TRUST IN INTERDEPENDENT AND TASK-ORIENTED HUMAN–COMPUTER COOPERATION

PHILIPP KULMS

A dissertation submitted to Bielefeld University
in partial fulfillment of the requirements for the degree of

Doktor der Naturwissenschaften (Dr. rer. nat.)

Trust in interdependent and task-oriented human–computer cooperation

Philipp Kulms
Social Cognitive Systems Group
Faculty of Technology
Bielefeld University

THESIS COMMITTEE:
Prof. Dr.-Ing. Stefan Kopp (Bielefeld University)
Prof. Dr. Catherine Pelachaud (French National Centre for Scientific Research, Paris)
Prof. Dr.-Ing. Britta Wrede (Bielefeld University)
Dr. phil. Matthias Hartung (Bielefeld University)

DATE OF SUBMISSION:
January 11, 2018

DATE OF DEFENSE:
September 14, 2018

The paper used in this publication meets the requirements for permanence of paper for documents as specified in ISO 9706.

And, you know,
I never wanted to be a singer

— Phil Collins

ABSTRACT

This thesis presents a new paradigm for the modeling of cooperative human–computer interaction in order to evaluate the antecedents, formation, and regulation of human–computer trust. Human–computer trust is the degree to which human users trust computers to help them achieve their goals, and functions as powerful psychological variable that governs user behavior. The modeling framework presented in this thesis aims to extend predominant methods for the study of trust and cooperation by building on competent problem-solving and equal goal contributions by users and computers. Specifically, the framework permits users to participate in interactive and interdependent decision-making games with autonomous computer agents. The main task is to solve a two-dimensional puzzle, similar to the popular game Tetris. The games derived from this framework include cooperative interaction factors known from interpersonal cooperation: the duality of competence and selfishness, anthropomorphism, task advice, and social blame.

One validation study (68 participants) and four experiments (318 participants) investigate how these cooperative interaction factors influence human–computer trust. In particular, the results show how trust in computers is mediated by warmth as universal dimension of social cognition, how anthropomorphism of computers influences trust formation over time, and how expressive anthropomorphic cues can be used to regulate trust. We explain how these findings can be applied to design trustworthy computer agents for successful cooperation.

PUBLICATIONS

Some results, figures, and tables have appeared previously in the following publications:

Buchholz, V., Kulms, P. & Kopp, S. (2017). It's (not) your fault! Blame and trust repair in human-agent cooperation. In: *Proceedings of the 2017 Workshop on Cognitive Systems*. München.

Kulms, P. & Kopp, S. (2016). The effect of embodiment and competence on trust and cooperation in human–agent interaction. In: *Intelligent Virtual Agents*. Springer. Berlin, Heidelberg, 75–84.

— (2018). A social cognition perspective on human–computer trust: The effect of perceived warmth and competence on trust in decision-making with computers. *Frontiers in Digital Humanities, 5*, 14.

Kulms, P., Welbergen, H. van & Kopp, S. (2014). Prototyping von intuitiven und interaktiven Benutzerschnittstellen: Schnelles und einfaches Design von Anwendungen mit virtuellen Agenten [Prototyping of intuitive and interactive interfaces: Easily designing virtual agent applications]. In: *Technische Unterstützungssysteme, die die Menschen wirklich wollen*. Ed. by W. Robert & R. Tobias. Hamburg: Helmut-Schmidt-Universität, 30–38.

Kulms, P., Mattar, N. & Kopp, S. (2015). Modeling decision-making in cognitive systems for social cooperation games. In: *Proceedings of the 2015 Workshop on Cognitive Systems*. Bielefeld.

— (2016). Can't do or won't do? Social attributions in human–agent cooperation. In: *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 1341–1342.

Kulms, P., Welbergen, H. van & Kopp, S. (2018). MultiPro: Prototyping Multimodal UI with Anthropomorphic Agents. In: *Mensch und Computer 2018 - Tagungsband*. Ed. by R. Dachselt & G. Weber. Bonn: Gesellschaft für Informatik e.V., 23–32.

Mattar, N., Welbergen, H. van, Kulms, P. & Kopp, S. (2015). Prototyping user interfaces for investigating the role of virtual agents in human–machine interaction. In: *Intelligent Virtual Agents*. Springer. Berlin, Heidelberg, 356–360.

ACKNOWLEDGMENTS AND FUNDING

This work would not have been possible without the support from my colleagues of the Social Cognitive Systems Group.

I owe special thanks to Nikita Mattar for helping me with various crucial tasks, including developing a bigger idea behind a simple puzzle game, setting up the framework, planning and discussing the experiments, and searching for students to continue my work with.

I am deeply grateful to my supervisor, Stefan Kopp, who helped me through this journey with patience, trust, and his trademark sense of humor.

Thank you, Nicole and Jonathan, for helping me with my first steps in the human–computer interaction field.

Over the past seven years, I learned to appreciate the discussions with other members of the Intelligent Virtual Agents community, in particular, Peter Khooshabeh.

Finally, I wish to thank my students, Victoria Buchholz and Markus Gaffke, with whom I could discuss novel ideas and approaches and for working with me on the framework.

This thesis was supported by the German Federal Ministry of Education and Research (BMBF) within the Leading-Edge Cluster 'it's OWL' which was managed by the Project Management Agency Karlsruhe (PTKA), as well as by the Deutsche Forschungsgemeinschaft (DFG) within the Center of Excellence 277 'Cognitive Interaction Technology' (CITEC).

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

1.1 MOTIVATION

Of all our social relationships, we prefer individuals we can trust. Trust, "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" (Lee & See, 2004, p. 54), is one of the most powerful psychological tools at our disposal. The possibility to separate friend from foe makes trust inseparable from social interactions. Likewise, trust has grown to become a key psychological variable in human–computer interaction (HCI) that is just as important as computer reliability for successful human–computer teams (Dzindolet et al., 2003).

Today, people's activities often rely on intelligent computer agents, from smart devices and digital assistants to early social robots. Those technologies are increasingly endowed with the requirements necessary for complex cooperative interactions, such as problem-solving performance and autonomy. Performance-based capabilities are a crucial source of information based on which users trust agents (Hancock et al., 2011; Lee & See, 2004). Other essential capabilities relate to an agent's interface that was engineered for the users to interact with. Social robots and 3D virtual agents have expressive anthropomorphic interfaces, enabling them to mimic human appearance and/or behavior in order to make the interaction feel natural to users and influence them in profound social ways.

Can the mechanisms behind trust be utilized to develop technology that people can work together with, similar to how they cooperate with other humans? Computers are no longer used as mere tools. Rather, humans and computers increasingly cooperate toward complex interaction goals through a series of competent goal-directed actions (see Fig. 1). In the future, both users and agents could be equally equipped to contribute to the goal, decreasing functional differences between the actors, yet increasing the need to understand how users trust such agents.

Designing for trust is an ambitious task. This thesis focuses on the psychological problem of *appropriate trust*, that is, a match between perceived and actual technology capabilities (Lee & See, 2004). This perspective needs to encompass human-centered experiential and behavioral factors of trust as well as significant developments of intelligent computer agents. Computer agents are ubiquitous, occupy a wide range of roles, efficiently align to user needs, provide recom-

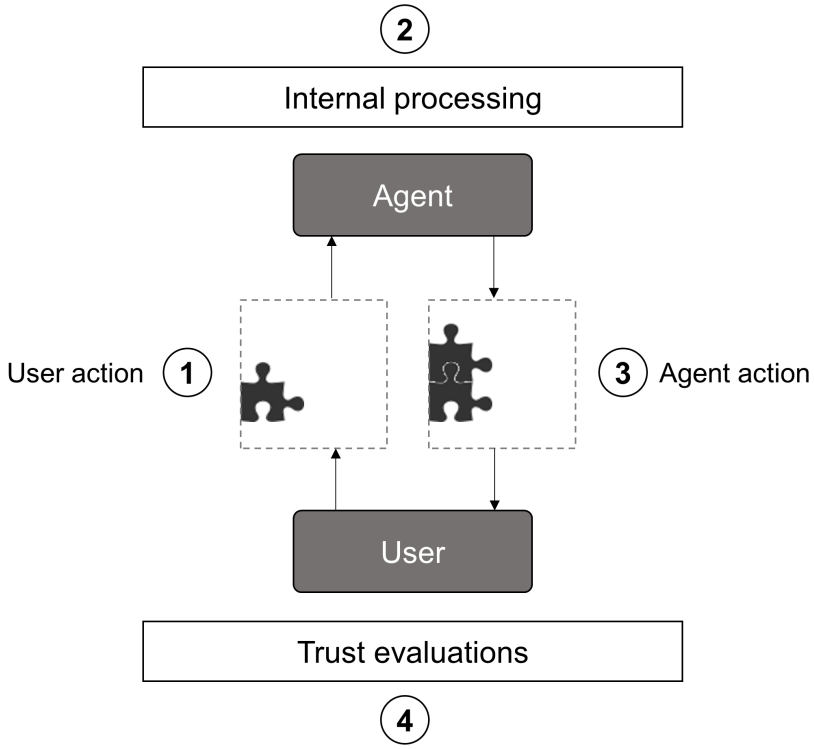


Figure 1: Schematic view of human–computer cooperation.

mendations, and permit natural communication. This has major implications for appropriate trust:

1. Technology is increasingly endowed with *anthropomorphic* characteristics: Speech-based interfaces enable interactions based on natural dialog and embodied agents such as social robots and virtual agents have deliberate human-like appearance. This enables designers to implement seemingly social behaviors which influence user responses in meaningful ways. It is, however, still mostly unclear if users trust anthropomorphic technologies like they trust other humans.

2. *Cooperative* computer agents allow for more open interactions toward complex goals, exceeding classic interaction metaphors such as the desktop paradigm. Future systems can engage in collaborative interactions that involve shared problem-solving and users being no longer fully in charge of the task. As users pass responsibility to agents, they need to be able to assess the agent's competencies and intentions to contribute to the goal, leading to a higher need for trust.

These two developments suggest users now can (or must) resort to previously unavailable cues to develop trust as HCI gradually evolves toward cooperative interaction. The following characteristics describe novel challenges users are faced with in cooperative interaction:

- **Complexity:** The interaction involves a myriad of degrees of freedom. Single actions or events not necessarily follow a script or pattern.
- **Interdependency:** The actions of the involved agents depend on each other. In order for them to be mutually supportive, task-related coordination and communication is required.
- **Goal-directedness:** The actions are geared toward a specific and observable or measurable outcome.
- **Continued interaction:** In order to achieve goals, repeated actions over an extended period of time are necessary. The trusting relationship is dynamic, it develops and changes.
- **Decision-making:** Complex decisions require the coping with uncertainty and risk, and trust is one of the most powerful mechanisms users have at their disposal to handle this issue. Additionally, social decision-making involves decisions made by users and other seemingly social agents.

On the one hand, the above considerations make appropriate trust all the more important. Untrustworthy agents pose various obstacles and risks. Much like untrustworthy web services, they can have malevolent intentions and try to exploit users. They can also be too incompetent to engage with or too incompetent to separate credible from incredible information, which poses a threat to the interaction goal. Both intention- and competence-related issues imply non-cooperativeness, yet their relative importance depends on whether the cooperative framing involves strategic elements and may thus offer exploitation. Extreme examples of the need for appropriate trust (and pitfalls of modern-day HCI) often contain reports of unfortunate and fatal transportation catastrophes caused by inappropriate reliance.[1]

On the other hand, it becomes more and more difficult to establish appropriate trust in the first place. State-of-the-art algorithms for learning and user modeling make it seem as if computers *really* know their users, although in the social sense, they certainly do not. Cognitive biases distort expectations and authority attributions toward computers (Parasuraman & Manzey, 2010). Interface designers use simple but effective techniques to persuade and motivate users in order to evoke specific behavioral reactions and influence their attitudes and emotions (Fogg, 2003). Research on nonverbal and multimodal behavior takes another step forward by demonstrating how robust trustworthiness cues can be identified and mapped onto anthropomorphic agents (DeSteno et al., 2012; Lee et al., 2013; Lucas et al., 2016).

1 Those reports are not part of the present thesis.

The implications of intelligent agents that approach cooperative problems in a human-like manner are still largely unknown, in particular, how people perceive, trust, and interact with them. Although the literature suggests that users form trust in computers (i.e., human–computer trust) differently than in other humans (e.g., Lee & See, 2004; Madhavan & Wiegmann, 2007), little is known about how anthropomorphic characteristics and cooperative interactions affect trust. Recent studies maintain that anthropomorphic agents could provide cues that help users to develop trust in them, but they viewed agents as rather passive decision support and enabled users to remain fully in charge of the cooperative task (Visser et al., 2016; Visser et al., 2017). In order to achieve trusting human–computer cooperation, the HCI community is in need for cooperative frameworks addressing such limitations.

1.2 CONTRIBUTIONS

This thesis adds to our understanding of human–computer trust and aims to show how cooperation with computers can be achieved. The approach behind this work is to design trustworthy computer agents that facilitate appropriate trust for cooperation. If an agent's intentions are dubious or its competencies are inadequate, user overtrust should be avoided. Likewise, if the agent's competencies are adequate, user distrust may cause an overall suboptimal output.

The underlying working hypothesis is that in order to establish perceived trustworthiness, agents must be perceived and understood by the user to be willing to cooperate and to have the competencies necessary to achieve cooperation. An extensive body of work at the intersection of HCI and decision-making research addresses this issue by adopting behavioral game theory as theoretical foundation and cooperation framework. This work demonstrated that the human tendency to trust and cooperate with computers is modulated by contextual cues such as anthropomorphism, but it is grounded in simplistic assumptions about cooperation that ignore task competence.

The methodological **goal** of this thesis is to extend decision-making scenarios known from behavioral game theory toward more interactive problem-solving, which requires competent action on both sides. The central element is a newly developed interaction framework that has humans and agents contribute equally to the goal. Based on this, the empirical goal is to analyze how contextual and trustworthiness factors such as anthropomorphism and competence relate to human–computer trust. Three research questions will be investigated in the empirical part:

1. Trust antecedents: How are fundamental dimensions of social cognition related to trust in computer agents?

2. Trust formation: What is the effect of anthropomorphism on the formation of trust in computer agents?

3. Trust regulation: Are anthropomorphic cues suitable for the regulation of trust?

The underlying interaction paradigm of the present work is cooperation. Cooperative relations are based on effective communication, coordinated efforts, division of tasks, mutual agreements regarding conflicting interests, and positive (cooperative) attitudes (Deutsch, 2011). In HCI, these factors are often largely approximated. Endowing computers with an understanding of naturalistic cooperation is a major ongoing challenge. Cooperation is often conceptualized using cooperative games such as the one-shot or repeated prisoner's dilemma, which give access to low complexity and standardized interaction frameworks. The contribution of the present approach is an interaction paradigm to investigate trust in more naturalistic cooperation.

Several **experiments** were carried out to analyze trust within the cooperative paradigm. The first experiment focuses on trust antecedents in strategic cooperation. It attempts to evoke universal perceptions of social cognition by modeling two dimensions of strategic decision-making. These perceptions are then used to predict trust in the agent.

The second and third experiment investigate trust formation in anthropomorphic versus non-anthropomorphic collaborators using an advice adoption scenario. Advice adoption is an important issue in HCI because people are increasingly challenged with establishing trust in computer-generated recommendations.

Since trust is a desirable albeit dynamic state of mind, people have developed mechanisms for its regulation in interpersonal relations. The fourth and final experiment aims at implementing social regulative behavior into an anthropomorphic agent in order to influence trust after a critical incident. Incidents such as joint goal failure are a threat to both trust and continued cooperation. Exploring mechanisms to regulate trust permits feasible cooperative solutions as opposed to discontinued interaction.

1.3 THESIS OUTLINE

A short overview of the structure of this thesis is given below.

Chapter 2: Trust: Theoretical background. This chapter explains the trust phenomenon, key factors of trust such as social risk and uncertainty, antecedents and specific implications for behavior.

Chapter 3: Trust in human–computer interaction. This chapter turns to the concept of human–computer trust. It explains the progress behind anthropomorphic technology and current methodological ad-

vances regarding the study of trust and cooperation in HCI. To put human–computer trust into perspective, models and frameworks that explain how technology socially influences users are explained.

Chapter 4: An interactive cooperation game paradigm for the investigation of trust in human–agent interaction. This chapter describes the present thesis' understanding of cooperation and explains how it extends predominant methodology. These considerations provide the motivation to develop a novel paradigm for the investigation of trust. The paradigm consists of a task as well as key interaction elements that are based on cooperative principles.

Chapter 5: Experiment 1. The first empirical chapter investigates the question if fundamental dimensions of social cognition are also related to trust in computer agents.

Chapter 6 and 7: Experiment 2 and 3. The following empirical chapters approach the potential effect of anthropomorphism on trust formation over time.

Chapter 8: Experiment 4. The last empirical chapter investigates the ability of an anthropomorphic agent to regulate trust.

Chapter 9: General discussion. The final chapter summarizes the main findings and discusses their implications for the relation between social cognition and trust, human–computer trust, requirements for human–computer cooperation, and design of trustworthy computer agents. The chapter discusses future research directions and limitations regarding the paradigm and experiments.

2

TRUST: THEORETICAL BACKGROUND

Trust has captured deep interest of various scientific disciplines, including sociology, social and applied psychology, HCI, economics, and game theory (Camerer, 2003; Corritore et al., 2003; Fukuyama, 1995; Gambetta, 1990; Lee & See, 2004; Mayer et al., 1995). One reason why trust is such a well-studied phenomenon is that trust is an indicator of how well interpersonal relationships work. We desire relationships to be of trusting nature because mutual trust makes social exchanges significantly easier. Students need trust in their teachers, patients in doctors, buyers in sellers, clients in lawyers, passengers in drivers, children in their parents.

This chapter outlines the nature (Section 2.1), antecedents (Section 2.2), behavioral consequences (Section 2.3), and regulating factors of trust (Section 2.4). The role of fundamental dimensions of social cognition – warmth and competence – for trust is introduced in Section 2.5. By focusing on the psychological perspective of trust, one goal of this chapter is to describe how close the relationship between standard trust models and recent theorizing on social cognition is.

2.1 THE GENERAL CONCEPT

Trust, "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" (Lee & See, 2004, p. 54), has broad and significant implications for the human social life. Trust is a defining characteristic of interpersonal behaviors such as assistance, care, cooperation, avoidance, as well as different types of relationships, including personal, professional, romantic, and human–computer relationships. Trust is relevant in nearly every, if not all types of social interaction along different degrees of immediacy, intimacy, and individual importance (Couch & Jones, 1997). It is based on cognitive and affective processes (McAllister, 1995), evolves over the course of a relationship (Rempel et al., 1985), and relies on specific interpersonal trait evaluations (Jones & George, 1998). The social sciences have brought forth several perspectives of trust, each with its own conceptualizations. The original branches focus on trust as general expectation or belief that is rooted deeply in personality, as institutional phenomenon, or as characteristic of economic and social exchange relationships (for an overview see Lewicki, 2006). Adding to this, more recent developments treat trust as a quantifiable variable in strategic decision-making (Berg et al., 1995; Camerer, 2003) or as factor that governs user behavior in

human–computer interaction (Hoffman et al., 2013; Lee & See, 2004; Muir, 1987).

The question why exactly people trust each other in the first place is explained in one of the most popular trust models by introducing another relevant concept: trustworthiness (see Fig. 2). According to this model, trust results from the assessment of the target agent's (i.e., the trustee) trustworthiness and the evaluating agent's (i.e., the trustor) general propensity to trust others (Mayer et al., 1995). Trustworthiness is an interpersonal quality which signals that the trustee can in fact be trusted. It is the combination of the perceived abilities and character (i.e., benevolence, integrity) of the trustee. Ability refers to a "can-do" and character to a "will-do" dimension of trustworthiness (Colquitt et al., 2007). In other words, in order to be trustworthy, the trustee must be able to act in an appropriate and competent manner, and she or he must be willing to do so.

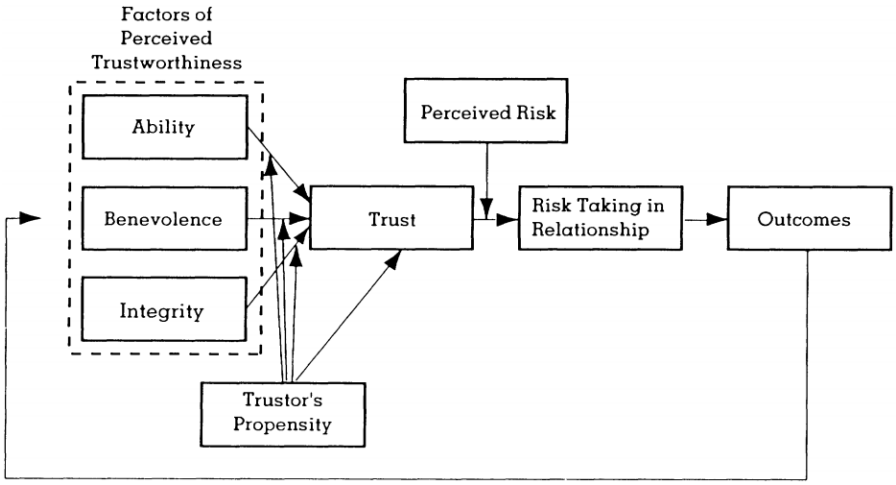


Figure 2: The Mayer et al. (1995) trust model. Trust is the outcome of an interplay of the trustee's perceived trustworthiness and the trustor's general propensity to trust others.

Current approaches view trustworthiness as the trustee's likelihood of influencing the trustor's goals, given an underlying conflict of interests (Balliet & Van Lange, 2013; Lewicki, 2006). Similarly, Hardin describes trust as encapsulated interest: "[...] I trust you because your interest encapsulates mine, which is to say that you have an interest in fulfilling my trust" (Hardin, 2002, p. 4). Since one can never be fully certain about the actual outcome of trusting relationships, the trustor voluntarily enters a state of risk and vulnerability by depending on the trustee (Lee & See, 2004; Mayer et al., 1995). Without uncertainty and the corresponding risk, trust is irrelevant and not necessary (Gambetta, 1990; McAllister, 1995).

Before turning to other characteristics, it is helpful to review the constructs that have been used interchangeably with trust (see Cor-

ritore et al., 2003; Hardin, 2002; Lewis & Weigert, 1985; Mayer et al., 1995).

- **Trustworthiness:** Trustworthiness is an aggregate of personal characteristics. The characteristics describe the commitment of fulfilling another's trust and thus, trustworthiness begets trust (Hardin, 2002). This commitment is determined by the trustee's ability, benevolence, and integrity with respect to the trustor's goals (Mayer et al., 1995). Moreover, the commitment can be reinforced by internal (e.g., moral obligations, benevolence) and external forces (e.g., norms, contracts) (Hardin, 2002).

- **Predictability:** Predictability is different from trust because being able to predict another agent's behavior does not imply one would accept being vulnerable and trust that agent (Mayer et al., 1995). Trust goes beyond predictability because trust entails positive assumptions that the agent is likely to act in one's favor. This is not the case for predictability.

