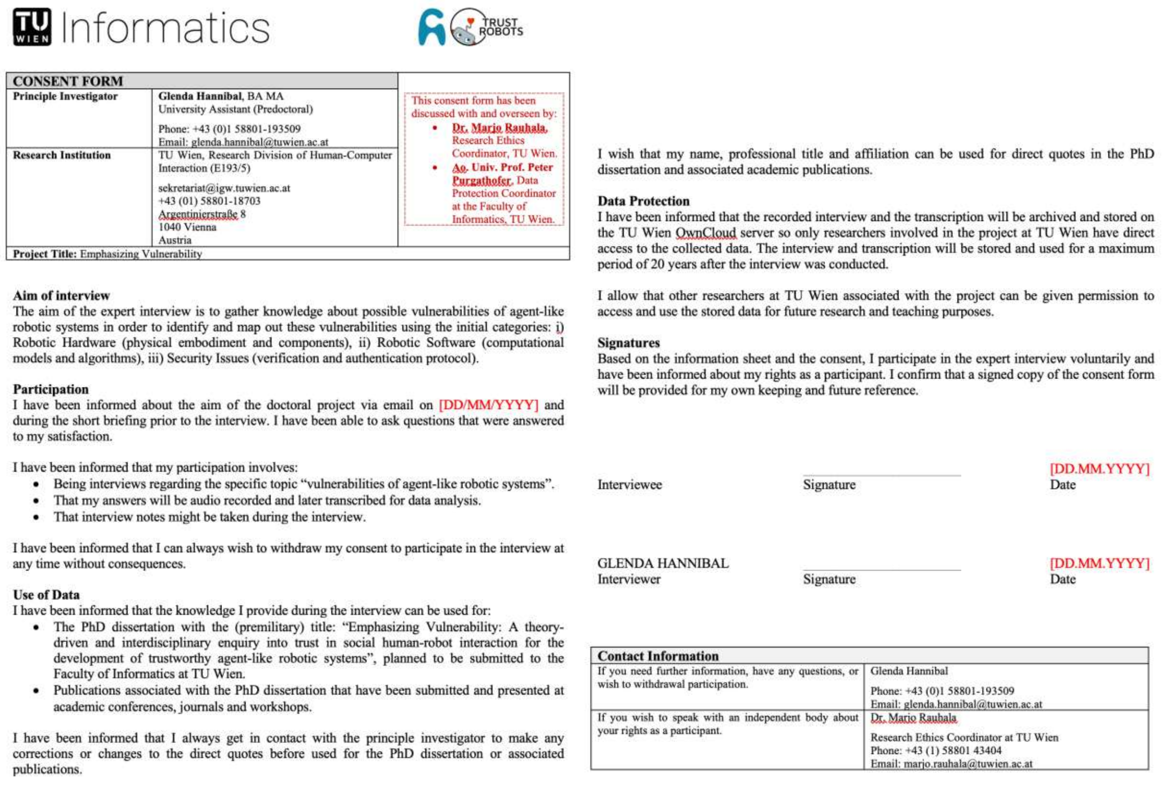
Project Aim
In my PhD project, I aim is to bring together insights from the philosophy of technology with themes in computer science by exploring issues of trust and trustworthiness in relation to computer-based technologies, especially those that concern interactions between humans and agent-like robotic systems. Therefore, the focus lays on exploring and describing the role of vulnerability in social human-robot interaction (sHRI) where attention will be directed towards identifying and mapping the different aspects of vulnerability from the perspectives of both agent-like robotic systems and humans. My overall aim can be translated into the following research question: How can an emphasis on vulnerability enable a better understanding of trust in social human-robot interaction and foster a development of trustworthy agent-like robotic systems?