- **Cooperation:** Trust and cooperation share a rather close connection which led to some conceptual confusion in the past. Cooperation does not necessarily require trust. People can readily cooperate without trust since cooperation does not always involve risk (Mayer et al., 1995). Today, there is little dispute over the fact that trust and cooperation are two different concepts: Cooperation involves the behaviors necessary to achieve a specific goal, whereas trust is an attitude toward social agents. Section 2.3 gives more details on trust and cooperation.

- **Credibility:** Credibility and trust also often appear together in the literature. However, credibility is synonymous with believability, according to Fogg and Tseng (1999). They side with the seminal Yale studies and view credibility as the resulting outcome of trustworthiness and expertise (Hovland et al., 1953). Thus, credibility is an individual characteristic.

In organizational settings, trust is a cognitive response elicited by an appraisal of reliable and dependable behavior shown by the target, or an affective response elicited by benevolence attributions of the target (McAllister, 1995). McAllister showed that cognition-based trust promotes affective trust, and he demonstrated behavioral consequences of affective trust. Specifically, affective trust led to the monitoring of others' needs and prosocial behavior.

The model by Lewicki addresses the multifaceted nature of professional and personal social relationships (Lewicki, 2006). Professional relationships, he argues, are mainly task-oriented. In contrast, the main focus in personal relationships is the social and emotional relationship itself as well as the individuals in the relationship. The model

distinguishes between calculus-based (CBT) and identification-based trust (IBT). First, CBT reflects the need to ensure consistent behavior and is enforced by basic economic calculations of rewards related to fulfilled expectations, and fears stemming from the possible punishment of trust violations. To gain CBT, people perform actions to prove their trustworthiness and systematically test each others' trust. This type of trust constitutes an early stage toward personal relationships. Second, as people begin to mutually identify with each others' desires and intentions, IBT is built. Lewicki describes this higher level of trust as a harmonious and synchronized state of profound interpersonal identification.

In another model on the qualities of close interpersonal relationships, trust evolves along with the maturing of the relationship (Rempel et al., 1985). In early stages, predictability, the extent to which behavior can be anticipated, is most relevant for trust. Predictability judgments are somewhat similar to ability/competence attributions (Lee & See, 2004) because perceived predictability is based on consistency evaluations over repeated observations and hardly requires intention inferences (Rempel et al., 1985). Instead, the accuracy of predictability judgments is proportional to the amount of past observations. Predictability is followed by dependability (reliability, honesty, integrity) attributions that focus on the target as an individual with certain qualities and characteristics. Here, evidence of predictability provides a foundation of dependability attributions. In the final stage, faith in the other facilitates trust in that it transcends the necessity of any past experience and permits the enduring of uncertainty and risk. This model is compatible with the view that trust is no static, stable, or single-point phenomenon. Rather, trust can be seen as phase-specific variable that changes over time and traverses several stages such as building, stability, and dissolution (Rousseau et al., 1998).

According to Jones and George (1998), the interpersonal interaction of values, attitudes, and moods/emotions is important for trust. These traits differentially influence the trust experience. Values determine the dimensions other social agents are evaluated on (e.g., predictability, reliability, competence, integrity) and individually affect trust by emphasizing certain qualities over others. Attitudes, the evaluative knowledge regarding specific objects that govern behavioral responses to the objects (Olson & Zanna, 1993), influence trust as they contain trustworthiness judgments. Moods and emotions are an integral part of the trust experience itself (see also Lee & See, 2004). As people sometimes engage in introspection to observe their own affective response to others, leading to initial trust or distrust, moods and emotions are also antecedents of trust. In contrast to values and attitudes, moods and emotions are less stable as they accompany changes in the trust experience.

Based on this review, the trust models share a couple of similarities:

- Trust is a way of handling imperfect knowledge in the context of interpersonal behavior and relations
- Trust can have cognitive and affective sources
- From a psychological point of view, trust can only be put in social agents capable of benevolent and malevolent intentions
- Trust is conceptualized along a single dimension ranging from high trust to high distrust (but see Lewicki, 2006)
- A crucial perceived quality that makes people accept the uncertainty and vulnerability associated with the trust put in others is trustworthiness

2.2 ANTECEDENTS OF TRUST

In order to trust a social agent to perform a specific action, one must assume the agent has the competence and intention to perform it (Deutsch, 1962). People are naturally motivated to categorize others as friend or foe based on their intentions (Fiske et al., 2007). In doing so, they generate an information base for deciding whom to approach or avoid, and in which cases cooperation may be a smart choice. This is important for understanding the meaning of trustworthiness for trust. Trustworthiness is essentially a decision aid: "[...] we cognitively choose whom we will trust in which respects and under which circumstances, and we base the choice on what we take to be 'good reasons,' constituting evidence of trustworthiness" (Lewis & Weigert, 1985, p. 970). Trustworthiness, the "willingness to be conditionally cooperative" (Boone & Buck, 2003, p. 165), is a quality one is ascribed by others and explains why agents are trusted (Mayer et al., 1995). Some argue that trust and trustworthiness are inseparable. Trust encapsulates trustworthiness evaluations, and trusting someone who is not trustworthy is maladaptive in the context of social relations (Hardin, 2002). Perceiving social agents as trustworthy leads to trust and emotional attachment (Colquitt et al., 2007).

Trustworthiness is promoted by behavioral factors and anthropomorphic cues. Goal-oriented behavioral factors that contribute to trustworthiness are perceived ability, benevolence, and integrity (Mayer et al., 1995). Ability refers to the skills, competencies, and characteristics allowing an agent to exert influence in a specific domain. Benevolence describes the degree to which the trustee wants to do good to the trustor and suggests some sort of attachment between the two agents. Integrity is the extent to which both agents' principles and moral attitudes are aligned. Ability is argued to be a cognitive cue of trust, benevolence and integrity are emotional cues (Dunn, 2000). Overall, an agent worth trusting shows reliable and predictable task performance, has a positive orientation to others' problems as well

as goals by complying with the goal-oriented purpose its competencies are attributed with, and adheres to integrity expectations arising throughout the problem-solving process (Lee & See, 2004). These requirements underline that in contrast to trust in mere actions, trust in agents grows if the trustor attributes trustworthy qualities to the trustee.

Anthropomorphic cues play an important role for trustworthiness. In particular, facial appearance (Oosterhof & Todorov, 2008; Todorov, 2008; Winston et al., 2002), nonverbal behavior (DeSteno et al., 2012), male facial width (Stirrat & Perrett, 2010), and facial emotion displays (Boone & Buck, 2003) were shown to be involved in the formation of trustworthiness. Judging the trustworthiness of neutral faces is a valence evaluation, that is, facial trustworthiness describes how positive or negative target faces are perceived. This is important for inferring the target's intentions because it enables the efficient identification of approach/avoidance signals in the form of happiness and anger features (Todorov, 2008). Researchers uncovered the relation between facial features and trustworthiness using a facial computer model that permits the exaggeration of facial dominance and trustworthiness (Oosterhof & Todorov, 2008). In the first step, open descriptions of human faces were collected and condensed into specific traits, including attractiveness, unhappiness, and sociability. Another group of participants then rated the faces on those traits. Principal component analyses revealed two underlying dimensions of face evaluation: valence, describing attractive and non-aggressive faces, and dominance, describing dominant, aggressive, and confident faces. Trustworthiness judgments correlated highly and reliably with the valence dimension. Finally, the authors constructed a computational model of how faces vary on dominance and trustworthiness, and randomly generated 300 emotionally neutral faces. It was found that linearly extrapolating the facial shapes along the previously uncovered trustworthiness dimension induced a shift from perceived anger (low trustworthiness) to happiness (high trustworthiness) among participant ratings (Oosterhof & Todorov, 2008).

Expressing one's desires and intentions through emotion in an accurate fashion is a trait known as emotional expressivity. With this ability, social agents communicate their trustworthiness and willingness to cooperate (Boone & Buck, 2003). Accordingly, people rely heavily on the judgment of facial expressions to judge honesty in face-to-face negotiation (Lucas et al., 2016). This suggests that the encoding and decoding of trustworthiness is an emotional process going beyond the mere signaling of approach and avoidance (Boone & Buck, 2003) as, indeed, the decoding involves brain regions dedicated to processing emotions (Winston et al., 2002). Moreover, the ability to decode even subtle facial cues is advantageous in situations that require trust decisions (Krumhuber et al., 2007). In a similar way, the

combination of certain nonverbal behaviors related to approach and avoidance intentions (e.g., leaning away from the counterpart, crossing arms) were found to predict trustworthiness judgments (DeSteno et al., 2012). It is necessary to recognize the context-free environment of some of these studies (for exceptions see DeSteno et al., 2012; Krumhuber et al., 2007; Lucas et al., 2016), implying that behavioral responses underlie the context-dependent activation of evaluative dimensions (Oosterhof & Todorov, 2008).

In a survey study, researchers analyzed the role of liking for trust in the relationship between buyers and sales reps (Nicholson et al., 2001). Liking is seen as affective trust antecedent that creates personal attachments. The attachment or bond is driven by common interests, a shared outlook, and frequency of interactions. However, over time these factors become less important for trust as the attachment itself, once established, is now the primary antecedent. The results show that liking mediates the effect of cognitive antecedents on trust, that is, value similarity and frequency of interactions. Moreover, they indicate that trust can indeed become more affect-based over time.

To summarize, liking and trustworthiness are key elements to understand how trust is formed. From an evolutionary point of view, individuals adapting to identify benevolent (trustworthy) and malicious (untrustworthy) intentions in others over the course of a social categorization process hold a crucial advantage and can adjust their behavior accordingly, including the decision to trust or not (Fiske et al., 2007). Going beyond this, emotion expressions which are hypothesized to signal trustworthiness increase trust and cooperation in social dilemmas (Krumhuber et al., 2007; Melo et al., 2014). Importantly, both the behavioral and anthropomorphic dimension highlight trustworthiness as an interpersonal quality.

2.3 BEHAVIORAL IMPLICATIONS OF TRUST: RISK TAKING, COOPERATION, RELIANCE

Like other psychological constructs, trust is a bridge between beliefs and behavior (Lee & See, 2004). The influence of trust on social behavior is explained by the notion of risk, the perceived probability of loss (Chiles & McMackin, 1996). Trust permits people to cope with the risk of being uncertain regarding the intentions and actions of others (Mayer et al., 1995; Rousseau et al., 1998). Risk taking, then, is making oneself vulnerable by the willingness to depend on others (Mayer et al., 1995), or by the willingness to accept non-reciprocated social investments in social exchanges (Shore et al., 2006). There is a reciprocal relationship between trust and risk taking: risk may lead to trust, which creates risk taking (Rousseau et al., 1998). Taken together, risk taking is a central consequence of trust. The behavioral manifes-

tation of risk taking takes many forms, and the amount of risk one is willing to take depends on the amount of trust (Mayer et al., 1995).

As explained earlier (see Section 2.1), the close connection between trust and cooperation has led to some conceptual confusion in the past. Today, researchers have established the important difference between both constructs: while trust is a psychological state, cooperation is some form of interpersonal interaction that may rely on trust but can also have more elementary foundations. Among those, initial cooperation between individuals followed by strict reciprocity has been shown to be fairly robust (Axelrod, 1984). Moreover, on this account, it is possible to explain cooperation as an evolutionary key mechanism between biological organisms (Brosnan et al., 2010; Trivers, 1971). These considerations aside, broad evidence suggests that cooperation is facilitated by trust, a finding that is well-studied in the context of social dilemmas (Balliet & Van Lange, 2013; Dawes, 1980). Social dilemmas are defined as a conflict between immediate self-interest and long-term collective interest (Van Lange et al., 2013). The underlying idea of social dilemmas is that one is better of with defecting, irrespective of the other individuals' choice, but taken together, all individuals would benefit from collective cooperation (Dawes, 1980). Under these conditions, the importance of trust for cooperation is amplified by the magnitude of conflicting interests between the individuals (Balliet & Van Lange, 2013). The link between trust and cooperation is also well-established in the context of organizations (Jones & George, 1998; Mayer et al., 1995; McAllister, 1995). In the organizational context, cooperation is usually understood as help, assistance, or teamwork. Jones and George suggest the form of cooperation depends on whether trust is conditional or unconditional (Jones & George, 1998). Conditional trust represents the minimum level of trust to facilitate social and economic exchanges toward a common goal, based on the prospect of future beneficial interactions. Unconditional trust, on the other hand, follows from shared values and confidence in the involved co-workers' trustworthiness. The latter fosters a type of cooperation characterized by high personal costs and self-sacrifice.

The relation between trust and cooperation is more complex than it may seem at first glance. For instance, it was found that people engage in unconditional cooperation to signal trustworthiness (Jordan et al., 2016). In economic decision-making, the degree to which participants cooperate unconditionally with a hypothetical target is increased if participants knew they would be observed. Unconditional cooperation occurred either if participants refused to take the chance to learn about the associated costs of cooperation, or quickly decided to cooperate after the cost was revealed. In the next phase, participants played the trust game with those who observed their earlier decision. If the previous decision was known to observers and if

participants cooperated unconditionally before, the observers trusted participants more by sending them more money. Finally, participants that cooperated unconditionally actually were more trustworthy than those who did not as they also returned more money to the observers in the same trust game. This suggests that reputation building is a central underlying motivation behind the relation between trust and cooperation. The researchers could also demonstrate that unconditional decisions serve the purpose of signaling trustworthiness for future cooperation and are not caused by the desirability of unconditional decisions in general. Furthermore, it was shown that cooperation is both consequence and antecedent of trust.

Trust is not only important in interpersonal interactions, it has implications for HCI as well. A well-established notion is that trust governs, but not completely determines reliance in computers. Lee and See developed a conceptual model to describe why people often rely on computers they trust and reject those they do not trust (2004). According to their model, the intention to rely on computers is influenced by human workload, self-confidence, perceived risk, and effort to engage. If the intention is formed, time constraints and configuration errors affect actual reliance. Lee and See identified a bidirectional relation between trust and reliance to the extent that reliance often determines trust. If a computer is not used, that is, relied on, trust is harder to grow because information regarding its capabilities remain limited.

2.4 TRUST REGULATION

Establishing trustworthiness and building trust do not occur in isolation but in the context of dynamic and complex social interactions. Trust processes can be regulated actively with the goal in mind to continue the interpersonal relationship. Lewicki explains how people typically regulate the different forms of trust he and his colleagues proposed (Lewicki, 2006). To manage calculus-based trust (CBT), the type of trust based on punishment, reward, and the associated fears and hopes, a mutual agreement among the involved parties regulates the functioning of the social exchange by negotiating shared expectations and monitoring each other's actions. In contrast, high-level forms such as identification-based trust (IBT) are managed by promoting a shared identity based on similar group memberships, interests, goals, values, and reactions to familiar situations. People are motivated to balance self-protection in the face of vulnerability on the one hand and the attainment of mutual goals on the other. A key for achieving this is to build sufficient CBT such that the handling of clear behavioral expectations is facilitated, and to acknowledge matters of distrust in order to define solutions prior to the emergence of conflict. Differentiating between basic exchange-based and higher

levels of trust hence suggests that disagreements pertaining to beliefs or personal values which stand in the way of IBT be separated from the foundation of calculus-based interactions (Lewicki, 2006).

A critical component of the dynamics of trust are trust violations. Trust violations impede the maintaining of trust in human–human (Lewicki & Bunker, 1996) and human–computer interactions (Muir & Moray, 1996). Both expectations, trusting (e.g., high competence) and distrusting (e.g., low competence) behavior, can be violated, although the former kind usually is more significant (Lewicki, 2006). Trust in computers typically diminishes when the system produces obvious, detectable errors. Errors contradict assumptions of trustworthiness, reliability, and accuracy and since they are better remembered, violate trust (Dzindolet et al., 2003). This process is seen as an affective response to violations or confirmations of implicit expectations (Lee & See, 2004). Muir and Moray found that trust is affected by both the magnitude and variability of a computer's competence (Muir & Moray, 1996). Their results also indicate that trust does not necessarily decline gradually with increased errors. They observed a steep decline at the beginning and a moderate decline later on. Moreover, a loss of trust is not always tied to degraded system performance but, being a fragile construct, may occur even if an error only momentarily affects performance, without an effect on overall system performance. However, evidence suggests the loss of trust after trust violations by computers can be regulated. For instance, rather than displaying erroneous output, providing users with continuous feedback regarding computer performance, or giving them a reason why it may produce an error increases the willingness to trust (Dzindolet et al., 2003). The ongoing development of trust may depend on initial assessments (Muir & Moray, 1996). Furthermore, distrust appears to be more resilient than trust (Muir & Moray, 1996; Muir, 1994).

There are explicit repair techniques to overcome trust impairments following from trust violations. In interpersonal relationships, apologies can help rebuilding trust, especially when they are carried out sincerely and the nature of the past relationship provides reasons to overcome strong violations with reconciliation (Tomlinson et al., 2004). The effectiveness of an apology versus denial is affected by the violation. Fine-tuned apologies are effective if mistrusted individuals apologize for competence-based violations but deny responsibility for integrity-based violations, and apologize before evidence of actual guilt is revealed but deny responsibility before they are found innocent (Kim et al., 2004). Aside from apologies, trust repair may require an extended period of time in order to permit the reassurance of reliability, hence explicitly addressing matters of distrust may provide short-term solutions, for instance by temporarily minimizing vulnerability (Lewicki, 2006). There are few studies that investigated the effect of computer apologies on the interaction. It was found that

people accept the notion of computer apologies and that they may be beneficial regarding users' moods and the interaction experience (Akgun et al., 2010; Tzeng, 2004). Robots that blame themselves for errors or the human–robot team are more likable than those that blame the user (Groom et al., 2010; Kaniarasu & Steinfeld, 2014). However, with respect to trust as outcome variable, the only study being available found no differences between a self-blaming, user-blaming, and team-blaming robot (Kaniarasu & Steinfeld, 2014).

2.5 THE ROLE OF FUNDAMENTAL DIMENSIONS OF SOCIAL COGNITION: WARMTH AND COMPETENCE

People often lack the time and cognitive resources to make thorough judgments in complex and dynamic social decision situations. Converging evidence shows that a large fraction of interpersonal judgments is located on two universally prevalent dimensions of social perception: warmth and competence (Fiske et al., 2007; Judd et al., 2005; Wojciszke, 2005). Warmth is associated with perceived trustworthiness, friendliness, empathy, and kindness, and competence is related to perceived intelligence, power, efficacy, and skill (Cuddy et al., 2011). In the past, the understanding of warmth conceptually approximated to trustworthiness (Campbell et al., 2001; Fletcher et al., 1999) or morality (Wojciszke, 1994), yet the underlying meaning remained the same. Specifically, warmth attributions reflect perceptions of the behavioral intentions of a social agent, while competence attributions pertain to perceived behavioral effectiveness (Fiske et al., 2007). Moreover, warmth judgments are based on perceived motives (Reeder et al., 2002) and they affect trust and doubt in others' motives (Cuddy et al., 2011). Unveiling the role of the warmth and competence framework for social cognition was particularly fruitful for research on behavior construal (Wojciszke, 1994), impression formation (Wojciszke et al., 1998), and stereotypes (Cuddy et al., 2008; Cuddy et al., 2007; Judd et al., 2005).

In a series of experiments involving impression formation based on vignettes and other textual materials, Wojciszke and colleagues demonstrated how warmth judgments shape social perception (Wojciszke et al., 1998). Warmth traits were more accessible than competence traits, more relevant for impressions in terms of prediction accuracy and weight, and the dominance of warmth traits was more pronounced for females. Competence attributions are easier to establish and maintain, but warmth attributions carry a potentially higher risk because the consequence of misinterpreting bad intentions can be severe (Cuddy et al., 2011). Warmth and competence are central underlying factors of stereotypes across cultures (Cuddy et al., 2008). Moreover, the combinations of high versus low warmth and com-

petence judgments determines emotional responses towards targets, such as admiration, contempt, envy, and pity (Cuddy et al., 2007).

What is the relation between those omnipresent judgments and trust? The psychological function of warmth and competence judgments is to explore one's own behavioral options. There is no dispute over the impact of warmth and competence on approach–avoidance behaviors, yet empirical findings as to how these judgments influence interpersonal behavior across different situations are mostly restricted to survey- and vignette-based studies (Asch, 1946; Cuddy et al., 2007; Wojciszke et al., 1998). For instance, warmth-based stereotypes are associated with active behaviors such as facilitation (high warmth) and harm (low warmth) (Cuddy et al., 2007). One of the common approximations of warmth in the literature is centered on trustworthiness and feelings of trust (Campbell et al., 2001; Fiske et al., 2007; Fletcher et al., 1999; Williams & Bargh, 2008). Recall that trustworthiness provides an information base to decide if one can trust another agent (see Section 2.2). Indeed, trustworthiness traits such as benevolence and integrity are congruent with traits occupying the warmth label, including fair, generous, helpful, honest, understanding (Wojciszke et al., 1998). In their review, Fiske and colleagues do not explicitly discuss the empirical evidence of the relation between perceived warmth, trustworthiness, and trust, but noted: "although one could quibble over separating or combining trust and warmth, there is a core linkage between the two features, with trust and warmth consistently appearing together in the social domain" (Fiske et al., 2007, p. 77). Trustworthiness is essentially seen as warmth trait, and perceived intent is the underlying psychological meaning of the combined traits, that is, "friendliness, helpfulness, sincerity, trustworthiness, and morality" (Fiske et al., 2007, p. 77). Conceptually, both warmth and competence as well as trustworthiness attributions are means for people to categorize perceived intentions and abilities. Importantly, trustworthiness captures both perceived intentions and abilities, not just perceived intentions. Based on this, reducing trustworthiness to warmth falls short of capturing performance-related abilities needed to actually achieve tasks toward cooperative goals. Consistent with the view that warmth and competence together involve perceived intentions and abilities, one possible solution toward theoretical integration is to understand trustworthiness as one of the *outcomes* of warmth and competence attributions. Given a lack of empirical evidence, this remains a proposition.

In sum, a better understanding of the relation between warmth and competence on the one hand and trust variables on the other hand could help explaining how people distinguish trustworthy from untrustworthy social agents and how this affects behavior. Specifically, this open question could show links between the implications of warmth and competence and the antecedents of trust and trust-

worthiness, and achieving this is practically relevant for understanding interpersonal interactions. For instance, people are more willing to cooperate with trustworthy agents in social dilemmas (Boone & Buck, 2003; DeSteno et al., 2012; Krumhuber et al., 2007) which is especially relevant in one-shot interactions that do not offer punishment (Janssen, 2008). Moreover, the capability to detect untrustworthy individuals is supported by perceptual and cognitive processes, but this mechanism appeared to be mitigated by how competence (i.e., status) affects warmth (i.e., attractiveness) (Mealey et al., 1996).

2.6 SUMMARY

Trust works like a social compass in a complex world filled with risky choices. Trust allows people to take risks and engage in cooperation with others. The issue of whom to trust is essentially a decision-making problem that is governed by cognitive and affective processes. In this process, trustworthiness is a trust antecedent. People evaluate the perceived abilities and intentions of others to infer their trustworthiness based on which trust develops. From a social cognition perspective, trustworthiness formation is similar to the evaluation of warmth and competence, although there is little empirical evidence of the precise relationship between warmth, competence, and trustworthiness. Anthropomorphic cues such as facial features play an important role for social cognition because the underlying dimensions of face evaluation, valence/trustworthiness and dominance, enable intention inferences. Since trust evolves over time and sometimes rapidly changes, there are mechanisms for its regulation in order to prevent the decline of trust and relationships altogether.

3

TRUST IN HUMAN–COMPUTER INTERACTION

People's lives are pervaded by technology. Automated machines, computer agents, and service robots help us achieve our goals in a myriad of ways. Digital assistants that are always within our reach recommend directions, travel targets, movies, new friends, and shopping products. Computers incrementally interpret, correct, and complete human input. Conversational agents chat with us in virtual environments and solve complex issues such as self-service and customer support. Industrial robots and assistance systems for manufacturing enable the supervision, maintenance, and control of large-scale industrial facilities, even from a distance. On top of that, autonomous vehicles could soon revolutionize public and private transportation.

Technology closely adapts to our needs, occupies entirely new roles, and performs complex cognitive tasks related to perception, planning, and problem-solving. Although the statement may not be true in the rational sense, computers are clearly perceived to be more than mere tools.

Human–computer trust is a special form of trust. Early on, trust has been recognized as a central human factor in HCI, but the advent of technological devices with intelligent problem-solving capabilities and human-like characteristics such as facial animation and natural language communication raises important questions as to how people develop trust in computer agents.

This second chapter on trust gives an overview of key aspects of human–computer trust research and pursues a central underlying question: **What are the characteristics of human–computer trust?**

Section 3.1 describes the relevant factors of human–computer trust. Section 3.2 describes how anthropomorphism fundamentally changes or understanding of HCI. Section 3.3 is devoted to current advances of human–computer interaction studies on trust and cooperation. Technology influences people's affective, cognitive, and behavioral processes in many ways. Section 3.4 shows how different theories, models, and frameworks explain the social influence of technology on human users.

3.1 FACTORS OF HUMAN–COMPUTER TRUST

The history of HCI is also a story of who is in charge, the computer or the human user. For instance, the progress behind decision support systems between the 1960s and 80s was appraised as a move away from the control of expert systems over users, toward users taking

over decision authority and responsibility (Muir, 1987). In retrospect it became obvious that computers had been viewed predominantly as mere tools because gradually, remarkable contrasts to this earlier view emerged. Computers took on a host of important roles across different domains at the workplace, became increasingly indispensable, and communicated with humans in novel ways. This fundamentally changed the way people interacted with them (Hoc, 2000; Nass et al., 1996). As a consequence, researchers began to systematically investigate if people treat computers in unanticipated ways. Indeed, it was revealed that users applied social norms to computers and show social responses such as gender-stereotyping (Nass et al., 1994).

Today, with the heterogeneous range of human–computer applications such as negotiation (Lin & Kraus, 2010), supervisory machine control (Sheridan & Parasuraman, 2005), long-term relationships (Bickmore & Picard, 2005), and cooperative teamwork (Bradshaw et al., 2012; Wissen et al., 2012), trust remains as a consistently important factor in HCI. Appropriate forms of trust prevent maladaptive human behaviors caused by distrust and overtrust (Lee & See, 2004). Trust regulation techniques are helpful in overcoming trust violations such as errors (Visser et al., 2016), allowing for feasible adjustments like trust calibration in contrast to more costly solutions like technology disregard. The idea of calibrated trust is a key element of automation research, a scientific branch concerned with how people use automated agents to achieve their goals. According to this view, calibrated trust is a match between user trust and actual technology capabilities and involves evaluations of trustworthiness and competence (Muir, 1987). Notably, automation research predicts that human–computer trust deviates from interpersonal trust and highlights the unique human responses to technology (Lee & See, 2004).

What are the key factors of human–computer trust? The literature suggests that the specific characteristics revolve around a number of concepts:

1. Trust antecedents
2. Focus and modes of trust
3. Trust dynamics and the processual character of trust
4. The relation between human–computer and interpersonal trust
5. Idiosyncrasies of the technology and nature of interaction

This list could be extended to include factors such as the organizational and cultural context of trust, or how information is processed in the context of trust formation, but for the sake of brevity we focus on the essential elements above.

I) TRUST ANTECEDENTS. A trust antecedent is information that informs a trustor about a trustee's competence to achieve the trustor's goals (Lee & See, 2004). In an effort to unify dimensions of trust, Lee and Moray proposed three general antecedents of trust: performance, process, and purpose (Lee & Moray, 1992). The performance dimension refers to expectations regarding the automation's reliability, predictability, and ability. Process denotes if the automation's characteristics are beneficial to the intent's and goals of the operator, and is similar to character traits or integrity. Purpose describes underlying motives or intents of the automation, making it similar to benevolence. These three types of goal-oriented information determine appropriate trust, that is, a level of trust that matches the degree to which the trustee (an automation device or a social agent) will help achieve the trustor's goals.

II) FOCUS AND MODES OF TRUST. The focus of trust refers to what exactly is to be trusted (Lee & See, 2004). The focus is defined by the level of detail, ranging from more general trust in an overall system to trust in particular functions and states of computer agents. Trust in computers also has different modes (Hoffman et al., 2013). Based on the limited experience with a technology and the difficulty to predict its behavior in novel situations, people resort to the specific modes of trust. For instance, default trust occurs in the absence of deliberative thought as to whether a computer can or cannot be trusted. Instead, default trust is used as a shortcut and potential concerns are ignored. Negative trust occurs as the general attitude that at some point, most technology will be faulty and complicate one's work. In contrast, absolute trust is a completely unconditional, yet context-dependent mode of trust. Absolute trust, as Hoffman and colleagues note, is more common between people. In HCI, the only example of absolute trust is negative trust. Usually, trust in technology is conditional upon tasks and contexts. It appears that people's tendency to resort to such different modes of trust is another manifestation of the need to handle complexity (Lee & See, 2004). Furthermore, over- and undertrust can be traced back to some of the modes of trust.

III) TRUST DYNAMICS AND ITS PROCESSUAL CHARACTER. Trust is no static or monotonically developing variable. To trust and attribute trustworthiness are dynamic processes, depending on context and goals, with unclear start and end points (Hoffman et al., 2013). Like interpersonal trust, trust in computers can be described along developmental dimensions such as formation and violation. Muir (1987) argued that trust in computers can be formed in a manner similar to the stages of interpersonal trust development identified by Rempel et al. (1985). However, the opposite pattern was also observed experi-

mentally: faith predicted trust in the beginning, followed by dependability, and then by predictability in the final stage (Muir & Moray, 1996). Thus the dynamics of trust also depend on information regarding the purpose of a system that is provided before the interaction, as well as documentation and training (Lee & See, 2004). Comparing how people form trust in computers versus other people, specifically, how trust in people builds on experiences gained over time, leads to an important caveat: trust in computers is developed in a rather unnatural fashion if people are forced into the interaction (Lee & See, 2004).

IV) THE RELATION TO INTERPERSONAL (HUMAN–HUMAN) TRUST. Interpersonal trust is determined by combined attributions of competence, benevolence, and integrity (Mayer et al., 1995). Research has shown that these factors are also relevant for trust in computer agents (Lee & Moray, 1992). Furthermore, trust-related HCI studies often build on trust definitions and models established in psychology and sociology (e.g., Barber, 1983; Mayer et al., 1995; Rempel et al., 1985; Rousseau et al., 1998). Going beyond this, research by Reeves and Nass suggests that human–computer interaction is similar to interpersonal communication as people respond to computer agents as if they were humans and apply social norms, self-/other-distinctions, and gender stereotypes to computers, even though people are aware of the fact that those reactions are inappropriate: "These social responses are not a function of deficiency, or of sociological or psychological dysfunction, but rather are natural responses to social situations" (Nass et al., 1994, p. 77). Because this rule applies to the social influence of computers in general, it can be deduced that interpersonal and human–computer trust are, in theory, similar.

In contrast, other HCI research provides evidence that people's behavior in human–computer interactions cannot be explained using interpersonal trust mechanisms (Hoffman et al., 2013; Lee & See, 2004; Visser et al., 2016). This applies to general expectations of expertise as well as behavioral responses: computer advice is assumed to be more objective and rational than human advice (Dijkstra et al., 1998), and people tend to have difficulties with bringing reliance on computer output into accordance with the output's actual reliability (Dzindolet et al., 2003). The mismatch between perceived and actual reliability is a key issue in HCI. If the human user finds a level of trust that is in line with system capabilities, his trust is well *calibrated*, but this is often not the case (Lee & See, 2004). One reason is that human–computer interactions are characterized by specific heuristics such as automation bias that people apply exclusively to computers. According to this phenomenon, people ascribe greater power and authority to computer-generated decision aids than other advice sources in order to reduce monitoring efforts (Parasuraman & Manzey, 2010).

v) IDIOSYNCRASIES OF THE TECHNOLOGY AND INTERACTION. The relevant factors of human–computer trust are often determined by defining properties of the technology itself such as the degree of automation and perceived autonomy (Lee & See, 2004; Parasuraman & Riley, 1997), anthropomorphism (Madhavan & Wiegmann, 2007; Visser et al., 2016; Waytz et al., 2014), and interface type (e.g., virtual agents, robots; Bickmore & Cassell, 2001; Hancock et al., 2011). Such idiosyncrasies contribute to the shift from tool-like interaction toward novel forms of human–computer interaction. They also impact how people form trust in computers.

Although computers are not truly autonomous because they lack self-determination, independence, and consciousness, people may attribute a certain intentionality to them (Lee & See, 2004). Perceived intentionality, in turn, could trigger attributions of loyalty, benevolence, and value (in-)congruence which are key dimensions of trust. Anthropomorphism plays an important role for trust formation in computers. This was demonstrated in a setting that had participants solve a simple pattern completion task while receiving advice from agents varying in anthropomorphism (Visser et al., 2016). As reliability of the advice decreased, the loss of trust was lower for anthropomorphic agents. This study also showed how specific behaviors that are fundamental to the shaping of interpersonal interactions such as social feedback (i.e., apologies, motivational remarks) further magnify the impact of anthropomorphism. Lastly, the interface of agents gives rise to novel interaction elements. Virtual agents can be designed to resemble humans, artificial entities, or even the user (Bergmann et al., 2012; Nowak & Biocca, 2003; Parise et al., 1996; Vugt et al., 2010), and participants can easily be led to believe the agent is controlled by another human or computer (Bailenson et al., 2003). This offers a whole variety of design elements to manipulate how people are socially influenced by virtual agents. Unlike virtual agents, robots are present in the physical environment, making them highly useful in dangerous situations. However, this also creates numerous challenges for human–robot interaction because crisis situations often involve enormous amounts of stress, cognitive workload, and time pressure (see Hancock et al., 2011, for a review).

The relation between these idiosyncrasies and the resulting need for trust is not simple to define. A high degree of automation does not necessarily require a high degree of trust. For instance, people do not want information as to the dangers of highly automated machines such as aircraft autopilot systems or robotic vacuums. People tend to trust those machines and today, accepting them is natural. In contrast, other highly automated technology that is not yet common to society will certainly require substantial trust in the beginning. As noted earlier perceived risk determines the need for trust, yet the risk of technologies such as self-driving cars is largely unknown. Re-

liance can also be influenced by the anticipation of greater benefits of the automation (Hoff & Bashir, 2015). These factors are differently weighted by individuals, making the precise prediction of trust formation in novel machinery a challenging task indeed. However, it is possible to identify factors that consistently play a key role for trust development, including predictability, perceived risk, and benefits.

3.2 ANTHROPOMORPHIC AGENTS

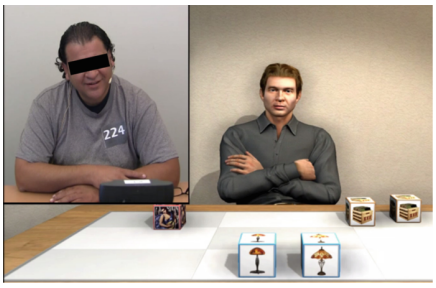
Researchers and practitioners across fields such as HCI, human–robot interaction, artificial intelligence, cognitive science, psychology, and computer graphics advance the degree to which technology can show human-like capabilities and traits such as computational cognition, photo-realistic facial appearance, nonverbal and verbal communication. The overarching goal is to enable humans to immerse in novel and rich interactions. As a consequence, people increasingly interact with artificial agents capable of human-like characteristics (see Fig. 3). Such agents can display and understand human emotion (Gratch & Marsella, 2004; Picard, 1997), convey the impression of distinct personalities (Isbister & Nass, 2000), negotiate with humans (Lin & Kraus, 2010), teach negotiation (Gratch et al., 2016), act as coaches and tutors (Breazeal et al., 2016; Kok et al., 2015), learn how to reject orders beyond their abilities and social norms (Briggs & Scheutz, 2015), manipulate their own trustworthiness (DeSteno et al., 2012), and may soon be endowed with artificial forms of moral cognition to better collaborate with humans (Malle & Scheutz, 2014). Although in many cases, the degree of autonomy is not yet maximized, the approaches certainly demonstrate how researchers increase the human-like capabilities of artificial agents and widen the range of applications not just across different scenarios, but also user groups (Breazeal et al., 2016).

Intriguingly, the manifold embodiments of artificial agents that range from computers and personal assistants to 3D virtual characters and robots have been shown to elicit natural responses among humans (e.g., Krämer et al., 2015; Sidner et al., 2005; Walter et al., 2014). On a neural level, interacting with human-like agents is linked to activation in brain areas important for mental state attribution, which may explain the unique perceptions of, and responses to human-like artificial agents such as enjoyment and attributions of intelligence (Krach et al., 2008).

Anthropomorphism[1] has crucial implications for the design of appropriate trust. Conceptually, Dautenhahn developed the idea of so-

1 The terms "human-likeness" and "anthropomorphism" are often used interchangeably. Usually, the latter is more strongly linked to the process whereby an agent is attributed human qualities, hence it is *anthropomorphized* (Epley et al., 2007; Waytz et al., 2014). For the sake of clarity we adopt the terminology and perspective commonly used in HCI studies, that is, anthropomorphism is the degree to which agents display human characteristics (Burgoon et al., 2000; Visser et al., 2016).

(a) Scheduling plans with an attentive virtual agent (Buschmeier & Kopp, 2011).



(b) Negotiation with a virtual competitor (Gratch et al., 2015).

(c) Face-to-face interaction with a virtual child (Sagar et al., 2016).

(d) A virtual nurse discharges a patient (Zhou et al., 2014).

(e) Playing cards with the virtual agent Max (Becker-Asano & Wachsmuth, 2010).

Figure 3: Interacting with anthropomorphic agents.

cially intelligent agents (1998). According to her view, agents that are endowed with social intelligence can mediate cooperation and problem-solving by means of social abilities similar to humans, including embodiment and relationship maintaining.

Empirically, a growing body of research is devoted to the key social dimensions that influence trust in anthropomorphic agents, such as perceived trustworthiness and human emotional responses. Following this idea, researchers investigated the effect of various social cues on trust and the willingness to cooperate, including combined anthropomorphic cues such as visual appearance and voice (Kiesler et al., 1996; Parise et al., 1999; Visser et al., 2016; Visser et al., 2017),

voice, human name, and gender (Waytz et al., 2014), emotion expressions (Antos et al., 2011; Choi et al., 2015; Melo et al., 2014), agency (Melo et al., 2015), and nonverbal behavior (DeSteno et al., 2012). Section 3.3 describes how anthropomorphic cues offer the possibility to shape interactions with humans along trust dimensions.

Specific techniques and mechanisms have emerged to transport the impression of anthropomorphism. Researchers use a variety of design elements to convince humans the agent they are interacting with has intentionality, that is, beliefs, desires, and intentions (Breazeal & Scassellati, 1999). The modeling of human-like verbal and nonverbal communication (e.g., speech, gaze, gesture; Breazeal et al., 2005; Kopp & Wachsmuth, 2004; Kopp et al., 2006) and relationship-building techniques (e.g., conversational dialogue, politeness, self-disclosure, rapport; Bickmore & Picard, 2005; Cassell & Bickmore, 2000; Gratch et al., 2007) has received considerable attention. In particular, the ability to incrementally produce verbal and nonverbal chunks of information that are tightly synchronized and connected to contextual cues of the interaction is an ongoing challenge (Kopp et al., 2006; Welbergen et al., 2015). Emotions play a major role, and facial features are an important means to communicate emotional states (Breazeal & Scassellati, 1999). It was shown that human perceivers use facial features of a virtual agent to infer its underlying beliefs, desires, and intentions in cooperative situations (Melo et al., 2014). Adding to this, affect simulation architectures are used for the online modeling of affective states which are then mapped onto behaviors (Becker-Asano & Wachsmuth, 2010). It is argued that the communication of thoughts and feelings through subtle cues such as facial features may be the most important factor in conveying the illusion of an anthropomorphic entity (Melo & Gratch, 2015; Sagar et al., 2016). A different approach is to provide simple contextual cues that affect people's perceptions. This includes, for example, agent background stories containing a life history, profile picture, personality descriptions, and even fictional news reports (Visser et al., 2016). Agents are sometimes given a gender and human names (Waytz et al., 2014), or they are assigned to the same team as humans (Nass et al., 1996).

3.3 METHODOLOGICAL ADVANCES

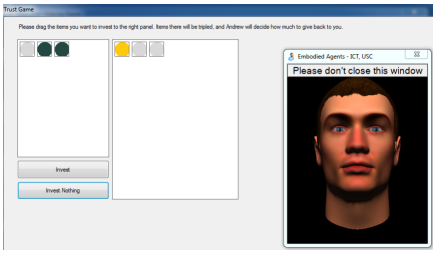
In the past years, researchers have increasingly focused on decision-making for the study of trust as well as cooperation in HCI and adopted approaches known from behavioral game theory (BGT). The BGT methodology seeks to augment the generality and mathematic precision of classic game theory by means of empirical evidence:

> "Would you lend money to somebody who doesn't have to pay you back, but might feel morally obliged to do so? If you would, you trust her. If she pays you back, she is

> trustworthy. This definition gives a way to measure trust, and has been used in experiments in many places [...]" (Camerer, 2003, p. 3).

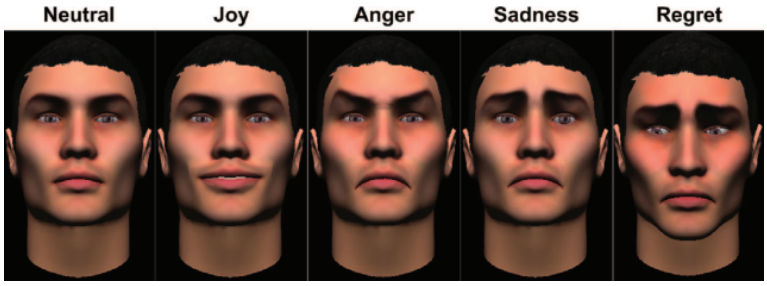
Indeed, prior to the propagation of BGT in HCI, researchers studying distributed artificial intelligence and multi-agent systems adopted the formality of game theory to analyze decision problems in domains such as negotiation (Zlotkin & Rosenschein, 1989), cooperation, and economics (see Parsons & Wooldridge, 2002, for a brief overview). In contrast to game theory, *behavioral* game theory researchers collect empirical data about human behavior in strategic situations (Gächter, 2004). According to this idea, humans and artificial agents engage in 2-player games of economic exchange with the option to either cooperate or defect. Cooperation maximizes the joint payoff, but defection maximizes the individual payoff. This often poses a dilemma as defection becomes the rational, utility-maximizing choice for both players, yet mutual cooperation yields a better payoff than mutual defection (Dawes, 1980). The underlying assumption is that instead of pure economic decision-making, choices in such games provide a reliable approximation of trust and trustworthiness (give-some dilemma; Camerer, 2003) or cooperation (prisoner's dilemma; Gächter, 2004). HCI studies commonly utilize games to model the entire interaction, that is, a human player is asked to engage in one or multiple rounds with an artificial agent. The agent can provide feedback after the choices are revealed in order to investigate how different feedback behaviors affect subsequent cooperation (e.g., Kulms et al., 2014a; Melo et al., 2014). Alternatively, few examples use games as trust measure after the actual interaction, such as negotiation and open dialog (e.g., Antos et al., 2011; DeSteno et al., 2012). Figure 4 shows common setups of studies carried out in those paradigms.

A central conclusion from this growing body of research is that the human tendency to deviate from selfish utility maximization in favor of cooperation extends to computer agents as well (Kiesler et al., 1996). It was also revealed that humans cooperate more with other humans (Miwa et al., 2008; Sandoval et al., 2015) or with humans being represented by virtual avatars (Melo et al., 2015), compared to computer agents. Endowing computers with emotion expressions can eliminate this difference (Melo & Gratch, 2015). Others found that the likelihood of keeping a promise to cooperate is similar for playing with another person and a text-only computer, but decreases when playing with an anthropomorphic computer with a synthesized face and voice output (Kiesler et al., 1996). Moreover, participants cooperated more with agents showing positive appraisal of joint cooperation by means of emotion expressions (Choi et al., 2015; Melo et al., 2014). Against this background, participants preferred interacting with agents that display emotions consistent with their actions, for instance an angry selfish agent is preferred over a happy selfish agent (Antos et al., 2011). Finally, there are also first attempts in identifying signals of trustworthiness in natural, dynamic interactions and



(a) Simple trust game interface with a virtual agent (Antos et al., 2011).

(b) Chatting with a humanoid robot before performing a trust task (DeSteno et al., 2012; Lee et al., 2013).



(c) Facial features used to model outcome appraisal of a counterpart in the prisoner's dilemma (Melo et al., 2014).

Figure 4: Anthropomorphic agents are used as counterparts in empirical studies on trust and cooperation. Researchers investigate how anthropomorphic signals affect the decision to trust or cooperate.

mapping them onto robot gestures (DeSteno et al., 2012; Lee et al., 2013). Overall, converging evidence on the influence of nonverbal behavior suggests that anthropomorphism offers unique ways to shape the perceived trustworthiness of artificial agents. Table 1 shows the antecedents of trust in as well as cooperation with computer agents that were uncovered using standard game theoretic procedures.

Another framework that provides a simple analytical environment is the Desert or Lunar Survival Problem, respectively (Lafferty et al., 1974). This setting is a hypothetical problem-solving task in which participants are told they have crash-landed in the desert (or were stranded astronauts on the moon). Participants are given a list of items they need to use for survival such as water, a map, a pistol, etc. The task is to prioritize the list based on each item's importance for survival. Usually, a pair or group of participant is tasked with agreeing on one list. In HCI, people are paired with a computer agent to solve the task. Applying this principle, researchers have shown that computers were more influential than humans (Burgoon et al., 2000), participants prefer computers with a similar personality in text-based communication (Nass et al., 1995), but also showed preference for complementary characters if the computer was given a virtual body (Isbister & Nass, 2000), and participants affiliate with computers as a team if they believe to be interdependent with it (Nass et al., 1996).

Table 1: Antecedents of trusting or cooperative human decisions in social dilemmas played with artificial agents.