Project Outcome and Methodology
Since the perspectives of both agent-like robotic systems and humans will be taken into consideration, I am using an interdisciplinary methodology to address the research question. The findings will be used to develop a trust matrix that can be used for future design of agent-like robotic systems. I will be using conceptual analysis for understanding and identifying aspects of trust and vulnerability in sHRI, which is also central to the recursive process of bridging the theoretical and empirical work. I will conduct several expert interviews in order to understand the possible vulnerabilities of agent-like robotic systems whereas an experimental HRI study will be carried out to explore the human-centered vulnerabilities.

Project Contributions
My PhD project will contribute to social robotics by closing a gap in the current literature on trust in sHRI through an emphasis on vulnerability that is so far an under-researched dimension. Moreover, I will also contribute to human-computer interaction (HCI) research by addressing the tendency to neglect how strongly computer-based technologies can be perceived by humans as agent-like and by developing specific design guidelines for future design of trustworthy agent-like robotic systems.

Figure C.2: The project information sheet used for the expert interviews.

C.3 Informed Consent Form



TU WIEN Informatics

TRUST ROBOTS

CONSENT FORM		This consent form has been discussed with and overseen by: • **Dr. Marjo Rauhala**, Research Ethics Coordinator, TU Wien. • **Ao. Univ. Prof. Peter Purgathofer**, Data Protection Coordinator at the Faculty of Informatics, TU Wien.
Principle Investigator	**Glenda Hannibal**, BA MA University Assistant (Predoctoral) Phone: +43 (0)1 58801-193509 Email: glenda.hannibal@tuwien.ac.at	
Research Institution	TU Wien, Research Division of Human-Computer Interaction (E193/5) sekretariat@igw.tuwien.ac.at +43 (01) 58801-18703 Argentinierstraße 8 1040 Vienna Austria	
Project Title: Emphasizing Vulnerability		

Aim of interview
The aim of the expert interview is to gather knowledge about possible vulnerabilities of agent-like robotic systems in order to identify and map out these vulnerabilities using the initial categories: i) Robotic Hardware (physical embodiment and components), ii) Robotic Software (computational models and algorithms), iii) Security Issues (verification and authentication protocol).

Participation
I have been informed about the aim of the doctoral project via email on [DD/MM/YYYY] and during the short briefing prior to the interview. I have been able to ask questions that were answered to my satisfaction.

I have been informed that my participation involves:

- Being interviews regarding the specific topic "vulnerabilities of agent-like robotic systems".
- That my answers will be audio recorded and later transcribed for data analysis.
- That interview notes might be taken during the interview.

I have been informed that I can always wish to withdraw my consent to participate in the interview at any time without consequences.

Use of Data
I have been informed that the knowledge I provide during the interview can be used for:

- The PhD dissertation with the (premilitary) title: "Emphasizing Vulnerability: A theory-driven and interdisciplinary enquiry into trust in social human-robot interaction for the development of trustworthy agent-like robotic systems", planned to be submitted to the Faculty of Informatics at TU Wien.
- Publications associated with the PhD dissertation that have been submitted and presented at academic conferences, journals and workshops.

I have been informed that I always get in contact with the principle investigator to make any corrections or changes to the direct quotes before used for the PhD dissertation or associated publications.

(a) Page 1

I wish that my name, professional title and affiliation can be used for direct quotes in the PhD dissertation and associated academic publications.

Data Protection
I have been informed that the recorded interview and the transcription will be archived and stored on the TU Wien OwnCloud server so only researchers involved in the project at TU Wien have direct access to the collected data. The interview and transcription will be stored and used for a maximum period of 20 years after the interview was conducted.

I allow that other researchers at TU Wien associated with the project can be given permission to access and use the stored data for future research and teaching purposes.

Signatures
Based on the information sheet and the consent, I participate in the expert interview voluntarily and have been informed about my rights as a participant. I confirm that a signed copy of the consent form will be provided for my own keeping and future reference.

Interviewee ____________ Signature [DD.MM.YYYY] Date

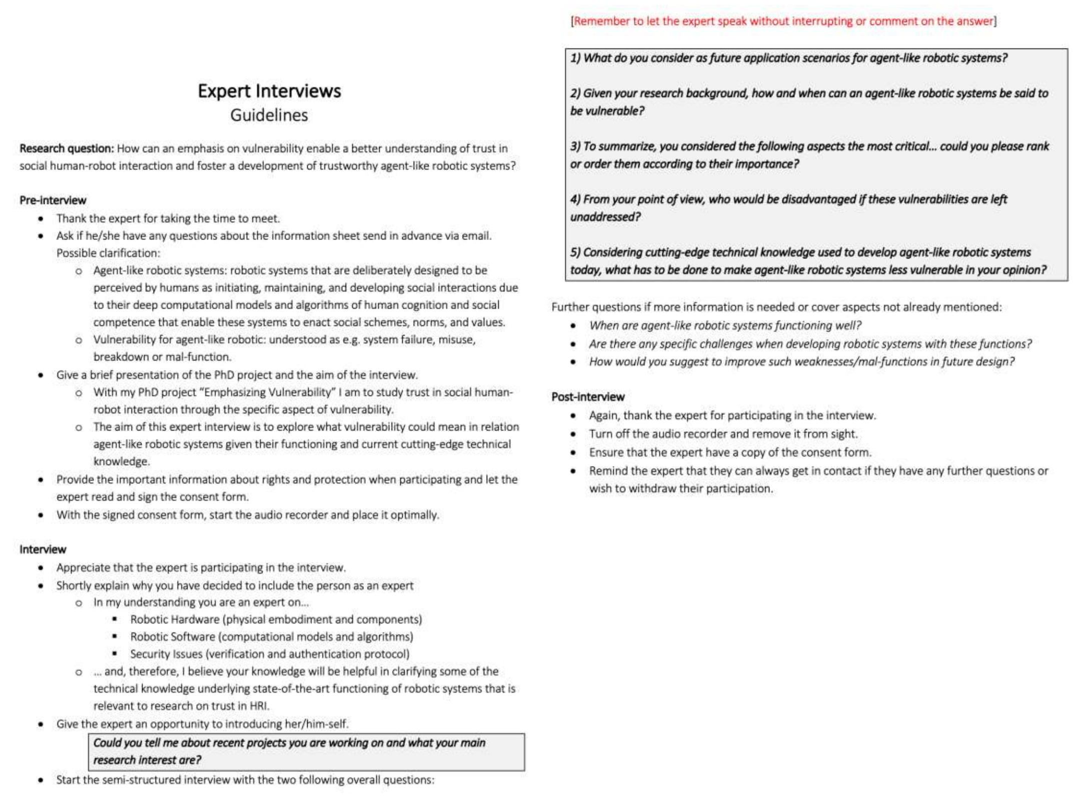
GLENDA HANNIBAL Interviewer ____________ Signature [DD.MM.YYYY] Date

Contact Information	
If you need further information, have any questions, or wish to withdrawal participation.	Glenda Hannibal Phone: +43 (0)1 58801-193509 Email: glenda.hannibal@tuwien.ac.at
If you wish to speak with an independent body about your rights as a participant.	Dr. Marjo Rauhala Research Ethics Coordinator at TU Wien Phone: +43 (1) 58801 43404 Email: marjo.rauhala@tuwien.ac.at

(b) Page 2

Figure C.3: The consent form all experts were required to sign before the interview could be conducted.

C.4 Interview Guideline



Expert Interviews
Guidelines

Research question: How can an emphasis on vulnerability enable a better understanding of trust in social human-robot interaction and foster a development of trustworthy agent-like robotic systems?