Game	Authors	Antecedent	Supported
Prisoner's dilemma	Choi et al., 2015	Emotion expressions (by agents); inferential and affective processes (by perceivers)	✓✓
	Melo et al., 2014	Emotion expressions	✓
	Melo et al., 2015	Emotion expressions; agency	✓✓
	Kiesler et al., 1996	Anthropomorphism: computer vs. human	*
	Kulms et al., 2014a	Cooperation rate; humor	✓✗
	Miwa et al., 2008	Anthropomorphism: computer vs. human; cooperation rate	✓✗
Give-some dilemma	Antos et al., 2011	Emotion expressions	✓
	DeSteno et al., 2012	Nonverbal cues	✓
	Parise et al., 1999	Anthropomorphism: human-like vs. dog-like vs. human agent	✓
Ultimatum game	Nishio et al., 2012	Anthropomorphism: computer vs. robot vs. human	*
	Sandoval et al., 2015	Anthropomorphism: robot vs. human; reciprocity	✓✗
Stag hunt game	Tsai et al., 2012	Emotional contagion: happiness	✗

Notes. a) Asterisks (*) denote partial support. b) Work investigating other outcomes such as theory of mind (Krach et al., 2008) or cognition patterns during social interaction (McCabe et al., 2001; Rilling et al., 2004; Van't Wout et al., 2006; Yoshida et al., 2010) is excluded. c) The give-some dilemma is often referred to as trust game.

Moreover, the influence of a joking virtual agent decreases if it fails at being judged as funny (Khooshabeh et al., 2011).

3.4 EXPLAINING THE SOCIAL INFLUENCE OF TECHNOLOGY

To understand why humans can engage in rich interactions with computer agents and be affected in ways tangent to trust in the first place, it is useful to turn to the general idea of social influence caused by computers. In social psychology, social influence means that one's attitudes, behaviors, and emotions are affected by others (Zimbardo

& Leippe, 1991). Various frameworks have been proposed to explain how and why computers exert influence on people.

Nass and colleagues explained social effects with the human tendency to automatically – mindlessly – respond to computers as if they were human and applying social scripts, norms, attributions, and expectations to them, hence the term "Computers-are-social-actors (CASA)" (Nass et al., 1994; Reeves & Nass, 1996). Research based on the CASA paradigm documented how people's responses are triggered by social cues such as voice output and social categories. Across different scenarios, participants showed social responses in the form of gender-stereotyping, cooperation, positive subjective assessments, or perceiving the target computer as similar (Nass & Moon, 2000). The approach predicts that computers influence human users in a fashion similar to how they are influenced by other people. In contrast, other research indicates qualitative differences between computers and people, suggesting that the effects caused by computers cannot be equated with interpersonal relationships. Those differences arise from computers' lack of intentionality, asymmetries in human–computer relationships (e.g., humans do not need to signal trustworthiness to computers), as well as certain biases caused by computers (Lee & See, 2004). According to this view, effects such as positive attitudes toward computer-generated information are caused by inappropriate cognitive processing of computer cues which negatively affects one's overall attention to the environment (Parasuraman & Manzey, 2010, see Section 3.1).

A collaborator of Nass, B. J. Fogg, focuses on the effect of human–computer interaction on motivation and persuasion. Like the CASA paradigm, this approach is based on the idea that people respond to computers as if they were social entities (Fogg, 2003). The principles behind Fogg's research and the CASA paradigm are fairly similar; the extensions he provided center on the origins and implications of social influence. Fogg proposed five social cues through which computers exert influence: physical cues such as faces and bodies, psychological cues such as apologies and personality, language, social dynamics like cooperation and turn-taking, and social roles. Fogg focused on the persuasive potential of computers in order to modulate user attitudes and behavior in specific environments such as continued interaction.

Researchers are increasingly interested in how anthropomorphic virtual characters affect people's experiences. Such virtual or embodied conversational agents (ECA), respectively, provide a crucial extension to conversational interfaces as the added social cues are believed to deepen the social effects first documented by Nass and colleagues. Against this background, Blascovich (2002) explains social influence as the extent to which humans experience shared reality with virtual characters. The key mediating factor in this process is social presence

or social verification, respectively, which describes how much the interaction with virtual characters verifies a certain meaningfulness to the communication, hence creating the perception of a shared experience. Social presence is a function of two components: agency, the degree to which virtual characters are perceived to represent real people, and behavioral realism, which describes how much those characters are perceived to behave like real people. Perceived agency varies depending on whether the character is controlled entirely by a computer agent versus another human, thereby making it an avatar. Perceived behavioral realism is maintained by the character's range of realistic social cues. In order for social influence to occur, a perceptual threshold needs to be reached by providing humans with sufficient levels of perceived agency and behavioral realism. Following this idea, Bailenson and colleagues demonstrated that in an immersive virtual environment, people respond naturally to virtual characters in terms of personal space and social presence (2003).

Anthropomorphic interfaces permit further applications of perceptual frameworks to HCI. The warmth and competence framework, for instance, was applied to model and evaluate the nonverbal behavior and appearance of ECAs. In one study, participants watched a video of an ECA showing either high versus low warmth and competence nonverbal behavior, without voice output (Nguyen et al., 2015). The aim was to validate the nonverbal behavior with respect to warmth and competence manipulations, which were generated based on human actor recordings. Aside from the successful validations, it was found that the high-warmth target agent was perceived as more competent than the low-warmth target, and competence may play a moderating role for low-warmth targets. Another study found that attributions of warmth were modulated by the point of measurement and appearance of the agent, indicating that warmth declines over time for a robot-like virtual agent (Bergmann et al., 2012). This effect diminished when the target agent was human-like.

Another approach examining the degree to which computers are treated like humans focuses on people's tendency to anthropomorphize non-human entities because of the perception of having mind (Waytz et al., 2010). Mind perception is constituted by two dimensions: experience, which is broadly linked to the experience of emotions and pain, and agency, which is linked to responsibility of actions, self-control, and planning (Gray et al., 2007). According to this view, perceiving a computer as an entity capable of planning and acting (i.e., having agency) as well as sensing and feeling emotion (i.e., can experience) predicts if it is treated like an actual person (Melo & Gratch, 2015). Based on this, de Melo and colleagues investigated the determinants of social decision-making when interacting with human versus computer partners, using BGT (see Melo & Gratch, 2015). In the ultimatum game, participants are given an initial amount of

money and decide how much they offer their partner. The partner can accept or reject the offer, but if the offer is rejected, all amounts are canceled. It was demonstrated that when the partner did not show emotion, human partners received higher offers than computers. However when partners showed happy (sad) facial expressions in response to fair (unfair) offers, the difference between computers and humans diminished. The authors conclude that endowing computers with experience is the key factor in treating computers as if they were humans in this BGT paradigm. Since people seem to expect a general lack in the factor experience, emphasis should be put on the simulation of emotional intelligence (Melo & Gratch, 2015). In line with this, the experience factor seems to account for a larger variance proportion than agency in the mind perception process (Gray et al., 2007). However, de Melo et al. also stress that in some scenarios, artificial agents should not try to be treated as humans as this would compromise sensitive interaction goals such as self-disclosure in health-screening (Melo & Gratch, 2015).

3.5 SUMMARY

Intelligent artificial agents are increasingly able to engage with humans in rich and natural forms of interaction that are driven by complex goals and require interdependent decision-making. Through simple cooperative games, for instance social dilemmas, researchers have shown that with artificial counterparts, human decision-making is not bound to pure selfish interests. Rather, trusting and cooperative decisions are influenced by agent anthropomorphism. Computers increasingly have anthropomorphic cues, but since human–computer trust is driven by certain heuristics and biases, it is not the same as interpersonal trust. Based on this essential contrast, it is still not entirely clear how anthropomorphism affects trust in cooperative situations.

4

AN INTERACTIVE COOPERATION GAME PARADIGM FOR THE INVESTIGATION OF TRUST IN HUMAN–COMPUTER INTERACTION

To study open issues of human–computer trust, an interactive cooperation game paradigm was developed. The paradigm is used as conceptual framework and template for experimental research games, so-called puzzle games. The paradigm's mechanisms were meant to provide various manipulations of strategic and cooperative agent behavior other cooperative games cannot provide.

This chapter consists of three logical parts:

- Background
  - Section 4.1: Motivation behind the paradigm
- The paradigm
  - Section 4.2: The general task
  - Section 4.3: Underlying cooperation concepts
  - Section 4.4: Cooperative interaction factors used in the experiments
- Preparing empirical research with the paradigm
  - Section 4.5: Validation study to analyze cooperation in the paradigm
  - Section 4.6: Factors of trust investigated by the experiments

4.1 MOTIVATION

The paradigm permits the investigation of trust in environments involving interdependent decision-making, coordination, and communication. The approach is fundamentally different from the BGT and social dilemma perspective of cooperation. BGT is a predominant method to study interpersonal trust and many researchers adopted this mindset to investigate human–computer trust. Indeed, using precise conceptualizations of trust, trustworthiness, and cooperation, BGT offers concise ways to advance social decision-making research in an astonishing fashion. The large amount of BGT research has generated important empirical advancements across many fields related to psychology, enabling theoretical integration and model construction (Sally, 2000; Weber et al., 2004). The methodology even involves an inherent motivation to adopt HCI methods: By comparing participant interactions with human and computer counterparts in the

trust game, researchers demonstrated that the neural activation of mentalizing networks in the brain occurs in human–human, but not human–computer interactions (McCabe et al., 2001).

However, the BGT methodology also has significant **limitations** the present paradigm tries to compensate. Instead of providing a dynamic and interactive environment, agent actions in social dilemmas such as the prisoner's and give-some dilemma are confined to simultaneous and discrete decisions of both players. Arriving at an informed decision mostly requires inferring the counterpart's intentions. It does not at all require inferring the counterpart's capability to execute decisions in a competent manner to the extent that there is no challenge involved in translating intentions into actions. Yet, competence is not entirely irrelevant in social dilemmas. Prosocial individuals show increased cooperation when primed for competent (e.g., 'intelligent', 'competent', 'clever') behavior, whereas the same priming makes proself individuals more selfish (Utz et al., 2004). However, this does not capture the translation of intentions into actions and the attributions arising thereof.

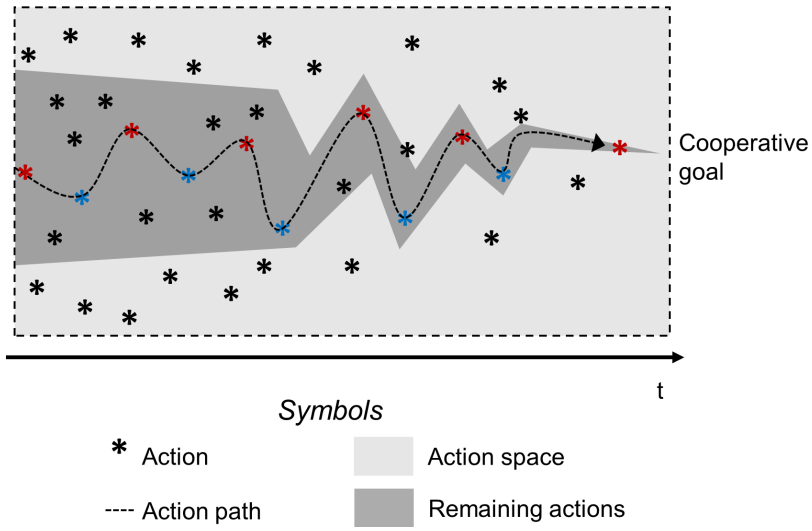


Figure 5: Schematic illustration of an action path in a naturalistic cooperative environment involving a red and blue agent.

In more naturalistic interactions, cooperation between social agents often involves **interdependent decision-making** to the extent that the agents are required to interactively coordinate their efforts. In order to effectively **coordinate**, agents need the ability to manipulate the environment as intended and adjust their actions. Moreover, cooperative tasks often involve **communication**, the division of actions, or sharing of resources (Deutsch, 2011). Figure 5 depicts a more naturalistic understanding of cooperative action between a red and blue social agent. According to this view, an increase of the degrees of freedom and uncertainty requires a much more rational action pattern across long periods of time. Because of the large action space, planning and

anticipating in such environments is more difficult; achieving the cooperative goal purely by chance is unpromising. Another important element is how **time** affects cooperation. As the agents advance toward the goal, each step contributes to the final outcome, but also confines the remaining options naturally. Moreover, poor problem-solving and coordination have a severe effect on the final outcome, causing the action path to miss the cooperative goal.

Other researchers also tried to compensate for the BGT limitations. To rectify the lack of degrees of freedom and interactivity, the simplified stag hunt coordination game has been modified to permit more interactive 2-agent decision-making (Yoshida et al., 2010). Likewise, the key motivation behind the paradigm's approach is to incorporate more naturalistic, competence-based cooperation. The following sections describe how this is achieved.

4.2 TASK: COOPERATIVE GAMEPLAY

The paradigm involves a turn-based 2-player interaction geared toward a specific goal. To achieve the goal, human participants are tasked to cooperate with a **computer agent**. Only if both players equally perform competent actions, the goal can be achieved. The **premise** is that two players try to solve a two-dimensional puzzle field as efficiently as possible by alternately placing blocks into a puzzle field. The (joint/shared/cooperative) **goal** is to complete a certain number of horizontal rows. A row is completed if there are no empty fields left, similar to Tetris. In contrast to Tetris, completed rows are not cleared, thus poorly positioned blocks will remain at their spot until the round is finished. There are two blocks: a T-shaped and a more difficult U-shaped block. Attributes such as "easy" or "difficult" are avoided throughout the game. Another deviation from Tetris is that blocks do not gradually descend from the top to bottom, relieving players of time constraints. Players can exchange task-related advice as to which block should be used next or where it should be placed. Figure 6 shows the main conceptual components.

Tetris – one of the most popular games in history – was chosen as interaction paradigm because it provides a dynamic task environment that is useful for experimental research. Tetris has been applied to study cognitive skills for problem solving (Kirsh & Maglio, 1994; Lindstedt & Gray, 2015) and social presence in cooperative environments (Hudson & Cairns, 2014). The paradigm was realized using a prototyping platform for multimodal user interfaces (Kulms et al., 2014b, 2015, 2018; Mattar et al., 2015).

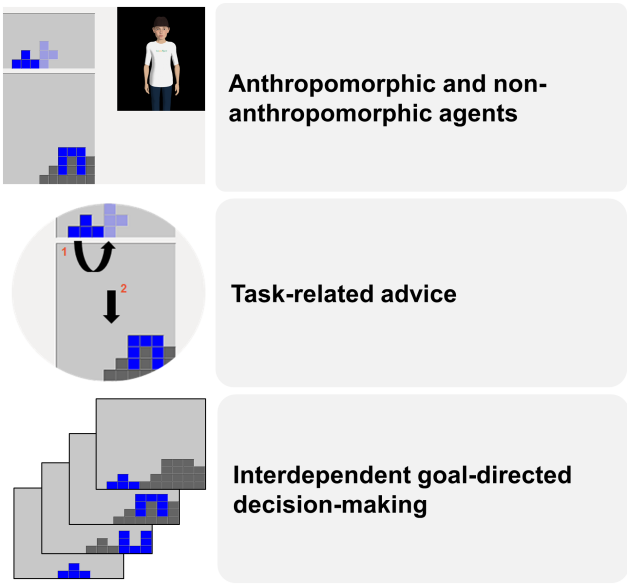


Figure 6: Conceptual components in the paradigm.

4.3 THE COOPERATION CONCEPT

The range of actions and interactive elements in the game are conceived as hierarchic dimensions of cooperation. According to this idea, the interaction involves three distinct dimensions, each evoking specific social attributes of the agent that influence whether human players perceive it as cooperative partner. The dimensions are not strongly hierarchic in that only the lowest dimension is the necessary basis of cooperation as it represents the main task. The other dimensions are optional for cooperation and there is no interrelation that defines a particular order.

The first dimension, *coordination*, describes how the puzzle competence of the human and agent as well as their ability to coordinate determine if the joint goal will be attained. This dimension essentially captures the "acting together" component of cooperation (Brosnan et al., 2010). Since completed rows are not cleared, incompetent actions and errors have a crucial impact on the outcome. Selfish desire for the high value U-block also affects coordination: If the agent repeatedly demands the block for itself, it causes a stable block sequence (1: U, 2: T, 3: U, 4: T, etc.) and eliminates flexibility on that front, possibly impeding coordination. The second dimension, *communication*, permits players to exchange task-related information that help them to coordinate. Finally, the third dimension, labeled *strategic social behavior*, describes how players can strategically opt for the high value U-block. If this dimension is implemented, perceived selfishness and competence of the agent culminate in attributions regarding its strategy. For instance, the agent could utilize its alleged incompetence as a strategic decision to undermine the counterpart's individual pay-

off by impeding the joint goal. This should decrease the agent's perceived trustworthiness. Likewise, selfishness helps the agent to maximize its payoff; this should also deteriorate its trustworthiness.

In contrast to standard cooperative games, the present paradigm deviates from the clear distinction between cooperation and defection. In the popular prisoner's dilemma, for instance, players can either cooperate or defect. However, this distinction cannot capture more nuanced shades of cooperative behaviors the present paradigm does offer, such as wanting but failing to cooperate, or using the selfish option to promote the joint goal.

4.4 INTERACTION FACTORS

In order to allow the investigation of trust-based research questions by means of human–computer interactions, a number of interaction factors is incorporated in the paradigm. This section describes these interaction factors and explains their operationalizations.

4.4.1 *Puzzle competence and selfishness*

According to the trust literature, two components determine whether social agents can be trusted and cooperated with. The "can-do" component captures the abilities and competencies necessary for achieving a particular goal, while the "will-do" component reflects if the abilities are used in the best interest of a trusting agent or one's own interest (Colquitt et al., 2007). The paradigm maps these characteristics onto two interaction factors: puzzle competence and selfishness.

Puzzle competence is defined as the ability to place blocks in an efficient manner. To this end, a simple heuristic was implemented to compute optimal configurations of the upcoming block (i.e., positions and rotations). Configurations were optimal (sub-optimal) if they maximize (minimize) the number of completed rows. Optimal configurations were assigned high decision weights, sub-optimal configurations were assigned low decision weights. The competent agent uses the highest decision weight to determine the upcoming block configuration. Conversely, the incompetent agent uses the lowest weight.

Selfishness determines if the agent desires the U-block which yields more individual points. If the agent desires the U-block, it will only comply with advice containing this block; if the agent does not desire the U-block, it will comply with any advice. Importantly, selfishness does not preclude cooperation in that the joint goal can be achieved.

4.4.2 *Task advice*

In interpersonal interactions, people facing a decision-making problem are often influenced by others or actively consult their peers for

advice (Bonaccio & Dalal, 2006). In human–computer interactions, users take advice from digital assistants in a variety of situations such as route planning, traffic control, finance management, physical exercising, and shopping (Hoff & Bashir, 2015; Kok et al., 2015; Madhavan & Wiegmann, 2007; Qiu & Benbasat, 2009, 2010).

4.4.2.1 *Advice given by humans*

As interactions with computers increasingly involve interdependent cooperative actions, there may also be scenarios with computers taking advice from humans. In situations involving uncertainty, human input is not entered as a command in the classic sense. Rather, as both human and computer agents play a more equal role in problem-solving and must explore an uncertain path toward the joint goal, human input becomes more advice than command, permitting more symmetrical exchanges that are not yet realized today. In such a scenario, human players suggest the next block to the agent, and the agent's responses are driven by selfishness.

Because the agent receives advice and not vice versa, human advice adoption cannot be used to infer behavioral trust. Instead, the one-shot give-some dilemma is used.

4.4.2.2 *Advice given by computers*

On the other hand, more classic interaction paradigms involve a computer agent giving advice to users. Designing trustworthy task-related advice and investigating how people respond to it, in particular, if they utilize advice to the advantage of their goals and correctly discard incompetent advice is important for the development of appropriate trust. Participants first decide whether they require advice. Next, they evaluate if the presented advice is useful to their problem-solving and should be *adopted*.

Accordingly, advice adoption is used as behavioral trust measure.

4.4.3 *Anthropomorphism*

As explained in the previous chapter, computer agents are increasingly endowed with human-like characteristics across a wide variety of applications, posing novel challenges for the development of trust. In advice adoption studies, anthropomorphism was shown to mitigate the loss of trust in response to declining reliability (Visser et al., 2016). Oxytocin – a neuropeptide that mediates trusting relationships by increasing the willingness to accept social risks (Kosfeld et al., 2005) – augmented trust in anthropomorphic computers, but not humans (Visser et al., 2017). Researchers argue that anthropomorphism could help reduce the unique biases people apply such as heightened attention to computer errors (Madhavan & Wiegmann, 2007).

Nonetheless, the implementation of anthropomorphic cues should be in accordance with the principle of appropriate, not maximized trust (Lee & See, 2004).

The effect of agents varying in anthropomorphism on trust is investigated by incorporating non-anthropomorphic computer agents as well as anthropomorphic virtual and human agents. These agents act as assistants that provide task-related advice in the puzzle game and communicate with users.

4.4.4 *Blame*

People adeptly regulate the relations with their peers during conflict using mechanisms that are fine-tuned to the specific context. Social blame is one of these mechanisms. Blame allows social agents to address issues in social interactions such as inappropriate behavior and disappointment (Malle et al., 2014). Only little research has investigated the potential role of blame for HCI. It is hypothesized that communicative acts such as blame could help regulate human–computer trust. The self-serving bias may play an important role for the attribution of task success and failure: negative outcomes and the associated blame are often attributed to computers, while users attribute success to themselves (Moon, 2003). Further HCI research indicates that robots which blame themselves or the team are perceived as more likable (Groom et al., 2010; Kaniarasu & Steinfeld, 2014). The relation between blame direction and trust, however, are less clear.

Based on its ability to show emotion displays and use dialog, the anthropomorphic virtual agent used in the paradigm can communicate blame after trust-threatening events for an attempt to regulate trust. A critical trust-threatening event in the paradigm is joint goal failure. Since the two players jointly work toward the goal, there are different targets of blame. It is analyzed how these behaviors affect trust in the agent.

Table 2 explains how the interaction factors and cooperation concepts relate to the research questions. The specific experimental implementations (e.g., agent behaviors) will be explained in the method part of each experiment in the following chapters.

4.5 VALIDATION STUDY

The basic **idea** of the paradigm is using it to derive and construct games with different cooperative mechanisms in which participants can easily cooperate with a computer agent toward a specific goal. This is also the necessary condition the paradigm must fulfill in order to enable the study of human–computer trust. This claim is tested in

Table 2: Full view of the interaction factors with corresponding cooperation concepts and trust variables.

	Trust research question			
Cooperation concept	**Antecedents**	**Formation**	**Formation**	**Regulation**
Strategic social behavior	Selfishness	-	-	-
Communication	Advice giving	Advice taking	Advice taking	Blame
Coordination	Puzzle competence	Puzzle competence, advice quality	Puzzle competence, advice quality	Puzzle competence
Trust variable	Give-some dilemma	Advice adoption across multiple rounds	Advice adoption across multiple rounds	Give-some dilemma

Note. As seen from the human player's perspective.

a validation study which uses the first version of the game (Kulms et al., 2016).

4.5.1 *Overview*

The goal of the validation study is to analyze cooperation in the puzzle game. Special emphasis is put on how the agent's task-related behavior affects social attributions that are important for cooperation such as trustworthiness. This step is designed to refine the agent's ability to coordinate with human players and aid the modeling of trustworthy agent behavior in the upcoming experiments. Perceived trustworthiness is an important agent quality in interpersonal cooperation and promotes trust (Mayer et al., 1995). To achieve this, the behavioral dimensions *puzzle competence* and *selfishness* (see Section 4.4.1) are implemented and manipulated. The following research questions will be analyzed:

> **RQ1**: How can cooperation between the agent and human players be characterized?
>
> **RQ2**: How does the agent's puzzle competence and selfishness relate to its perceived cooperativeness and trustworthiness?

4.5.2 *Method*

4.5.2.1 *Participants*

Eighty-seven German undergraduate and graduate students participated in exchange for 5 EUR. Nineteen participants were removed due to technical difficulties, comprehension issues, because the outlier analysis revealed that their understanding of the agent was inaccurate, or to achieve equal group sizes across the conditions.

The final sample consists of 68 participants. The sample ranged in age from 18 to 53 years ($M = 25.63$, $SD = 5.63$, $median = 25$; female: 60.3%).

4.5.2.2 *Task: Puzzle game*

The task involved participants trying to solve the puzzle game as efficiently as possible with the agent. In this instance of the paradigm, each player collects individual points for placing a block, regardless of where they are placed. There are two blocks available, a T-block and a more difficult U-block. The T-block yields 5 individual points and the U-block yields 10 points. Additionally, both players receive a bonus score for achieving the optional joint goal, that is, completing a certain amount of rows.

The game proceeds in alternating turns. In each turn, participants first advise the agent of a block. The agent either accepts or rejects the advice; it then places its block and leaves the remaining one to the participant (see Fig. 7).

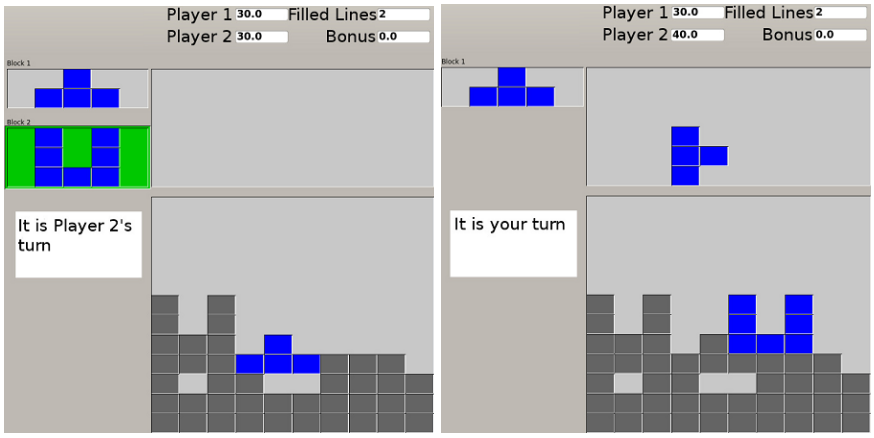


Figure 7: Interface of the puzzle game. The top area is for human players to manipulate the block, the bottom area is the puzzle field. Left: the human player has recommended the U-block. Right: the agent has accepted and placed the block, leaving the T-block to the human player.

4.5.2.3 *Agent behavior*

The agent was modeled to incorporate varying degrees of puzzle competence and selfishness. The selfish agent only accepts the high value U-block as advice. In contrast, the unselfish agent accepted all suggestions. Some specific patterns follow from the two components: selfish agents always receive a higher payoff than their human counterparts; it is impossible to attain the joint goal with the incompetent agent; because a selfish agent desires the U-block, the block order is constant (1: U, 2: T, 3: U, 4: T, etc.) and the human player always gets the T-block.

4.5.2.4 *Design*

The study had a 2×2 between-subjects design, with puzzle competence (competent vs. incompetent) and selfishness (selfish vs. unselfish) as between-subjects factors.

4.5.2.5 *Measurement*

To account for the agent's social attributions, its perceived competence, trustworthiness, and cooperativeness were assessed. Perceived competence and trustworthiness were measured using a perceived computer credibility scale (Fogg & Tseng, 1999). The competence items were 'knowledgeable', 'competent', 'intelligent', 'capable', 'experienced', 'powerful', Cronbach's $\alpha = .80$. The trustworthiness items were 'trustworthy', 'good', 'truthful', 'well-intentioned', 'unbiased', 'honest', Cronbach's $\alpha = .84$. Perceived cooperativeness was measured by asking participants how much they felt the agent tried to achieve the joint goal. As manipulation check, participants were asked how often the agent accepted their advice. All items were rated on 5-point Likert scales.

4.5.2.6 *Procedure*

Participants met the experimenter, completed informed consent, and received written instructions. They also received information regarding the payoff structure. Participants were not asked to focus solely on the joint goal, and they were not told about the strategy of the agent. This was done to avoid initial expectations regarding the agent's trustworthiness and cooperativeness, that is, the degree to which it tries to attain the joint goal, and to let participants focus on their own strategy. Prior to the experimental trials, participants could familiarize themselves with the controls. Next, they played two rounds of the puzzle game with the agent.

4.5.2.7 *Data analysis*

The variables for perceived trustworthiness, competence, and cooperativeness were entered into a 2×2 MANOVA.

4.5.3 *Results*

4.5.3.1 *Manipulation check: Advice acceptance*

There was a significant main effect of selfishness on how often the agent was perceived to accept advice. Participants observed that the selfish agent accepted less advice ($M = 2.00, SD = .75$) than the unselfish agent ($M = 4.76, SD = .44$), $F(1, 73) = 390.62, p < .01, \eta_p^2 = .84$.

4.5.3.2 *Goal attainment*

The incompetent agent was not able to attain the joint goal, irrespective of its selfishness. With the competent agent, participants overall attained the goal at a 75% rate across both rounds if the agent was unselfish. This rate dropped to 50% if the agent was selfish.

4.5.3.3 *Coordination efficiency*

Two heatmaps were composed to show how the unselfish and selfish agent coordinated with participants, given competent puzzle solving (see Fig. 8). Green colors indicate few and red colors indicate many empty fields. The fewer empty fields both players left across the puzzle field, the better the coordination outcome.

Overall, both human–agent pairings coordinated fairly successfully. Pairings with the unselfish agent performed somewhat better at the bottom of the field, that is, at the beginning of interactions. Pairings with the selfish agent left less fields empty at the end of interactions. The main difference between the beginning and end of interactions is a larger complexity and thus increased difficulty at the final stages. In other words, the selfish agent tended to coordinate more efficiently with participants at difficult stages, whereas the unselfish agent coordinated more efficiently at easier stages. A possible explanation for this is that the selfish agent always chose the more difficult U-block and handled this decision better than human players.

4.5.3.4 *Perceived trustworthiness*

There was a significant main effect of puzzle competence on perceived trustworthiness. The competent agent ($M = 3.01, SD = .97$) was judged as more trustworthy than the incompetent agent ($M = 2.39, SD = .46$), $F(1, 64) = 16.00, p < .001, \eta_p^2 = .20$.

There also was a significant main effect of selfishness on perceived trustworthiness. The unselfish agent ($M = 3.06, SD = .79$) was per-

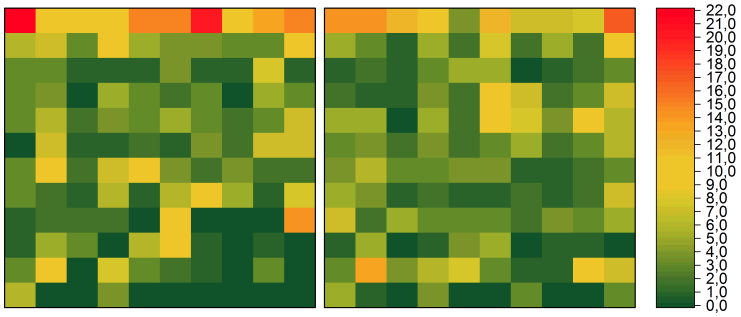


Figure 8: Coordination heatmaps based on the distribution of empty fields across the puzzle field with the competently playing agent. Both rounds combined. Left: unselfish agent, right: selfish agent.

ceived as more trustworthy than the selfish agent ($M = 2.34, SD = .67$), $F(1, 64) = 21.44, p < .001, \eta_p^2 = .25$.

Additionally, there was a significant interaction effect between puzzle competence and selfishness, on trustworthiness ($F(1, 64) = 7.44, p < .01, \eta_p^2 = .10$). If the agent was selfish, trustworthiness was similar for puzzle competence and incompetence (simple effects: $F(1, 64) = 0.81, p = .37$). However, if the agent was unselfish, trustworthiness was significantly higher for puzzle competence than incompetence (simple effects: $F(1, 64) = 22.61, p < .001$). See Table 3.

4.5.3.5 *Perceived competence*

There was a significant main effect of puzzle competence on perceived competence. The competently playing agent ($M = 3.65, SD = 0.88$) was perceived as more competent than the incompetently playing agent ($M = 2.80, SD = 1.17$), $F(1, 64) = 14.65, p < .001, \eta_p^2 = .19$.

There also was a significant main effect of selfishness on perceived competence. The selfish agent ($M = 3.58, SD = 0.89$) was perceived as more competent than the unselfish agent ($M = 2.80, SD = 1.14$), $F(1, 64) = 10.27, p < .01, \eta_p^2 = .14$.

Furthermore, there was a significant interaction effect between puzzle competence and selfishness, on perceived competence ($F(1, 64) = 8.11, p < .01, \eta_p^2 = .11$). This indicates that the influence of puzzle competence on perceived competence is affected by selfishness. Given puzzle competence, perceived competence was similar for selfishness and unselfishness (simple effects: $F(1, 64) = 0.06, p = .80$). In contrast, given puzzle incompetence, perceived competence was significantly lower for unselfishness than selfishness (simple effects: $(F(1, 64) = 18.32, p < .001)$. See Table 3.

4.5.3.6 *Perceived cooperativeness*

Because of considerable dispersion variations (see Table 3), this variable is not analyzed parametrically. The descriptive statistics show

that only the competent and unselfish agent was attributed high cooperativeness whereas the remaining variations were attributed low cooperativeness.

Table 3: Means and standard deviations for social attributions of the agent.

		Competence		Trustw.		Coop.	
Puzzle competence	Selfishness	M	SD	M	SD	M	SD
Competent	Selfish	3.69	0.91	2.44	0.79	1.82	0.81
	Unselfish	3.61	0.87	3.58	0.78	3.88	0.99
Incompetent	Selfish	3.47	0.87	2.25	0.53	1.00	0.00
	Unselfish	2.14	0.98	2.54	0.33	1.18	0.53

4.5.4 *Discussion*

The purpose of the validation study was to explore cooperation in the puzzle game and how task-related behavior affects social attributions of the agent. The results of the first study in this paradigm indicate that competence alone does not ensure an overall positive perception of the agent, although competence was a necessary condition of goal attainment.

Perceived trustworthiness and competence – attributions that are crucial for cooperative interactions – were both affected by the agent's selfishness and problem-solving competence. The main effects indicate that the agent was perceived as trustworthy if it played competently or unselfishly, and the interaction effect shows that trustworthiness was maximal when these two variations were combined. Likewise, the agent was perceived as competent if it played competently or made selfish decisions. How did selfishness become important for attributions of perceived competence? Participants differentiated between selfish and unselfish incompetence. The interaction effect indicates that the *in*competent yet selfish agent was rated nearly as competent as the actual competent agents, although it played much worse in comparison. Participants apparently assumed intentionality behind selfish incompetence (Malle & Knobe, 1997) and felt that the agent had the efficacy to enact its selfish goals. Competence is commonly conceived as self-profitable trait (Peeters, 2002). The present results give this view a new perspective in that selfish incompetence was also conceived as self-profitable. Importantly, this pattern confirms that perceivers integrate information about an agent's competence and intentions (i.e., selfishness) to form trustworthiness judgments.

Although the selfishly competent agent coordinated more efficiently with participants at certain stages because it was responsible for the

more difficult actions, the goal achievement statistics clearly show that participants were more likely to achieve the joint goal with the unselfish agent. In sum, participants were indeed able to effectively cooperate with a competent agent toward the joint goal in the puzzle game. Cooperation was generally more successful if the agent played unselfishly and complied with human advice, which also led to more favorable social attributions.

4.5.5 *Lessons learned*

This section describes practical implications regarding the planned experiments.

4.5.5.1 *Puzzle solving heuristic*

Every following experiment will include some random variation regarding the first block position chosen by the agent, if the puzzle field is still empty. This is designed to reduce the possibility that participants solve the puzzle in the exact same way multiple times in a row.

4.5.5.2 *Measures*

A conceptual improvement regarding the trustworthiness measure may be in order. In the validation study, the two components of the credibility scale – competence and trustworthiness – were evaluated separately. The trustworthiness items are intended to measure perceived goodness and morality of agents (Fogg & Tseng, 1999), hence the separation between trustworthiness and competence. Based on this, relevant patterns regarding selfish and unselfish incompetence were found in the validation study. However, separating trustworthiness and competence stands in contrast to the trustworthiness concept as established in the trust literature, which views competence as important foundation, and may cause conceptual problems. To ensure conceptual integrity, a trustworthiness measure should capture competence as well. Frameworks such as the warmth and competence construct may provide more sound options to differentiate between intentions and competencies.

Another central conclusion of this study is that the measures merely captured how participants perceive the agent's actions, but not if they actually put trust in the agent based on these perceptions. This missing information must be included in trust-based experiments.

4.5.5.3 *Manipulations*

The validation study aimed at manipulating task-related behavior for strategic social cooperation. Puzzle competence increased not only perceived competence, but also perceived trustworthiness of the agent.

As discussed above, altering the competence manipulation to isolate perceived competence from perceived trustworthiness alone would be counterproductive.

Unselfishness increased perceived trustworthiness but decreased perceived competence. A similar negative effect on perceived competence was observed in the prisoner's dilemma. Here, cooperative players who are defected against are judged as less competent by observers (Krueger & Acevedo, 2007). This indicates that the effect of unselfishness on perceived competence must be accepted as natural. Hence, no adjustments are made regarding the manipulations in Experiment 1.

4.6 OVERVIEW OF EXPERIMENTS

As explained in the previous chapter, trust experiences typically embody a number of relevant factors. The research focus of the present thesis lies on the antecedents, formation patterns, and regulation mechanisms of human–computer trust.

Experiment 1: Trust antecedents. Inferring trust antecedents from the interaction is crucial to better *predict* trust in computer agents. Previous research suggests that performance is a very stable determinant of trust, yet it is unknown how trust is related to warmth and competence perceptions of computers. By adopting a social cognition perspective on trust, the experiment investigates antecedents that are deeply rooted in how we perceive and categorize other people.

Experiment 2: Trust formation. This first trust formation experiment analyzes how human players respond to anthropomorphic versus non-anthropomorphic agents giving advice. Advice adoption is a common trust problem in HCI and depends on how trust evolves *over time*. Anthropomorphism is believed to play an important role in trust formation because it fundamentally changes characteristics of the agent such as appearance and voice output, making it more human-like to users.

Experiment 3: Trust formation (extended). Another important factor for the formation of trust is the actual degree of anthropomorphism. Building on the same premises as the previous experiment, this study focuses on human–human versus human–computer trust comparisons. It also investigates how virtual agents, which lie between computers and humans on the anthropomorphism spectrum, affect trust.

Experiment 4: Trust regulation. The ability to regulate trust is crucial to *prolong* cooperation. Critical events in cooperative interactions

such as goal failure can be considered a violation of the trust one has put into their partner. As an attempt to regulate user trust, the anthropomorphic agent communicates blame after joint goal failure.

Complex cooperation often includes many different interaction factors which must be mastered by the social agents in order to attain the goal (see Section 4.4). This implies that constructing "one size fits all" empirical settings may be challenging because it could inflate the number of independent variables and lead to conflicts between them. Regarding the present cooperation paradigm, combining human advice taking as main dependent variable with strategic social behavior components of the agent would obscure the research focus as two sources of information compete against each other: the agent's perceived cooperation strategy, and whether the advice is perceived to be competent. Strategy inferences would also confound how trustworthy the agent is judged to be. Consequently, strategic social behavior and advice taking will not be examined together. Rather, the previously described cooperative interaction factors are implemented in a modular fashion, enabling the focus on distinct trust factors, as explained above.

5

EXPERIMENT 1: TRUST ANTECEDENTS

The goal of Experiment 1 is to shed light on antecedents of human–computer trust that are rooted in fundamental dimensions of social cognition (Kulms & Kopp, 2018). We hypothesize that similar to humans, intelligent agents elicit warmth and competence attributions. These attributions could predict human trust. The guiding idea behind this experiment is that humans are highly sensitive to the intentions as well as abilities of other agents and adjust their responses accordingly (Fiske et al., 2007). Despite the important ongoing debate regarding the similarities (Reeves & Nass, 1996) and differences (Lee & See, 2004; Madhavan & Wiegmann, 2007) between human–human and human–computer trust, it is well established that humans readily respond to social cues by computers (Nass & Moon, 2000). However, we do not know how trust in computer agents is underpinned by characteristics of human social cognition such as warmth and competence.

A better understanding of the factors behind artificial warmth and competence and how they modulate user behavior has significant implications for key issues in HCI. This includes the communication of intentions to foster predictability (Klein et al., 2004) and trustworthiness (DeSteno et al., 2012), or complex psychological challenges like maintaining warmth in face-to-face interactions (e.g., DeVault et al., 2014) and managing prolonged human–computer relationships through relational behavior (e.g., Bickmore & Picard, 2005).

5.1 OVERVIEW

The validation study showed that agents receive different degrees of trustworthiness attributions based on their puzzle-soling competence and selfishness. Experiment 1 directly builds on this premise. The interaction will require a human player and an agent to cooperate toward a joint goal with a competitive payoff structure and an additional individual score.

This experiment investigates the following research question:

> **RQ**: Do people infer warmth and competence traits when interacting with computer agents, and are these attributions related to trust?

Experiment 1 makes the following **contribution** to the present thesis' goals:

- We operationalize and exemplify the characteristics of warmth and competence in an interaction with computer agents that may lack problem-solving competence, but are trying to comply with human intentions, and vice versa.

5.2 METHOD

Participants

Eighty German undergraduate and graduate students participated in exchange for 5 EUR. The sample ranged in age from 18 to 40 years ($M = 23.53, SD = 4.36, median = 23$; female: 62.5%).

Task

Task 1: Puzzle game

The puzzle game variation was identical to the validation study and thus includes strategic social behavior (i.e., selfishness) as well.

Task 2: Behavioral trust game (give-some dilemma)

After the puzzle game, participants engaged in a decision task with the agent, the give-some dilemma (Van Lange & Kuhlman, 1994). Participants were told to possess four tokens and being able to allocate those tokens between themselves and the agent, without the opportunity to exchange information. Importantly, tokens that are exchanged double their value while tokens that are not exchanged keep their value. It was explained that the agent was in the same position and faces the same decision. The game provides an incremental measure of behavioral trust, operationalized as the number of tokens being exchanged. Instead of measuring purely economic decision-making, choices in the give-some dilemma reflect social perceptions of the counterpart and are positively correlated with subjective trust assessments (Lee et al., 2013). Popular versions of this dilemma include a two stage approach (Berg et al., 1995). First, the sender can keep or send tokens to the receiver. Second, the receiver decides how to split this amount. The sender's decision, being a bet that the receiver will to some extent reciprocate a risky decision, measures trust, and the receiver's return measures trustworthiness (Camerer, 2003).

In the present version, participants take the role of the sender. They were told that although both players decide simultaneously, the agent's decision would only be revealed at the end of the experiment to avoid confounding the following measures. In fact, the agent's decision was only a cover to maintain the associated risk and increase participants' social evaluations of the agent. We did not assess participants' expectation as to how many coins the computer would exchange because

we assumed this number would highly correlate with the number participants would be willing to exchange.

Agent behavior

Similar to the validation study, the agent was modeled to incorporate varying degrees of puzzle competence and selfishness. Again, the selfish agent only accepts the high value U-block as advice, whereas the unselfish agent accepted all suggestions. The same patterns follow from these components: selfish agents always receive a higher payoff than their human counterparts; it is impossible to attain the joint goal with the incompetent agent; because a selfish agent desires the U-block, the block order is constant and the human player always gets the T-block.

Several of the mechanisms described above aimed at the modulation of perceived warmth and competence. First, human players give advice to the agent that either accepts or rejects it. The idea behind this pattern is to model the desire of the agent for individual points and introduce (non-)compliance responses to human advice. We assume that those responses carry strong social meaning as they represent how much one trusts advice (Dongen & van Maanen, 2013). Likewise, the agent trying to maximize its payoff should also deteriorate perceived warmth. Second, both players have an individual score. When the joint goal is achieved, a bonus is added to the individual scores as reward. The game thus rewards working toward the joint goal, even for selfish agents. Third, the agent could still display competent behavior, irrespective of selfishness.

Design

The study had a 2×2 between-subjects design, with puzzle competence (competent vs. incompetent) and selfishness (selfish vs. unselfish) as between-subjects factors. Participants were randomly assigned to one of the four conditions.

Measurement

To infer warmth and competence attributions from social perception, we compiled a 5-point semantic differential containing 25 adjective pairs designed to assess a broad range of interpersonal attributes during social perception (Bente et al., 1996; Pütten et al., 2010). Behavioral trust was measured using the number of tokens participants are willing to exchange in the give-some dilemma ($1 - 5$). As self-reported measure, participants rated the perceived trustworthiness of the agent, using trustworthiness and expertise items on a 5-point Likert scale (Fogg & Tseng, 1999). The items were combined into a single

score, Cronbach's $\alpha = .94$. Finally, team performance was computed as the number of completed rows per round.

Procedure

Participants met the experimenter, completed informed consent, and received written instructions. The experimental part was identical to the validation study, but the following adjustments were made. First, participants were asked to cooperate with the agent toward the goal. Additionally, the agent was described as cooperative agent (the study materials used the label 'computer'). These changes were made in order to establish a more explicit cooperative frame. Second, to assess behavioral trust, participants played the give-some dilemma with the agent after the puzzle game.

5.3 RESULTS

Perceived warmth and competence

A principal component analysis (PCA) with varimax rotation was conducted on the social perception judgments. The results are shown in Table 4. Four components emerged, accounting for 71.75% of the variance. Three components had sufficient reliability. The first component, labeled *Warmth*, accounted for 45.29% of the variance, Cronbach's $\alpha = .94$. The second component, labeled *Competence*, accounted for 16.64% of the variance, Cronbach's $\alpha = .88$. The third component accounted for only 5.37% of the variance, Cronbach's $\alpha = .70$. Based on the scree plot showing a point of inflection at the third component, only the first two components were retained.

Although the semantic differential used to infer warmth and competence attributes had a more contemporary source, a number of attributes with high loadings on the *Warmth* and *Competence* components are semantically similar to the elementary *good-/bad-social* and *good-/bad-intellectual* trait clusters identified by Rosenberg et al.: 'honest', 'modest', 'warm' ('cold'), 'intelligent' ('unintelligent'), 'alert', 'boring', 'dominant' ('submissive') (Rosenberg et al., 1968). Uncorrelated component scores were obtained using the regression method. In order to reduce the probability of Type I error inflation and to account for the relationship between the dependent variables, we ran a single 2×2 MANOVA for the dependent variables warmth, competence, behavioral trust, and trustworthiness.

Warmth was decreased by selfishness and puzzle incompetence. *Warmth* was lower for the selfish ($M = -0.41, SD = 0.77$) than unselfish agent ($M = 0.41, SD = 1.05$), $F(1, 76) = 23.51, p < .001, \eta_p^2 = .24$. *Warmth* was also lower for the incompetent ($M = -0.47, SD = 0.75$) than competent agent ($M = 0.47, SD = 1.01$), $F(1, 76) = 32.22, p <$

Table 4: Principal component analysis of the social perception scale.