Pre-interview

- Thank the expert for taking the time to meet.
- Ask if he/she have any questions about the information sheet send in advance via email. Possible clarification:
  - Agent-like robotic systems: robotic systems that are deliberately designed to be perceived by humans as initiating, maintaining, and developing social interactions due to their deep computational models and algorithms of human cognition and social competence that enable these systems to enact social schemes, norms, and values.
  - Vulnerability for agent-like robotic: understood as e.g. system failure, misuse, breakdown or mal-function.
- Give a brief presentation of the PhD project and the aim of the interview.
  - With my PhD project "Emphasizing Vulnerability" I am to study trust in social human-robot interaction through the specific aspect of vulnerability.
  - The aim of this expert interview is to explore what vulnerability could mean in relation agent-like robotic systems given their functioning and current cutting-edge technical knowledge.
- Provide the important information about rights and protection when participating and let the expert read and sign the consent form.
- With the signed consent form, start the audio recorder and place it optimally.

Interview

- Appreciate that the expert is participating in the interview.
- Shortly explain why you have decided to include the person as an expert
  - In my understanding you are an expert on...
    - Robotic Hardware (physical embodiment and components)
    - Robotic Software (computational models and algorithms)
    - Security Issues (verification and authentication protocol)
  - ... and, therefore, I believe your knowledge will be helpful in clarifying some of the technical knowledge underlying state-of-the-art functioning of robotic systems that is relevant to research on trust in HRI.
- Give the expert an opportunity to introducing her/him-self.

  Could you tell me about recent projects you are working on and what your main research interest are?

- Start the semi-structured interview with the two following overall questions:

(a) Page 1

[Remember to let the expert speak without interrupting or comment on the answer]

1) What do you consider as future application scenarios for agent-like robotic systems?

2) Given your research background, how and when can an agent-like robotic systems be said to be vulnerable?

3) To summarize, you considered the following aspects the most critical... could you please rank or order them according to their importance?

4) From your point of view, who would be disadvantaged if these vulnerabilities are left unaddressed?

5) 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?

Further questions if more information is needed or cover aspects not already mentioned:

- *When are agent-like robotic systems functioning well?*
- *Are there any specific challenges when developing robotic systems with these functions?*
- *How would you suggest to improve such weaknesses/mal-functions in future design?*

Post-interview

- Again, thank the expert for participating in the interview.
- Turn off the audio recorder and remove it from sight.
- Ensure that the expert have a copy of the consent form.
- Remind the expert that they can always get in contact if they have any further questions or wish to withdraw their participation.

(b) Page 2

Figure C.4: The interview guidelines used for conducting the expert interviews.

Curriculum Vitae

Glenda ***HANNIBAL***

Email: glendahannibal@gmail.com
Webpage: glendahannibal.weebly.com
Twitter: @ordinary_robot

EDUCATION:

- MA in Philosophy, Aarhus University, 2015
  Specialization in Epistemology, Metaphysics and Cognition

- BA in Philosophy, Aarhus University, 2012
  Minor in Cognitive Semiotics

ACADEMIC EMPLOYMENT:

- Scientific Staff, Ulm University, 2022-2023
  Research Group on Explainable AI
  Institute of Artificial Intelligence

- University Assistant, TU Wien, 2018-2022
  Division of Human-Robot Interaction (HCI) + Trust Robots Doctoral College
  Institute of Visual Computing and Human-Centered Technology

- Expert, EU Commission, 2019-2020/2021-2022
  Human Behaviour and Machine Intelligence (HUMAINT) project
  Joint Research Center (JRC) for Advanced Studies/unit for Digital Economy

- University Assistant, University of Vienna, 2015-2017
  Research area of Culture and Knowledge
  Institute of Sociology

RESEARCH EXPERIENCE:

- Internship, Danish Technological Institute (DTI), 2014
  Center for Welfare and Interaction Technologies

- Student Volunteer, Aarhus University, 2012-2015.
  Research group "Philosophical and Transdisciplinary Enquiries into Social Robotics" (PENSOR), Department of Culture and Society – Philosophy

PUBLICATIONS:

- Hannibal, G., Dobrosovestnova, A. & Weiss, A. (2022). Tolerating Untrustworthy Robots: Studying Human Vulnerability Experience within a Privacy Scenario for Trust in Robots. Proceedings of the *31st IEEE International Conference on Robot & Human Interactive Communication* (pp. 821-828). Naples, Italy: IEEE.
- Hannibal, G., Rabb, N. Law, T., & Alves-Oliveira, P. (2022). Towards a Common Understanding and Vision for Theory-Grounded Human-Robot Interaction (THEORIA). Proceedings of the *17th ACM/IEEE International Conference on Human-Robot Interaction* (pp. 1254–1257). Sapporo, Japan (online): IEEE Press.
- Hannibal, G., Weiss, A. & Charisi, V. (2021). "The robot may not notice my discomfort" – Examining the Experience of Vulnerability for Trust in Human-Robot Interaction. Proceedings of the *30th IEEE International Conference on Robot & Human Interactive Communication* (pp. 704-711). Vancouver, BC (online): IEEE.
- Dobrosovestnova, A., Hannibal, G. & Reinboth, T. (2021). Service Robots for Affective Labor: a Sociology of Labor Perspective. *AI & Society*, pp. 1-13.
- Hannibal, G. (2021). Focusing on the Vulnerabilities of Robots through Expert Interviews for Trust in Human-Robot Interaction. Proceedings of the *16th ACM/IEEE International Conference on Human-Robot Interaction* (pp. 288-293). Boulder, CO (online): ACM.
- Dobrosovestnova, A. & Hannibal, G. (2020). Working Alongside Service Robots: Challenges to Workplace Identity Performance. Proceedings of *Robophilosophy 2020 – Culturally Sustainable Social Robotics* (pp. 148-157), Frontiers in Artificial Intelligence and Applications Series, Vol. 335. Aarhus, Denmark (online): IOS Press.

- Dobrosovestnova, A. & Hannibal, G. (2020). Teachers' Disappointment: Theoretical Perspective on the Inclusion of Ambivalent Emotions in Human-Robot Interactions in Education, Proceedings of the *15th ACM/IEEE International Conference on Human-Robot Interaction* (pp. 471–480). Cambridge, UK (online): ACM.
- Hannibal, G. & Weiss, A. (Eds.) (2020). Envisioning Social Robotics: Current Challenges and New Interdisciplinary Methodologies. *Interaction Studies* [Special Issue], 21(1), pp. 1-6.
- Hannibal, G. & Lindner, F. (2018). Transdisciplinary Reflections on Social Robotics in Academia and Beyond, Proceedings of *Robophilosophy 2018 – Envisioning Robots in Society–Power, Politics, and Public Space* (pp. 23-27), Frontiers in Artificial Intelligence and Applications Series, Vol. 311. Vienna, Austria: IOS Press.
- Weiss, A. & Hannibal, G. (2018). What Makes People Accept or Reject Companion Robots? A Research Agenda, Proceedings of the *11th Pervasive Technologies Related to Assistive Environments Conference* (pp. 397-404). Corfu, Greece: ACM.
- Hannibal, G. (2016). Bringing the Notion of Everyday Life Back to the Center of Social Robotics and HRI, Proceedings of *Robophilosophy 2016 – What Social Robots Can And Should Do* (pp. 67-75), Frontiers in Artificial Intelligence and Applications Series, Vol. 290. Aarhus, Denmark: IOS Press.
- Bertel, L. & Hannibal, G. (2015). The NAO robot as a Persuasive Educational and Entertainment Robot (PEER) - a case study on children's articulation, categorization and interaction with a social robot for learning. *Læring & Medier*, 8(14), 22 pages.
- Hannibal, G. (2014). 'Dynamic' Categorization and Rationalized Ascription: A Study on NAO, Proceedings of *Robophilosophy 2014 – Sociable Robots and the Future of Social Relations* (pp. 343-347), Frontiers in Artificial Intelligence and Applications Series, Vol. 273. Aarhus, Denmark: IOS Press.