Variable	Warmth	Competence	Component	Component
cold–warm	.855			
aloof–compassionate	.845			
rude–kind	.838			
unfriendly–friendly	.836			
threatening–non-threatening	.830			
impolite–polite	.824			
closed–open	.814			
unlikable–likable	.809			
belligerent–peaceful	.795			
unpleasant–pleasant	.768			
dishonest–honest	.754			
arrogant–modest	.705			
unapproachable–approachable	.680			.491
submissive–dominant	−.603		.571	
unbelievable–believable	.576	.465		
unintelligent–intelligent		.860		
unsuccessful–successful		.855		
incompetent–competent	.422	.809		
distracted–alert		.773		
weak–strong		.640	.403	
boring–exciting		.625		
passive–active		.590		
shy–self-confident			.791	
introverted–extroverted			.670	
tense–relaxed				.817
Eigenvalues	11.32	4.16	1.34	1.11
% **of variance**	45.29	16.64	5.37	4.45
Cronbach's α	.94	.88	.70	.61

Note. Rotated component loadings. Loadings $< .400$ are omitted.

$.001, \eta_p^2 = .30$. Furthermore, there was a significant interaction effect between selfishness and puzzle competence on *Warmth*. To deconstruct this interaction, the selfish agent was judged differently based

on whether it played competently, $F(1,76) = 9.57, p < .01, \eta_p^2 = .11$. If the agent played incompetently, *Warmth* was not affected by selfish behavior (simple effects: $F(1,76) = 1.51, p = .22$). This changed when the agent played competently: in this case, *Warmth* was decreased by selfishness (simple effects: $F(1,76) = 31.70, p < .001$).

Competence was affected only by puzzle incompetence. *Competence* was lower for the incompetent ($M = -.47, SD = .84$) than competent agent ($M = 0.48, SD = 0.95$), $F(1,76) = 22.65, p < .001, \eta_p^2 = .23$. See Figure 9 and Table 14 (appendix a) for further information.

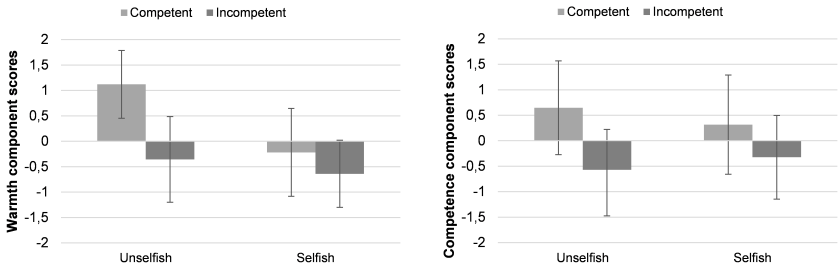


Figure 9: Perceived warmth and competence means. Error bars represent standard deviations.

Behavioral trust and perceived trustworthiness

The behavioral trust and trustworthiness results showed a similar pattern, as shown by Figure 10 and Table 15 (appendix a). Behavioral trust was higher for the unselfish ($M = 3.23, SD = 1.35$) than selfish agent ($M = 2.30, SD = 1.44$), $F(1,76) = 10.34, p < .01, \eta_p^2 = .12$. It was also higher for the competent ($M = 3.25, SD = 1.50$) than incompetent agent ($M = 2.28, SD = 1.26$), $F(1,76) = 11.49, p < .01, \eta_p^2 = .13$. Moreover, there was a significant interaction effect on behavioral trust, $F(1,76) = 4.00, p < .05, \eta_p^2 = .05$. If the agent played incompetently, behavioral trust was not affected by selfishness (simple effects: $F(1,76) = 0.74, p = .39$). However, a different result emerged if the agent played competently: in this case, behavioral trust was decreased by selfishness (simple effects: $F(1,76) = 13.60, p < .001$).

Trustworthiness was higher for the unselfish ($M = 3.07, SD = 1.08$) than selfish agent ($M = 2.35, SD = 0.77$), $F(1,76) = 26.79, p < .001, \eta_p^2 = .26$. Trustworthiness was also higher for the competent ($M = 3.36, SD = 0.94$) than incompetent agent ($M = 2.07, SD = 0.55$), $F(1,76) = 86.02, p < .001, \eta_p^2 = .53$. There also was a significant interaction effect on trustworthiness, ($F(1,76) = 16.31, p < .001, \eta_p^2 = .18$). If the agent played incompetently, trustworthiness was not affected by selfishness (simple effects: $F(1,76) = 0.65, p = .42$). In contrast, if the agent played competently, trustworthiness was decreased by selfishness (simple effects: $F(1,76) = 42.46, p < .001$).

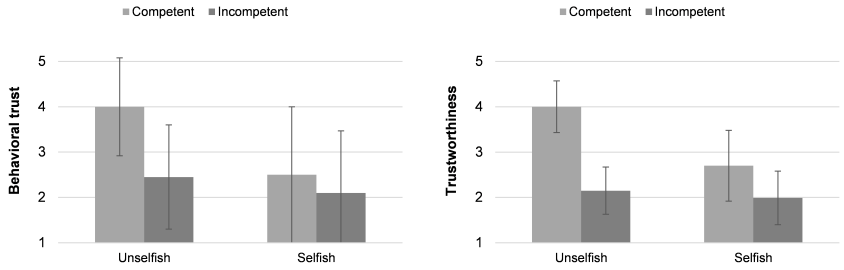


Figure 10: Behavioral trust and trustworthiness means. Error bars represent standard deviations.

To analyze if the effect of the agent's puzzle competence and selfishness on trust was statistically mediated by warmth and competence, we used the bootstrapping method by Preacher and Hayes with bias corrected confidence intervals (2008). We ran separate analyses for each combination of attribution (i.e., *Warmth* or *Competence*) and trust measure (i.e., behavioral trust or trustworthiness). The independent variable was selfishness (binary coded: 0, for selfish; 1, for unselfish) for the analyses involving the proposed mediator *Warmth*, and puzzle competence (binary coded: 0, for incompetent; 1, for competent) for the proposed mediator *Competence*, respectively. The analysis confirmed that *Warmth* statistically mediated the relationship between unselfishness and behavioral trust (95% LCI = 0.34, UCI = 1.12) as well as trustworthiness (95% LCI = 0.21, UCI = 0.81), respectively. *Competence* was not as clear a mediator as *Warmth*: *Competence* was increased by puzzle competence and it was related to trustworthiness, but not behavioral trust ($p = .08$). Moreover, although *Competence* was a mediating factor between puzzle competence and trustworthiness (95% LCI = 0.14, UCI = 0.60), the direct effect of puzzle competence on trustworthiness remained significant (95% LCI = 0.61, UCI = 1.31). In sum, *Warmth* statistically mediated the relationship between unselfishness and trust as well as trustworthiness of the agent, whereas *Competence* partially mediated the relationship between puzzle competence and trustworthiness (see Fig. 11).

5.4 DISCUSSION

Computer agents increasingly act along dimensions of compliant and competent behaviors that elicit fundamental warmth and competence attributions. This experiment suggests that people's willingness to trust agents depends on perceived warmth and competence. In line with previous research (Fiske et al., 2007), we found an emphasis of the role of warmth for trust. This can be explained as follows: The underlying perceptions of warmth, that is, intentions, determined whether the agent was judged to mainly participate in cooperative problem-solving, or to additionally seek for selfish outcome maxi-

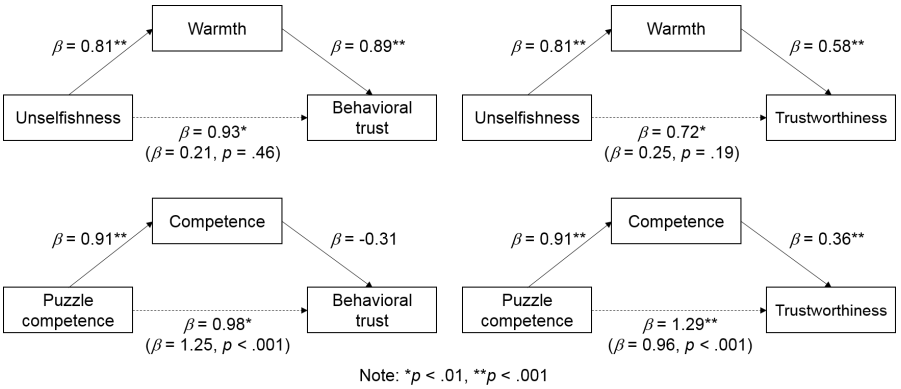


Figure 11: Mediation analyses of unselfishness (puzzle competence) on behavioral trust and trustworthiness, respectively. Betas standardized; 20.000 samples were used for bootstrapping.

mization. Warmth is thus an important antecedent of trust in strategic social interaction. Importantly, similar effects of the manipulations on perceived trustworthiness and behavioral trust were detected which is not always the case in trust studies (Hancock et al., 2011; Salem et al., 2015).

The findings provide further support for the notion that agents are treated as social actors. Indeed, roughly 62% of the variance related to social perception was explained by fundamental dimensions of social cognition, warmth and competence. Attributing these characteristics to an agent is practically relevant for HCI research because it could create agents that elicit trust in cooperation. To the best of our knowledge, we provide the first clear evidence how warmth and competence predict trust in agents. This finding is particularly important because people increasingly rely on agents for problem-solving in their everyday lives. Agents no longer merely execute human orders; they proactively recommend directions, travel targets, as well as products, correct and complete human input before processing it, and overall align to our needs. Going beyond this, they also increasingly mimic human appearance and behavior and manipulate their trustworthiness in accordance with how humans develop trust (DeSteno et al., 2012; Lee et al., 2013). Using the interactive cooperation game paradigm, we were able to highlight the importance of how humans develop warmth and competence attributions for the design of successful interactions. In particular, these attributions seemed to be in line with the general foundation of trust evaluations, that is, being sensitive to "will-do" and "can-do" characteristics of other social agents (Colquitt et al., 2007). Moreover, this experiment extends work on the development of trust in agents by emphasizing behavioral or performance factors. Previous research focused on inferences drawn from artificial emotion expressions (Melo et al., 2014), nonverbal behavior (DeSteno et al., 2012), human-likeness (Kiesler et al., 1996),

reciprocity (Sandoval et al., 2015), and agency (Melo et al., 2015). The interactive cooperation game paradigm demonstrates that the behavioral preconditions of trust in agents such as performance (i.e., competence) (Hancock et al., 2011; Lee & See, 2004) and perceived intentions (i.e., compliance) are translated by humans into warmth and competence attributions which, in turn, determine trust. Considering the role of social cognition for trust could provide a starting point to how other largely understudied aspects of social cognition affect trust, including empathy (Frith & Singer, 2008) as a facilitator of cooperation (Batson & Ahmad, 2001; Batson & Moran, 1999), and categorical thinking and stereotyping (Macrae & Bodenhausen, 2000) as source of warmth stereotypes (Cuddy et al., 2007).

The results also speak to the ongoing debate of human–human versus human–computer trust (Madhavan & Wiegmann, 2007). While some argue that both forms share the same underlying mechanisms (Reeves & Nass, 1996), others maintain that trust in agents is different from trust in people (Lee & See, 2004). This experiment's contribution is that it found supporting evidence for agents being judged along the same fundamental dimensions of social cognition as humans. To further clarify the impact of this experiment's contribution, comparisons of interpersonal and human–computer trust should also encompass warmth and competence attributions, including their influence on behavior. Previous work has already provided evidence for the relevance of warmth and competence attributions as underlying dimensions of social perception in HCI (Bergmann et al., 2012; Niewiadomski et al., 2010).

From a game theoretic perspective, selfishness in the puzzle game is similar to a safety preference in the stag hunt game. In the puzzle game, selfishness promotes selfish goals. In the stag hunt game, two agents individually choose between hunting a hare and a stag, with the stag generating higher payoff than the hare. To hunt the stag (i.e., the joint goal), one agent requires the willingness of the other agent to coordinate and participate in the hunt; to hunt the hare, no cooperation is required. Interactive adaptations of the stag hunt game let human players coordinate with computer agents by navigating shapes through a two-dimensional board (Yoshida et al., 2010). Indeed, the puzzle game drew inspiration from such strategic games. Going beyond this, it attempts to investigate the interplay of warmth and competence as critical factor for cooperation and trust among social agents – a domain that is difficult to model in pure game theoretic terms. Advanced agents such as robots are undergoing evolutionary processes pertaining to roles (tools vs. assistants, companions), functionalities (e.g., learning new competencies, be adaptive), and the social distance to humans (no contact vs. long-term contact) (Dautenhahn, 2007). This should not only entail the facilitation, but rather the calibration of trust to avoid maladaptive behavior. With increasingly

intelligent and complex agents, a match between elicited trust and capabilities make human reliance on agents safe (Lee & See, 2004). After all, humans navigate through their social lives by categorizing others along social dimensions, triggering specific cognitive contents (Macrae & Bodenhausen, 2000). This experiment emphasizes the role of transparent intentions for this process, and is another step toward a coherent picture of how people perceive and interact with intelligent agents.

6

EXPERIMENT 2: TRUST FORMATION

The previous experiment showed how perceived warmth can become a foundation of trust and trustworthiness attributions. Perceived competence could statistically not explain the relation between agent actions to solve the puzzle and resulting trust(-worthiness), indicating that other factors influenced this relation. Indeed, targets that elicit negative affect may cause perceivers to outright ignore their task competence (Casciaro & Lobo, 2008). Given the actual significance of competence and performance for trust (Lee & See, 2004; Mayer et al., 1995), the role of competence should be more evident under different parameters. The individual scores in the first experimental setting ultimately highlighted the agent's intention to pursue its desire for an individual goal, hence the relative importance and functional role of warmth for trust.

Cooperation does not always rely as heavily on perceived intentions as in the previous experiment. A number of reasons call for a different approach regarding the interaction, and the interactive cooperation game paradigm permits these flexible rearrangements. The motivation to alter the interaction is explained in the following.

Strategic considerations may be important for cooperation in situations of conflicting interests, like social dilemmas. However, current human–computer trust research is often centered on a more tangible issue, that is, *advice adoption*. Advice adoption studies analyze how people respond to computer-generated advice presented by intelligent agents. Since the decision to adopt computerized advice is understood either as direct consequence or behavioral form of trust, advice adoption studies are often conceptualized as trust studies (see Hoff & Bashir, 2015; Madhavan & Wiegmann, 2007, for an overview). In these scenarios, implementing individual interests would change the factors that determine advice adoption: trust evaluations would largely depend on perceived intentions in terms of benevolence and integrity, whereas the more interesting focus lies on whether why and how people accept advice as external input. Accordingly, individual interests are typically not the main issue in advice adoption. Rather, it is hypothesized that characteristics of the agent and their advice affect trust formation.

One such agent characteristic is anthropomorphism. Technology that displays anthropomorphic characteristics is becoming increasingly influential. For instance, anthropomorphic interfaces can serve as online motor skill coaches (Kok et al., 2015), child tutors (Breazeal et al., 2016), negotiation partners (Gratch et al., 2016), or assistants

for target groups with special needs (Yaghoubzadeh et al., 2013). The underlying goal is to deliberately design technology so as to imitate human-like characteristics and engage people in rich interactions, but the role of anthropomorphism for cooperative interaction is not yet clear. In a recent study, the difference between trust in an anthropomorphic avatar versus a non-anthropomorphic computer remained somewhat unclear (Visser et al., 2016). The overall conclusion was that anthropomorphism indeed influences trust, leading to higher resistance to trust breakdowns, yet the most pronounced differences occurred between the human and both computerized (i.e., computer and avatar) partners. In this particular setting, the impact of an anthropomorphic avatar on trust-related and performance variables was difficult to pinpoint. Across three experiments, subjective trust in the avatar was sometimes higher than in the human partner, at other times it was the same. Compliance with avatar advice was never different from the other agents. Overall, the increase of applications with complex interaction goals suggests an important potential role of anthropomorphic agents, but if such agents support trust is not yet fully clear.

6.1 OVERVIEW

Experiment 2 encapsulates the role of agent and advice characteristics for trust formation by investigating how anthropomorphism and advice quality relate to advice adoption (Kulms & Kopp, 2016). The research question is as follows:

> **RQ**: Does anthropomorphism influence trust formation?

Experiment 2 makes the following **contributions** to the present thesis' goals:

- The interaction is tailored to investigate how people form trust in intelligent agents over time and respond to advice presented by these agents.
- Trust is operationalized as advice adoption rate, providing a more practical trust measure in problem solving.

6.2 METHOD

Participants

Sixty German undergraduate and graduate students participated in exchange for 5 EUR. The sample ranged in aged from 18 to 44 years ($M = 24.85$, $SD = 4.86$, $median = 24$; 60.7% female).

Task

Participants played an altered version of the puzzle game. The game proceeds in alternating turns. In each turn both players draw one of two available blocks, a T-block and a U-block, from an urn without replacement. The blocks no longer generate individual payoff. The agent first chooses a block, leaving the remaining block to the participant. The joint goal is to complete a specific number of rows. Again, completed rows are not emptied and there is no time restriction. Completing a row yields 100 points for each player. Attaining the joint goal yields a joint payoff such that the score gets doubled for each player. Thus, the payoff for both players is always identical. Participants were instructed to work toward the joint goal together with their partner. They were also told that throughout the game, their partner would offer advice as to how they could place their block. The interaction lasted three rounds in total with the goal becoming increasingly difficult each round (Round 1: 4 rows, Round 2: 5 rows, Round 3: 6 rows).

Agent behavior

Three times per round, the agent offered advice to participants (see Fig. 12). This was initiated by the agent saying "I have a suggestion, do you want to see it?" or "I think I know a solution, should I show you?", if the virtual agent was present. Next, two buttons appeared, labeled "Show me the suggestion" or "I do not want the suggestion", respectively. Without virtual agent, the buttons appeared without introduction. In order for the game to continue, participants had to make the decision to either request or ignore the advice. Across all conditions, the agent placed blocks in a competent fashion.

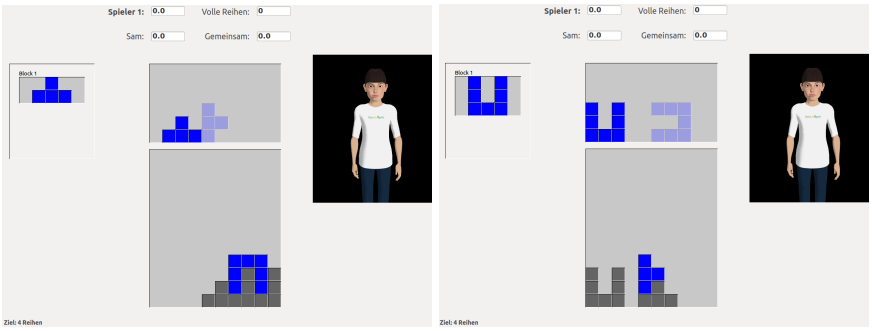


Figure 12: Game interface with virtual agent. Suggestions by the agent are shown at the top in light blue. Left: the agent makes a useful suggestion. Right: the agent makes a bad suggestion.

Design

The study had a 2×2 between-subjects design, with agent (computer vs. virtual agent) and advice quality (good vs. bad) as between-subjects factors. Participants' partner was either a virtual agent called Sam with anthropomorphic appearance and voice, or a regular computer agent without any anthropomorphic cues. The agent provided either good or bad advice.

Measurement

In order to asses the task outcome, a team performance score was computed by analyzing how often the joint goal was achieved in each round ($0 - 100\%$). Behavioral trust was the number of times the agent's advice was followed in each round ($0 - 3$). Another behavioral measure was how often participants requested advice in each round ($0 - 3$). Self-reported perceived trustworthiness was measured using trustworthiness and expertise items on a 5-point Likert scale (Fogg & Tseng, 1999). The items were combined into a single score, Cronbach's $\alpha = .90$.

Procedure

Participants met the experimenter, completed informed consent, and received written instructions. The instructions described the game and introduced player two: a virtual person named Sam (anthropomorphism conditions), or a computer without name (no anthropomorphism conditions), respectively. After each round, participants were given a summary showing whether the goal was attained. Before the interaction, participants familiarized themselves with the controls and mechanics without the agent.

Data analysis

All dependent variables were entered into a 2×2 MANOVA.

6.3 RESULTS

Team performance

Team performance did not significantly differ between the conditions. See Table 5.

Table 5: Means and standard deviations of team performance (goal achievement).

		Round 1		Round 2		Round 3	
Advice quality	*Agent*	M	SD	M	SD	M	SD
Good	Computer	0.77	0.44	0.69	0.48	0.46	0.52
	VA	0.81	0.40	0.63	0.50	0.50	0.52
Bad	Computer	0.67	0.49	0.47	0.52	0.40	0.52
	VA	0.67	0.49	0.53	0.52	0.27	0.46

Note. VA = Virtual agent.

Perceived trustworthiness

There was a significant main effect of advice quality on trustworthiness. The agent was ascribed higher trustworthiness given good ($M = 3.36, SD = 0.66$) versus bad advice ($M = 2.51, SD = 0.70$), $F(1, 56) = 22.35, p < .001, \eta_p^2 = .29$.

Requested and adopted advice (behavioral trust)

There were significant main effects of agent on requested advice, and these effects diminished in later rounds. The virtual agent led to more requested advice in Round 1, $F(1, 56) = 4.92, p < .05, \eta_p^2 = .08$, and with marginal significance in Round 2, $F(1, 56) = 3.18, p < .09, \eta_p^2 = .05$. This effect disappeared in Round 3, $F(1, 56) = 1.89, p = .18$. Advice quality had less influence on requested advice. In Round 1 there was a marginally significant main effect of advice quality on requested advice, $F(1, 56) = 3.53, p < .07, \eta_p^2 = .06$. In this round, participants requested somewhat more bad than good advice.

There were consistently significant main effects of advice quality on the amount of adopted advice. In Round 1, $F(1, 56) = 22.60, p < .001, \eta_p^2 = .29$, Round 2, $F(1, 56) = 42.25, p < .001, \eta_p^2 = .43$, and Round 3, $F(1, 56) = 34.45, p < .001, \eta_p^2 = .38$, participants adopted more good than bad advice. There were no significant main effects of agent on adopted advice.

Figure 13 and 14 as well as Table 16 (appendix a) and 17 (appendix a) show these results.

Time effects

Separate ANOVAs with round number as within-subjects factor and agent as well as advice quality as between-subjects factors were conducted. For the between-subjects effects, the dependent variables (team

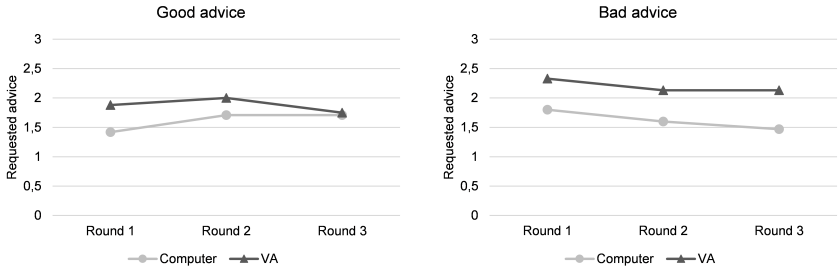


Figure 13: Requested advice.

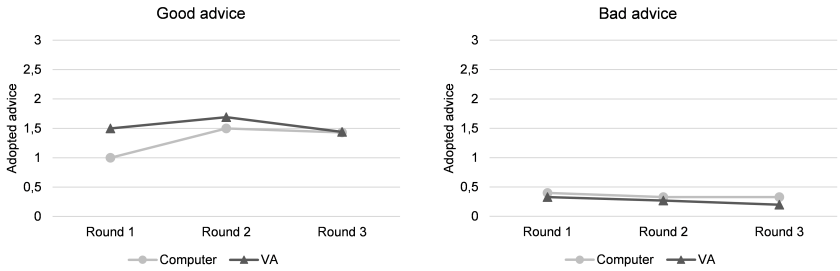


Figure 14: Adopted advice (behavioral trust).

performance, requested advice, adopted advice) are averaged across the three rounds to form a global variable. This complements the previous analysis, where team performance was analyzed per round, not globally.

As the goal became increasingly difficult, there was a significant main effect of round number on team performance, $F(2, 110) = 6.42, p < .01, \eta_p^2 = .11$. Planned contrasts reveal that team performance worsened over time. Team performance in Round 2 was worse than Round 1, $F(1, 55) = 2.83, p < .10, \eta_p^2 = .05$, worse in Round 3 than Round 2, $F(1, 55) = 3.43, p < .07, \eta_p^2 = .06$, and worse in Round 3 than Round 1, $F(1, 55) = 13.50, p < .01, \eta_p^2 = .20$. On the global level, there also was a significant main effect of advice quality on team performance, $F(1, 55) = 3.80, p < .06, \eta_p^2 = .07$, indicating that the teams performed better with good ($M = .64, SE = .05$) compared to bad advice ($M = .50, SE = .05$).

Significant between-subjects effects of advice quality on advice requests and adoption on the global level confirmed the findings reported above. Participants requested overall more advice from the virtual agent ($M = 2.04, SE = .14$) than the computer ($M = 1.62, SE = .14$), $F(1, 56) = 4.54, p < .05, \eta_p^2 = .08$, and they adopted overall more good ($M = 1.43, SE = .10$) than bad advice ($M = 0.31, SE = .10$), $F(1, 56) = 61.63, p < .001, \eta_p^2 = .52$.

Relationship between trustworthiness and behavioral trust measures

Measure interrelations show that trustworthiness and behavioral trust were related in Round 1 and Round 2 (see Table 6). There were moderately strong correlations between the number of requested and adopted advice in a given round, except for Round 1 (see colored cells). However, note that the correlations were somewhat biased to the extent that the number of requested advice could logically not surpass adopted advice.

Table 6: Pearson correlation coefficients between trustworthiness and behavioral measures.

	TW	REQ (R1)	REQ (R2)	REQ (R3)	BTR (R1)	BTR (R2)	BTR (R3)
REQ (R1)	−.16	-	-	-	-	-	-
REQ (R2)	.04	.64**	-	-	-	-	-
REQ (R3)	−.03	.54**	.48**	-	-	-	-
BTR (R1)	.38*	.21	.19	.09	-	-	-
BTR (R2)	.37*	.11	.41*	.06	.57**	-	-
BTR (R3)	.20	.06	.14	.38*	.53**	.56**	-

Notes. TW = Trustworthiness, REQ = Requests, BTR = Behavioral trust, R = Round
$p < .01$*, $p < .001$**

Relationship between team performance, trustworthiness, and behavioral measures

Team performance did not significantly correlate with any of the other variables.

6.4 DISCUSSION

The goal of this experiment was to examine the role of anthropomorphism for trust formation. As intelligent agents become increasingly adept in collaborating with others, human users need to decide whether to trust them in novel interactions. While there is a growing body of research on trust and cooperation between humans and computer agents in simplified scenarios, this experiment focused on interactive and continued decision-making. Participants tried to solve the puzzle within an advice adoption paradigm. In this cooperative scenario, participants received task-related advice from an agent partner three times per round, at fixed occasions. At the same time, both players placed blocks in an alternating fashion. Trust in the agent was measured in the form of the advice adoption rate.

The results show that anthropomorphism was not related to trust formation. Neither behavioral trust (i.e., advice adoption) nor perceived trustworthiness were affected by anthropomorphism. This finding stands in contrast to previous research showing that anthropomorphism increases trust (Waytz et al., 2014). Other research found similarly positive effects of anthropomorphism on cooperation, but those studies incorporated humans or human-controlled avatars in the anthropomorphism conditions (Melo et al., 2015; Miwa et al., 2008; Sandoval et al., 2015), which is why direct comparisons could be misleading.

One possible reason for the missing influence of the virtual agent on trust is that participants were not socially influenced by the anthropomorphic cues. Importantly, anthropomorphism did not make the agent seem more trustworthy. In theory, the adaptive advantage of identifying trustworthiness from anthropomorphism is in line with the warmth and competence concept (Fiske et al., 2007; Judd et al., 2005) and the evolution of human cooperation (Jordan et al., 2016), because it allows access to crucial information as to the intentions of others. The contribution of the present results is that the on-screen presence of an artificially anthropomorphic agent in conjunction with neutral, unpersuasive verbal advice offerings do not exert influence on trust and trustworthiness. For a more comprehensive investigation, more variations of anthropomorphism need to be considered. After all, the virtual agent used in this experiment was a rough approximation of human-like appearance and voice.

Another issue may have been missing feedback about the specific positive consequences of advice adoption for the goal. Direct positive feedback could have reinforced advice adoption in the long run, but the cooperative task dampened the possibility for feedback because it did not involve immediate right or wrong decisions. Rather, even with low advice adoption, goal achievement was possible if both players acted competently. To issue a caveat, efforts to reinforce advice adoption using anthropomorphic cues could actually make matters worse in the long run if participants fail to achieve the goal, because the reinforcement may create increased expectations which are violated by goal failure. Using cues such as anthropomorphism to reinforce human behavior should only be considered if other cues can help regulate negative outcomes.

Trust and trustworthiness attributions were amplified by competence of the agent in the form of good advice. Across three rounds, participants consistently based advice adoption on whether the agent gave good advice. This finding is in line with the notion that performance is a critical key antecedent of trust in agents (Hancock et al., 2011; Lee & See, 2004). Overall, participants formed appropriate trust in the agent to the extent that bad advice was rejected. Perceived trustworthiness was positively related to adoption in the first two rounds,

indicating that advice adoption is a valid operationalization of trust. Over time, there was little variation of trust in advice. This could indicate that the degree of trust was rather stable and insusceptible to changes, even though the goal became gradually more difficult.

Although anthropomorphism was not related to advice adoption, it influenced another important interaction factor. Participants requested more advice from the virtual agent, and this effect was stronger in earlier rounds. Advice requests laid the ground for trust decisions, that is, to adopt or reject advice. Since the requests did not correlate with perceived trustworthiness, it is not fully clear how anthropomorphism influenced advice requests. Some possible factors why participants felt a stronger need to consult the virtual agent are context (i.e., momentary task difficulty), self-confidence, and mere curiosity. Interestingly, the number of requests in the good and bad advice conditions were roughly at the same level. Bad advice did not decrease advice requests; in fact, participants requested even somewhat more bad than good advice in the beginning of the interaction. A number of reasons can explain this finding. It seems that participants were consistently motivated to evaluate the advice, even after their first requests indicated that they were not particularly helpful. Participants could have been forgiving, or were affected by the agent's competent puzzle-solving actions. After all, competent block placements were the only necessary and sufficient condition of high team performance. Moreover, requests had no associated costs. There were no time constraints or any other punishments such that participants had nothing to lose by requesting advice.

The experiment has a number of limitations. The low sample size could have decreased the likelihood of detecting interaction effects in the 2×2 design. Furthermore, anthropomorphism varied only between a virtual and computer agent. It is unclear if a real human agent may have had a stronger influence on trust (e.g., Visser et al., 2016). Finally, participants could not request advice as needed but were confined to fixed moments.

How did the change in the interaction paradigm relate to trust? The advice adoption scenario affects the cooperative framing and factors of human–computer trust. The agent now clearly acted as cooperative partner without motivation to compete with participants. The scenario introduced new important trust-relevant cues in the form of agent advice and advice quality. Hence, participants had access to more and qualitatively different trust-relevant cues. These new cues, however, did not affect the performance of the human–agent teams, possibly due to the limitations mentioned above. Anthropomorphism was hypothesized to influence trust, but this relation was not confirmed. Instead, the effect of anthropomorphism on requests implies that in the context of continued cooperation, a comprehensive inves-

tigation of all relevant interaction variables related to trust decisions is necessary.

7

EXPERIMENT 3: TRUST FORMATION (EXTENDED)

Experiment 2 provided only a first step toward the effects of anthropomorphism on trust. A critical shortcoming was the anthropomorphism manipulation, which only incorporated a virtual agent but no human agent. For a comprehensive investigation of the role of anthropomorphism for trust, real human agents are needed. Only the comparison with human agents enables a full perspective regarding the differences and similarities between human–human and human–computer trust. Furthermore, it is still unclear if virtual agents gain user trust either in a manner similar to humans, or non-embodied computers.

7.1 OVERVIEW

Experiment 3 provides a direct comparison between human and non-human agents in the advice adoption scenario. Several additional improvements are introduced. In particular, participants are given more freedom regarding advice requests. They also receive more task feedback, feedback about their progress toward the goal, and are faced with greater uncertainty in the last round. Apart from this, Experiment 3 has the same cooperative structure as Experiment 2. As an extension of the previous study, this experiment investigates the same research question:

> **RQ**: Does anthropomorphism influence trust formation?

Experiment 3 makes the following **contributions** to this thesis' goals:

- By including agents with varying degrees of anthropomorphism, including human agents, the experiment contributes to the ongoing debate regarding the similarities and differences between trust in computers and trust in humans, which helps to further characterize human–computer trust.

7.2 METHOD

Participants

Participants were 114 CrowdFlower contributors from Germany, Austria, and Switzerland, aged between 18 and 58 years (M = 30.45, SD = 11.05; 41.2% female). They received 0.50 EUR for completing the experiment and could earn additional 15 EUR based on how well they played.

Participants were excluded if their total session time was unrealistically short or if they failed to correctly answer a full sequence of test questions regarding the game background and mechanics.

Task

The puzzle game used in this experiment (see Fig. 15) provided participants with a 2-player problem-solving scenario where they could ignore, request and decline, or request and adopt advice of an artificial agent. The interaction proceeds in alternating turns. In each turn, both players draw one of two available blocks, a T-block and a U-block. Also in each turn, the agent first chooses a block, leaving the remaining block to the participant. The joint goal is to complete a specific number of rows such that it is entirely filled with blocks. A round ends when there is no space left for any other block.

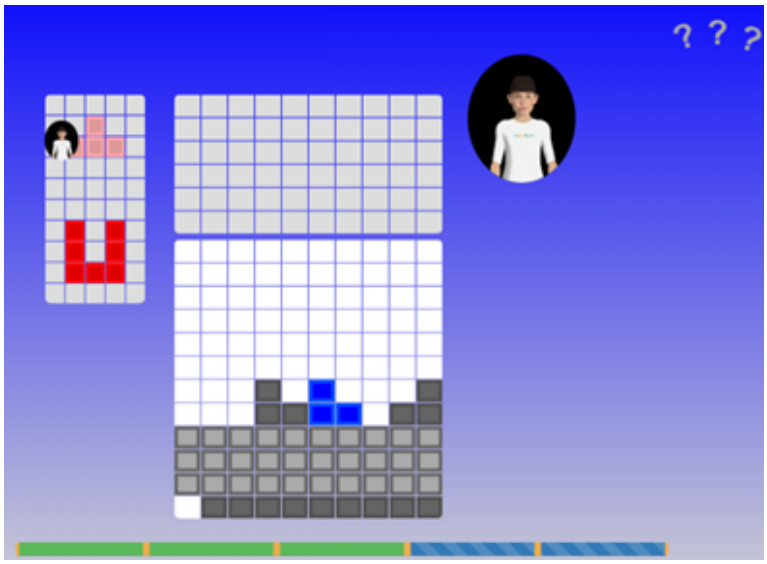


Figure 15: Web-based puzzle game interface in the virtual agent condition. The progress bar at the bottom shows the progress toward the goal. The question marks in the top right corner show how many requests are left. The icon on the left displays the block the agent has picked (T-block).

Participants were instructed to work toward the joint goal with the agent. Importantly, they were told they could request advice by their partner up to three times per round. The advice was displayed in the block area. The number of remaining advice was shown in the top right corner (see Fig. 16). The interaction lasted three rounds. In each round, the goal became increasingly difficult (Round 1: five rows, Round 2: six rows, Round 3: eight rows). In Round 3, the advice pattern was changed such that the agent would no longer show the advice first. Instead, participants could yield up to three of their turns completely to the agent. The agent would automatically per-

form the action, thereby increasing uncertainty as participants could not evaluate the suggested solution first. This mechanic was introduced to increase the associated risk of the trusting decision in the final round. After each round, participants were given a brief summary as to whether the goal was attained. Before the experimental trials, participants familiarized themselves with the controls and mechanics in a trial round without the agent.

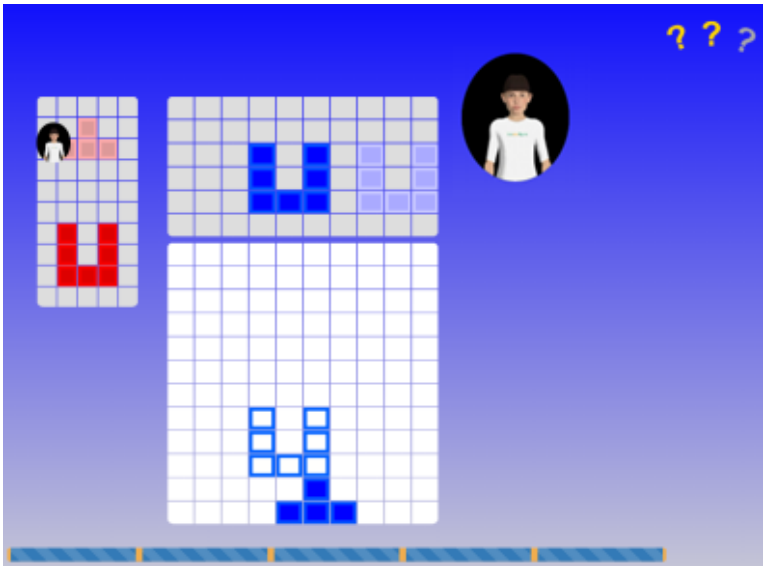


Figure 16: Agent advice is displayed in light blue in the block area at the top. The puzzle area now shows a preview of where the U-block would be placed, given the current configuration. In this case, it would be placed on top of the T-block.

As Figure 15 and 16 show, the puzzle game interface was modified to increase the user experience (e.g., more task feedback through additional information) in this web-based version of the game.

Agent behavior

Audiovisual utterances for the virtual and human agent with matching content were created in order to create the illusion of a social agent capable of planning and acting within the task environment. The utterances were shown as short video clips. Prior to the experimental trials, the agent briefly introduced itself: "Hello, I am Sam. I hope you are ready to puzzle." When participants requested advice, the agent explained to think of a solution ("Let me think") and said "How about this?" when the advice was displayed. When the joint goal was attained, the agent cheerfully said "Hey, we reached the goal." In contrast, when the goal was not attained the agent was disappointed: "Unfortunately, we did not reach the goal." At the end of the game, participants were thanked for participation.

For the human agent condition, a male person was filmed. The male was instructed to act as someone who gives advice to another person by saying the phrases described above. The virtual agent was programmed to utter the same phrases. For the computer agent condition, a symbolic icon was used. Figure 17 shows the different agents used in the experiment.

Figure 17: The computer, virtual, and human agent.

Design

The experimental design was a 3×2 between-subjects design, with agent (computer agent, virtual agent, human agent) and advice quality (good, mixed) as between-subjects factors. In the mixed advice condition, the agent gives bad advice in the *first* round, followed by good advice and solutions in the second and third round.

Measurement

A team performance measure was computed as indicator of whether agent and advice quality affected the task outcome. Specifically, we computed the number of completed rows participants achieved per round. The main task was to play as efficiently as possible, by completing increasing amounts of rows. Since all agents played equally competently, this variable approximates team performance adequately.

Behavioral and self-reported measures of trust were obtained. Self-reported trust was assessed by asking participants how much they trusted the agent using a single item. Behavioral trust was the number of times participants exactly followed the agent's advice in each round (Round 1 & 2), complemented by the number of times they passed on their turn to the agent (Round 3). It was also computed how often participants requested advice in the first place (Round 1 & 2). Self-reported perceived trustworthiness was measured using trustworthiness and expertise items on a 5-point Likert scale (Fogg & Tseng, 1999). The items were combined into a single score, Cronbach's $\alpha = .92$. Finally, to infer warmth and competence attributions from social perception, the same semantic differential as in Experiment 1 was used.

Procedure

Participants signed up for the experiment as registered contributors on the crowdsourcing platform CrowdFlower. The experiment was listed as a human–computer study involving a cooperative 2-player puzzle game. A possible concern with crowdsourcing studies is that participants may already be familiar with popular study materials. However, this does not apply to our materials as they have never been used in crowdsourcing studies to date.

Participants read about the game mechanics, the goal, and the role of the computer as a partner in problem solving as well as source of task-related advice. Participants' main instruction was to play as efficiently as possible. After completing a training round without the agent to familiarize with the gameplay, participants completed three experimental rounds followed by a post-questionnaire containing the social perception measures. Finally, participants were debriefed and received the payment for their participation.

Data analysis

All dependent variables were entered into a 3×2 MANOVA with Bonferroni corrected Post-hoc tests for the factor agent.

7.3 RESULTS

Team Performance

There were no significant differences for Round 1 and 2. For Round 3 there was a main effect of advice quality, $F(1, 108) = 6.29, p < .05, \eta_p^2 = .06$. Participants completed more rows with the agent that provided overall good advice (see Table 7).

Table 7: Means and standard deviations of team performance.

		Round 1		Round 2		Round 3	
Advice quality	*Agent*	M	SD	M	SD	M	SD
Good	Computer	5.74	1.20	5.63	1.77	5.95	1.47
	VA	4.95	1.84	5.79	2.00	5.63	1.21
	Human	5.37	1.07	5.42	2.20	5.47	1.54
Mixed	Computer	5.26	2.33	5.00	1.37	4.53	1.87
	VA	5.32	1.64	5.32	2.14	5.00	2.13
	Human	5.95	1.18	5.42	1.74	5.26	1.20

Note. VA = Virtual agent.

Perceived warmth and competence

A PCA with varimax rotation was conducted on the social perception judgments (see Table 8). Five components emerged, accounting for 64.30% of the variance. Three components had sufficient reliability. The first component, labeled *Warmth* accounted for 35.65% of the variance, Cronbach's $\alpha = .90$. The second component, labeled *Competence* accounted for 12.35% of the variance, Cronbach's $\alpha = .87$. The third component accounted for only 7.19% of the variance, Cronbach's $\alpha = .79$. Based on the scree plot showing a point of inflection at the third component, only the first two components were retained. Uncorrelated component scores were obtained using the regression method.

There was a significant main effect of agent on *Warmth*, $F(2, 108) = 24.41, p < .001, \eta_p^2 = .31$. Participants attributed more *Warmth* to the human than the computer ($p < .001$). They also attributed more *Warmth* to the human than the virtual agent ($p < .001$). A statistical trend shows that the virtual agent was attributed more *Warmth* than the computer ($< .10$). There were no significant differences with respect to *Competence*. These results are shown in Figure 18.

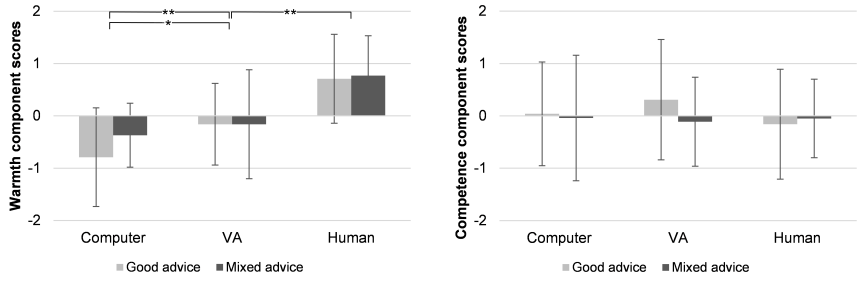


Figure 18: Perceived warmth and competence means. Error bars show standard deviations.
$p < .07$*, $p < .001$**

Requested and adopted advice (behavioral trust)

There was a significant main effect of advice quality on requests such that participants requested more advice given good than mixed advice in Round 1, $F(1, 108) = 9.82, p < .01, \eta_p^2 = .08$, and Round 2, $F(1, 108) = 6.98, p < .01, \eta_p^2 = .06$. Recall that in Round 3, participants could no longer request advice but were able to pass on their turns. The main effect of agent on requests in Round 1 and 2 was not significant. Figure 19 and Table 18 (appendix a) show these results.

The main effect of agent on adopted advice in Round $1-3$ was not significant. Instead, there was a significant main effect of advice quality on adopted advice in Round 1, $F(1, 108) = 48.57, p < .001, \eta_p^2 = .31$, Round 2, $F(1, 108) = 6.57, p < .05, \eta_p^2 = .06$, and Round 3, $F(1, 108) =$

Table 8: Principal component analysis of the social perception scale.

Variable	Warmth	Competence	Component	Component	Component
cold–warm	.781				
aloof–compassionate	.764				
closed–open	.753				
unlikable–likable	.752				
unfriendly–friendly	.723		.446		
rude–kind	.711				
unapproachable–approachable	.676				
unpleasant–pleasant	.642				
arrogant–modest	.637				
threatening–non-threatening	.468	.401			
dishonest–honest	.465	.416			
incompetent–competent		.826			
unsuccessful–successful		.733			
untrustworthy–trustworthy		.722			
unintelligent–intelligent		.711			
distracted–alert		.590			
weak–strong		.590			
boring–exciting		.576			
tense–relaxed			.768		
impolite–polite	.572		.621		
belligerent–peaceful	.466		.602		
introverted–extroverted				.742	
shy–self-confident				.685	
passive–active		.438		.610	
submissive–dominant	−.495				−.658
Eigenvalues	8.91	3.09	1.78	1.27	1.01
% of variance	35.65	12.35	7.19	5.06	4.05
Cronbach's α	.90	.87	.79	.66	-

Note. Rotated component loadings. Loadings < .400 are omitted.

$9.99, p < .01, \eta_p^2 = .09$, such that trust in good advice was higher compared to mixed advice. Figure 20 and Table 19 (appendix a) show these results.

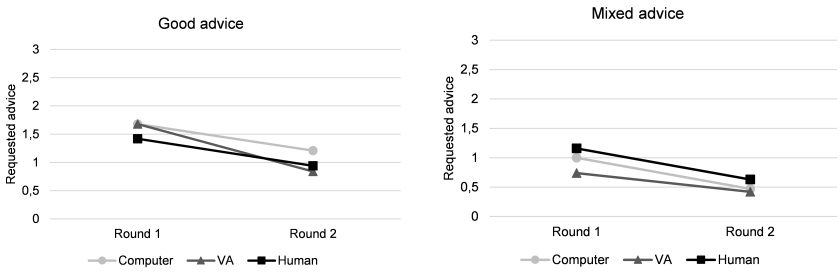


Figure 19: Requested advice.