AWARDS/PRIZES/GRANTS:

- Kurzfristige Wissenschaftliche Arbeiten (KUWI) grant (3.130€), 2020. International Office, TU Wien
- Best Paper Award (workshop at PETRA'18) + Prize (1.000€), 2018. Arab-German Young Academy of Sciences and Humanities.

- Financial Support for Research Activities (850€), 2017. Institute of Sociology, University of Vienna.
- Financial Support for Research Activities (560€), 2017. Institute of Sociology, University of Vienna.
- Travel Stipendium (500€), 2015. Research Network for Transdisciplinary Studies in Social Robotics (TRANSOR), Aarhus University.

INVITED TALKS:

- Goethe University Frankfurt am Main, 10/2019. "From Trust to Vulnerability: Philosophical perspectives on trust in sHRI" at the international symposium *Diffracting AI and Robotics: Decolonial and Feminist Perspectives*. Frankfurt, Germany.
- Colombian School of Engineering Julio Garavito, 06/2018. "Using social robots in rehabilitation and healthcare - philosophical and sociological perspectives" at the international seminar on *Personal Robots for Rehabilitation* (SORR). Bogota, Columbia.

CONFERENCE PRESENTATIONS:

- "Tolerating Untrustworthy Robots: Studying Human Vulnerability Experience within a Privacy Scenario for Trust in Robots", RO-MAN'22. Naples, Italy.
- "Towards A Questions-centered Approach to Explainable Human-Robot Interaction", Robophilosophy'22 – Social Robots in Social Institutions. Helsinki, Finland.
- "'The robot may not notice my discomfort' – Examining the Experience of Vulnerability for Trust in Human-Robot Interaction", RO-MAN'21. Vancouver, BC (online).
- "Everyday Life Centered Approach (ELCA): Sociological Perspectives on Agent-like Robotic Systems" (with Astrid Weiss), STSgraz'21. Graz, Austria (Online).
- "Emphasizing Vulnenability: Investigating Conceptualizations of Trust in Human-Robot Interaction", SPT'21. Lille, France (Online).
- "Social Robotics and the Nature of Trust", DoRoTa Symposium at the AISB'21 Convention. London, United Kingdom (online).

- “Co-working with Social Service Robots” (with A. Dobrosovestnova), Robophilosophy’20 – Culturally Sustainable Social Robotics. Aarhus, Denmark (online).
- “Teachers’ Disappointment: Theoretical Perspective on the Inclusion of Ambivalent Emotions in Human-Robot Interactions in Education” (with A. Dobrosovestnova), HRI’20. Cambridge, UK (online)
- “Acknowledgment of Workers in a Robot-Supported Society” (with S. Andersen), Robophilosophy’18 – What Robots Can and Should Do. Vienna, Austria.
- “What Makes People Accept or Reject Companion Robots? A Research Agenda”, Social Robots Workshop at PETRA’18. Corfu, Greece.
- “Everyday life centered approach to social robotics”, ESA’17. Exeter, United Kingdom.
- “Brining the notion of everyday life back to the center of social robotics and HRI”, Robophilosophy’16 – What Robots Can and Should Do. Aarhus, Denmark.
- “The invitation from social robotics to philosophy”, 2016 Annual conference of the Danish Philosophical Society. Copenhagen, Denmark.
- “How should we conceptualize social robots?”, 2015 Annual conference of the Danish Philosophical Society. Aarhus, Denmark.
- “Dynamic’ categorization and rationalized ascription: a study on NAO”, Robophilosophy’14 – What Robots Can and Should Do. Aarhus, Denmark.

WORKSHOP PRESENTATIONS:

- HRI’21, “Emphasizing the relation between trust and vulnerability”, position paper for the TRAITS workshop. Boulder, CO (online).
- HAI’18, “From mere reliance to the human condition”, extended abstract for the Measuring and Designing Trust (MDT) workshop. Southampton, UK.

TEACHING EXPERIENCE:

- Lecture (WS2022/23) – “Explainable AI” (BA/MA level), Institute of Artificial Intelligence, Department of Computer Science, Ulm University.

- Lecture (WS2022/23) – "AI Ethics" (BA/MA level), Institute of Artificial Intelligence, Department of Computer Science, Ulm University.
- Guest Lecture (SS2022) – "Trusting Robots in a Privacy Scenario" (PhD level), Karl Popper Kolleg (KPK) for Responsible Safe and Secure Robotic Systems Engineering (SEEROSE), University of Klagenfurt.
- Guest Lecture (SS2022) – "Humans in the Loop: Robots, Automation and Humanities Research" (MA level), Research Unit for Human-centered AI, Department of Communication and Psychology, Aalborg University.
- Guest Lecture (WS2020) – "Ethics and Philosophy" (MA level), Center for Human-Computer Interaction, Department of Computer Sciences, University of Salzburg.
- Guest Lecture (WS2020) – "Methodological Approaches" (MA level), Division for Gender Studies, Department of Social and Cultural Anthropology, University of Vienna.
- Guest Lecture (WS2019) – "Theories and Methodics" (MA level), Division for Gender Studies, Department of Social and Cultural Anthropology, University of Vienna.
- Guest Lecture (SS2019) – "Ethics in HRI" (MA level), Complex Dynamical Systems group, Department of Automation and Control (ACIN), TU Wien.
- Lecture (WS2017/18) – "Knowledge, Everyday Culture and Society" (MA level), Division for Culture and Knowledge, Department of Sociology, University of Vienna.
- Lecture (SS2016 & WS2017/18) – "Diagnosis of Society – Social Robotics" (BA level), Division for Culture and Knowledge, Department of Sociology, University of Vienna.
- Guest Lecture (WS2016/17 & WS 2015/16) – "Material Culture" (MA level), Division for Culture and Knowledge, Department of Sociology, University of Vienna.
- Teaching Assistant (SS2016) – "Social Constructivism as Paradigm?" (PhD level), Division for Culture and Knowledge, Department of Sociology, University of Vienna.

ACADEMIC SERVICE:

- Organizing Committee (OC) Member: Social Media Chair at HRI’22.
- Program Committee (PC) Member: CEPE/IACAP’21.
- Workshop Organizer: Robophilosophy’18 (“Transdiciplinary Reflections on Social Robotics in Academia and Beyond”), HRI’22 (“Towards a Common Understanding and Vision for Theory-Grounded HRI (THEORIA)”).
- Manuscript Reviewer: Interaction Studies, International Journal of Social Robotics (SORO/IJSR), Transactions on Human-Robot Interaction.
- Paper Reviewer: HRI’17/HRI’20/HRI’21), ICSR’19, CEPE/IACAP’21.
- Book Reviewer: Routledge (section on sociology).

PUBLIC OUTREACH:

- Public Presentation, 2021. “Trust in HRI: Probing Vulnerability as an Active Precondition”, Talking Robotics seminar (Online).
- Public Presentation, 2017. “Social Robotics and Visual Narratives in Tandem” (with Solmaz Farhang), Angewandte Innovation Laboratory at the University of Applied Arts Vienna.
- Pilot Project, 2017. “Visual Narrative and Social Robotics in 360 degree: a dialogue between academic research and art” (with Solmaz Farhang), WTZ Ost at the University of Vienna.
- Blog post, 2017. “Robot-supported health”, Contribution to the Semester Question “Health from the Laboratory – How?”, University of Vienna.
- Blog post, 2016. “When an interface becomes the face”, Contribution to the Semester Question “How do we live in a digitalized future?”, University of Vienna.

CERTIFICATES:

- Basic Qualification for Junior Staff – Teaching in Higher Education, 10/2020. Granted by the Faculty of Social Sciences at the University of Vienna.
- Ethical Conduct for Research Involving Humans (TCPS 2), 06/2018. Granted by the Canadian Panel on Responsible Conduct of Research (PRCR).