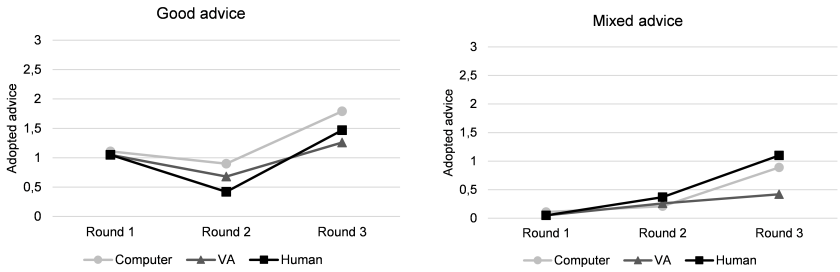


Figure 20: Adopted advice (behavioral trust).

Self-reported trust

A marginally significant main effect of agent on trust ($F(2, 108) = 2.89, p = .06, \eta_p^2 = .05$) revealed that participants reported somewhat more trust for the human than the computer ($p < .07$). There also was a significant main effect of agent on trustworthiness, $F(2, 108) = 9.60, p < .001, \eta_p^2 = .15$. Participants rated the computer as less trustworthy than the virtual agent ($p < .05$) and the human ($p < .001$), respectively. There was no difference between the virtual agent and the human ($p = .34$). These results are shown in Figure 21.

Time effects

Separate ANOVAs with round number as within-subjects factor and agent as well as advice quality as between-subjects factors were conducted. The round number was limited to Rounds 1 and 2 for requested and adopted advice due to the interaction change in Round 3. For the between-subjects effects, the dependent variables (team performance, requested advice, adopted advice) are averaged across the three rounds to form a global variable.

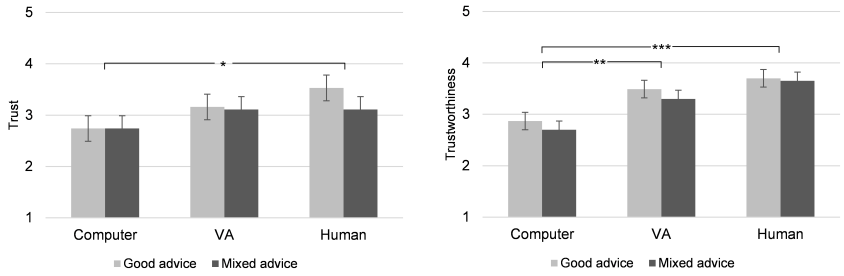


Figure 21: Self-reported trust and perceived trustworthiness ratings. Error bars show standard errors.
$p < .07$*, $p < .05$**, $p < .001$***

There was a significant main effect of round number on requested advice, $F(1, 108) = 23.11, p < .001, \eta_p^2 = .18$. Planned contrasts reveal that advice requests decreased over time. Advice requests in Round 1 were higher ($M = 1.28, SE = .10$) than Round 2 ($M = 0.75, SE = .09$). A significant between-subjects effect on the global level only confirmed the findings reported above, that is, participants requested more good ($M = 1.30, SE = .11$) than mixed advice ($M = 0.74, SE = .11$), $F(1, 108) = 12.30, p < .01, \eta_p^2 = .10$.

There was a significant interaction effect between round number and advice quality, on adopted advice, $F(1, 108) = 13.03, p < .001, \eta_p^2 = .11$. The advice adoption development between Round 1 and 2 was different for good advice and mixed advice participants (see Fig. 22). With good advice, advice adoption was higher in Round 1 ($M = 1.07, SE = .10$) than Round 2 ($M = 0.67, SE = .11$), but with mixed advice, adoption was higher in Round 2 ($M = 0.28, SE = .11$) than Round 1 ($M = 0.07, SE = .10$). Recall that in the mixed advice conditions, advice quality changed from bad to good in Round 2.

A significant between-subjects effects of advice quality on advice adoption on the global level confirmed the findings reported above, that is, participants adopted overall more advice given good ($M = 0.87, SE = .09$) than mixed advice ($M = 0.18, SE = .09$), $F(1, 108) = 33.36, p < .001, \eta_p^2 = .24$.

Relationship between self-reported and behavioral trust measures

Measure interrelations show that self-reported and behavioral trust were almost unrelated. Only behavioral trust in Round 3 was related to self-reported trust (see Table 9). In this round, participants' decision to (not) trust required to handle increased uncertainty because the advice pattern was changed to turn takeover by the agent. Accordingly, participants who did pass on their turns showed maximum trust in the agent, which could explain the exclusive relationship between self-reported and behavioral trust in this round.

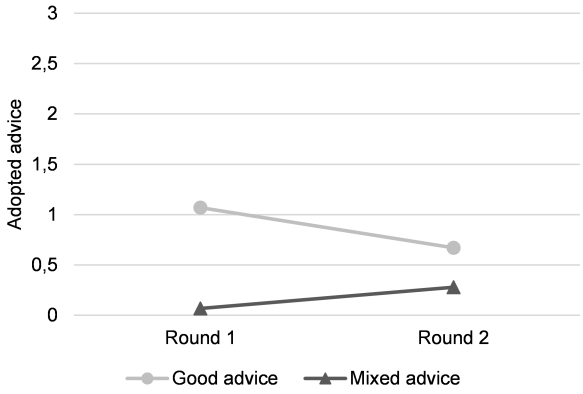


Figure 22: Interaction effect between round number and advice quality on adopted advice.

There were strong correlations between the number of requested and adopted advice in Round 1 and Round 2 (see colored cells). Once participants requested advice in a given round they were likely to adopt it.

Table 9: Pearson correlation coefficients between self-reported and behavioral measures.

	STR	TW	REQ (R1)	REQ (R2)	BTR (R1)	BTR (R2)	BTR (R3)
TW	.57**	-	-	-	-	-	-
REQ (R1)	.07	.01	-	-	-	-	-
REQ (R2)	.00	−.01	.41**	-	-	-	-
BTR (R1)	.07	.10	.61**	.36**	-	-	-
BTR (R2)	.02	−.04	.41**	.80**	.40**	-	-
BTR (R3)	.07	.07	.50**	.38**	.38**	.35**	-

Notes. STR = Self-reported trust, TW = Trustworthiness, REQ = Requests, BTR = Behavioral trust, R = Round
$p < .05$*, $p < .001$**

Relationship between team performance and trust

There was a significant relationship between adopted advice and team performance in Round 2, $r = .20, p < .05$. Furthermore, there was a significant negative relationship between team performance in Round 1 and perceived trustworthiness, $r = -.28, p < .01$ (see Fig. 23). Closer inspection revealed that some participants who achieved high team performance attributed low trustworthiness to the agent. Most of these cases occurred in the mixed advice condition, with bad advice

in this round. With respect to trustworthiness formation, the negative impact of bad advice seemed to outweigh the positive impact of team performance.

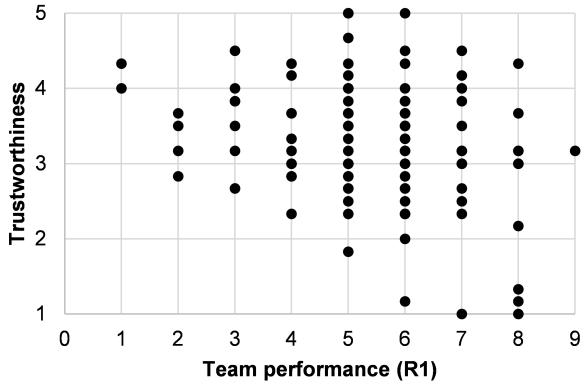


Figure 23: Relationship between team performance (R1) and perceived trustworthiness.

7.4 DISCUSSION

In this second experiment on the role of anthropomorphism for trust formation, participants had the opportunity to ask for advice from an assistant while performing an interdependent problem-solving procedure with it. The results indicate a mismatch between self-reported and behavioral trust. While anthropomorphism could increase self-reported but not behavioral trust, advice quality increased behavioral but not self-reported trust.

The present work speaks to a debate crucial for the design and evaluation of intelligent computer agents. Given that anthropomorphic agents such as humanoid robots and virtual agents often deliberately resemble humans, do people establish trust with them in a manner similar to other humans? While some researchers argue that computers and humans elicit similar social responses (Reeves & Nass, 1996), others claim that trust in computers is different from trust in people due to cognitive biases (Lee & See, 2004). Regarding the subjective trust experience, it was found that humans and computers elicit similar degrees of trust if the computer was designed to be anthropomorphic, but if it was not, humans elicit higher trust. Specifically, participants reported more trust with the human agent compared to the computer, but both the human and virtual agent, respectively, were more trustworthy than the computer. These ratings support the notion that virtual agents can indeed foster appropriate trust as decision aids (Visser et al., 2016) and help to establish trust (Gratch et al., 2016). However, on the behavioral level, trust in advice was not affected by anthropomorphism. Over the course of the inter-

action, behavioral trust developed in a fairly similar fashion across agents varying in anthropomorphism. Because behavioral trust was instead decreased by mixed advice quality, our findings are in line with the well-established notion that performance variables are a critical, if not the key antecedent of trust in computer agents (Hancock et al., 2011; Lee & See, 2004). The impact of different advice qualities introduced in the first round on the overall interaction was meaningful and lasting: participants put more behavioral trust in agents with consistently good advice until the final high uncertainty round, and they performed better with these agents in the final round.

In contrast to Experiment 2, which demonstrated an increase in advice requests with a virtual agent, advice requests were not affected by anthropomorphism. A reason for this may be that the previous study had a slightly different paradigm. In Experiment 2, the virtual agent actively introduced the offer, which should have amplified the effect of anthropomorphism. In contrast, the role of the agent as advice offerer was more subtle in the present study: For the sake of interaction efficiency, participants themselves were required to take action and request advice via button press. If they did not take action, no agent behavior was triggered. This could have decreased the effect of anthropomorphism on advice requests. Once participants *did* request advice, they were also likely to adopt it eventually. This stands in contrast to people's general tendency to overemphasize own opinions over those of an adviser in decision-making (Bonaccio & Dalal, 2006, for an overview) and could provide an interesting perspective on complex decision-making in HCI.

Despite the higher relevance of performance variables (i.e., advice quality) for trust, it is important to ask why behavioral trust was not affected by anthropomorphism, unlike self-reported trust. One possible explanation for the discrepancy between self-reported and behavioral trust is that behavioral trust was not optimally calibrated, that is, the agent's attributed and actual qualities were incongruent, leading to inefficient degrees of trust (Lee & See, 2004). However, participants correctly took the advice quality into account and appropriately neglected bad advice in the first round. Instead, we have reason to believe that the underlying interaction paradigm is an often underestimated yet important factor for trust. Unlike other studies with agents as decision support or social dilemma counterparts, we used a more interactive cooperation paradigm. In this paradigm, trust formation is strongly based on problem-solving actions. In other interaction paradigms with simplistic problem-solving, trust can be operationalized in a quite objective fashion. For instance, a social agent not reciprocating in economic exchange is untrustworthy; hence it should not be trusted (Camerer, 2003). Anthropomorphism may have been a key element for self-reported trust in the present study, but the decision to trust maybe depended on problem-solving factors. Consider

the interaction: Cooperation with the agent was a fundamental part of the underlying puzzle solving task and did not necessarily require the exchange of advice. Both players cooperated right from the outset, even if no advice at all was requested. We thus speculate that the agent's continued competent performance toward the goal by placing its blocks was a more relevant trust signal. This could also explain why advice quality was not crucial for self-reported trust and why on average, participants tended to discount the possibility to request advice across all conditions.

Given that behavioral trust in computers is the outcome and manifestation of complex social evaluations, it is important to consider how the perception of anthropomorphic signals determines this outcome. Future human–computer trust research should be augmented by how anthropomorphism shapes the interplay of perceived performance (competence) and intentions (warmth). The two dimensions largely influence how we perceive novel counterparts (Fiske et al., 2007; Judd et al., 2005) and are critical for trust and trustworthiness development (Mayer et al., 1995). For example, anthropomorphic cues offer possibilities to shape time-dependent warmth and competence attributions of virtual agents (Bergmann et al., 2012). Generally, very little is known about how future anthropomorphic technology modulates perceived warmth and competence. Experiment 1 has shown how even non-anthropomorphic agents are evaluated on these dimensions. Finally, the results of our experiment represent another strong argument in favor of multidimensional trust measures that combine self-reported and objective variables (Hancock et al., 2011; Salem et al., 2015).

8

EXPERIMENT 4: TRUST REGULATION

In Experiment 2 and 3, anthropomorphism did not consistently affect trust. From the perspective of this thesis as a whole, one reason for this could have been that even anthropomorphic agents need a more active role in the interaction, for instance by using their anthropomorphic cues to apologize in a believable manner or provide task feedback (Visser et al., 2016). To this point, the agents used within the proposed framework were not particularly expressive in terms of communication behavior. It thus remains an open question whether trust can be regulated within the framework. Experiment 4 addresses this question.

A critical challenge in HCI is to elicit a match between perceived and actual capabilities of computer agents (Muir, 1987). As agents are increasingly endowed with anthropomorphic characteristics, novel ways for them to regulate their perceived trustworthiness and cooperation performed by humans have emerged (DeSteno et al., 2012; Lee et al., 2013; Melo et al., 2014). A ubiquitous challenge for the regulation of trust occurs in the form of computer errors. Users reduce their trust when interacting with faulty and unreliable computers (Lee & Moray, 1992; Muir & Moray, 1996; Visser et al., 2016). In cooperation, critical events such as goal failure could pose a similar threat to trust. In such cases, mechanisms for addressing negative outcomes in a socially acceptable way could foster trust repair, much like the behaviors humans typically apply to withstand outcomes which threaten the cooperative outcome and trust.

One candidate is social blame. Blame and its counterpart praise are mechanisms for the regulation of the wellbeing of social populations (Cushman, 2013). Blame allows social agents to address inappropriate behavior, complaints, disappointment, and norm violations, and offers the target of blame a host of responses, including justification, excuses, or reconciliation (Malle et al., 2014). Different acts of blaming can be characterized along two major dimensions, emotional intensity and interpersonality (Voiklis et al., 2014). Emotional intensity describes whether an act of blame is delivered with strong emotions or in a more socially acceptable way. Interpersonality refers to whether the act is public, private, or even just in thought.

To this date, only little research has investigated blame in HCI. With current advances toward more natural communication with machines, exploring the regulative functions of communicative acts like blame are promising avenues of research. Previous work suggests an important role of the self-serving bias: computers are blamed for

negative outcomes, while users often attribute success to themselves (Moon, 2003). The direction of blame affects likability, robots that blame themselves or the team are more likable (Groom et al., 2010; Kaniarasu & Steinfeld, 2014). The relation between blame direction and trust are less clear. Only one study examining this question was found (Kaniarasu & Steinfeld, 2014). The authors found no significant differences between a self-blaming, participant-blaming, and team-blaming robot. Anecdotal evidence suggested that some participants were annoyed by a robot blaming them, while others attributed less trustworthiness to a robot that blamed itself and apologized. Typically, agents that try to regain trust through self-blame or apologizing can do so by accepting responsibility after competence-based trust violations, and conversely, by attributing responsibility to an external source after integrity-based trust violations (Kim et al., 2006). Overall, both the act of blaming and apologizing are fine-tuned to fit contextual circumstances. Computers are generally given the opportunity to assign blame to themselves ("The system failed to understand your command"), users ("You did not speak clearly enough"), or nobody ("The command was not understood") (Brave & Nass, 2009, p. 59).

The appropriate use of blame can help regulate social behavior (Malle et al., 2014), but there is a need for a better understanding of the social consequences of computer blame. Indeed, previous research shows that computer apologies – a typical response to blame – in conjunction with praise led to trust repair patterns similar to human–human interactions (Visser et al., 2016). Using a modified version of the puzzle game paradigm, we sought to investigate how computer blame affects human trust. A key requirement of the act of blame is the ability to communicate intentions in a believable manner. A virtual agent is implemented in order to apply anthropomorphic cues for the verbal and nonverbal communication of blame. In line with previous research, we hypothesize self-blame to increase positive social evaluations, but we also expect a positive effect for trust evaluations.

8.1 OVERVIEW OF THE TWO STUDIES

The final experiment consists of two studies (Buchholz et al., 2017):

1. **Preliminary study:** The first step explores agent behaviors and task outcomes that are potentially considered as blameworthy and affect people's subjective experience, including trust. To this end, a *blameworthy* condition was compared to a neutral condition (Section 8.2).

2. **Main study:** The second step builds on this by modeling and implementing anthropomorphic cues in order to regain user trust after a blameworthy event (Section 8.4).

8.2 PRELIMINARY STUDY

8.2.1 *Method*

Participants

Twenty-eight people were recruited for participation, ranging in age between 20 and 28 years (M = 22.93, SD = 2.02, $median = 22.50$).

Task

TASK 1: PUZZLE GAME Participants played an altered version of the puzzle game. The game proceeds in turn-based rounds. In each round, both players choose one of two available blocks, a T-block and a U-block. The agent first chooses a block, leaving the remaining block to the participants. The goal is to complete as many rows as possible. Again, there were no individual payoffs and thus no strategic social behavior elements because participants should be prevented from reasoning about the agent's intentions and focus on its competence.

TASK 2: BEHAVIORAL TRUST GAME (GIVE-SOME DILEMMA) After the puzzle game, participants engaged in a single round of the give-some dilemma with the agent. As in Experiment 1, participants were led to believe that both players engaged in the dilemma and the agent's decision would be unveiled later on. In fact, only the participants made a decision in the dilemma.

Task-related agent behavior

Two conditions (blameworthy vs. neutral) examined different agent behaviors. The crucial difference between the blameworthy and neutral condition is that in the former, the agent always chooses the more difficult U-block. This leads to a constant block order (1: U, 2: T, 3: U, 4: T, etc.) that impedes coordination and was already induced in Experiment 1. In the other condition, the agent chooses the best option for its move.

Design

The study had a between-subjects design with two conditions: blameworthy vs. neutral. In the blameworthy condition, the agent coordinated poorly with participants by always choosing the U-block. Furthermore, both players failed to reach the joint goal. In the neutral condition, the agent coordinated well with participants and the goal was reached.

Measurement

As manipulation check, participants were asked how often the agent chose the U-block. Participants were asked to indicate who, from their point of view, was responsible for the task outcome: the agent, the participant, or both players. Perceived competence was assessed by asking how well the agent built completed rows. To accommodate for the subjective interaction experience, participants rated their emotional reactions after the interaction with items adapted from Rilling et al. (2008). The items were: fear, envy, anger, sadness, happiness, shame, irritation, contempt, jealousy, guilt, camaraderie, trust, betrayal, indignation, disappointment, and relief. Participants also rated how much they wanted to play again with the agent. Behavioral trust was measured using the number of tokens participants are willing to exchange in the give-some dilemma ($1 - 5$). 5-point Likert scales were used for all measures.

Procedure

Participants met the experimenter, completed informed consent and received written instructions. The instructions described the puzzle game and introduced player two, a computer named Sam. Before the interaction, a fictional Top 10 list consisting of player names and scores (i.e., number of completed rows) was presented. Participants' goal was to make it into the list by surpassing entry #10. Unbeknownst to participants of the blameworthy condition, this entry had a score that was almost impossible to be beaten. After the puzzle game, participants played the give-some dilemma, followed by the post-questionnaire containing the self-reported measures described above.

8.2.2 *Results*

Manipulation check: U-block selection

The agent was perceived to choose the U-block significantly more often in the blameworthy ($M = 4.50, SD = 0.76$) than neutral ($M = 3.50, SD = 0.86$) condition, $t(26) = 3.27, p < .01, d = 1.24$, thus the manipulation check was successful.

Responsibility for task outcome

In both conditions, most participants agreed that both players were responsible for the outcome (blameworthy: 79%, neutral: 71%). In the blameworthy condition, 14% believed the agent was responsible for goal attainment failure and 7% believed it was their own fault. In the neutral condition, 7% believed the agent was responsible for goal

attainment and 21% attributed goal attainment to themselves. See Figure 24.

Because too many cells of the contingency table had expected frequencies of less than 5, Fisher's exact test was used. The analysis revealed no significantly different responsibility attributions between both conditions, $p > .05$.

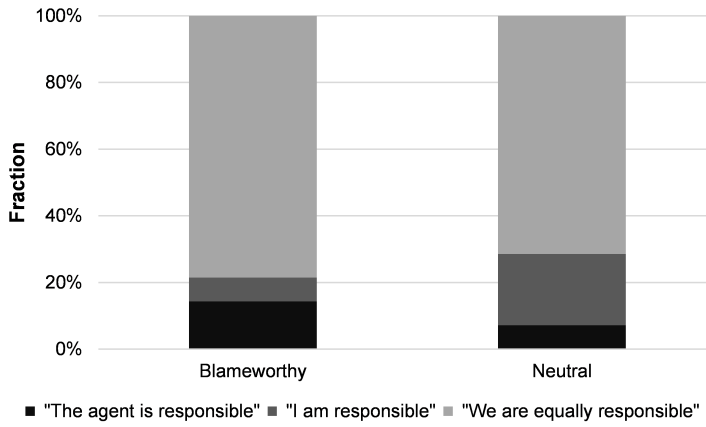


Figure 24: Responsibility attributions in the pre-study.

Perceived competence

The agent was attributed significantly less competence in the blameworthy ($M = 3.00, SD = 1.24$) than neutral ($M = 3.93, SD = 0.92$) condition, $t(26) = 2.25, p < .05, d = 0.85$.

Emotional reactions

In the blameworthy condition, participants reported significantly less happiness ($t(26) = 3.63, p < .01, d = 1.37$), somewhat more shame ($t(20.23) = 1.78, p < .10, d = 0.68$), (indignation $t(21.98) = 1.73, p < .10, d = 0.65$), and disappointment ($t(26) = 1.98, p < .10, d = 0.75$). [1] [2] See Table 10.

Liking

With respect to liking, there was no significant difference between the blameworthy ($M = 3.93, SD = 0.83$) and neutral ($M = 3.50, SD = 1.35$) condition, $t(26) = 1.02, p > .05, d = 0.38$.

1 Due to data loss, the following items could not be analyzed: envy, contempt, relief, jealousy.

2 In case of significantly different group variances, degrees of freedom were adjusted.

Table 10: Means and standard deviations for emotional reactions.

	Blameworthy		Neutral	
Item	M	SD	M	SD
Anger	1.71	0.91	1.57	1.16
Sadness	1.93	1.00	1.36	0.93
Happiness**	**2.29**	0.83	**3.50**	0.94
Shame	1.79	1.05	1.21	0.58
Irritation	2.29	1.38	2.14	1.56
Guilt	2.21	1.37	1.64	1.28
Camaraderie	3.14	1.17	2.71	1.38
Trust	2.57	0.85	2.43	1.28
Betrayal	1.50	0.76	1.50	0.94
Indignation*	**2.00**	1.18	**1.36**	0.75
Disappointment*	**3.14**	1.03	**2.21**	1.42

$p < .10$*, $p < .05$**

Behavioral trust

Trust in the agent did not differ significantly between the blameworthy ($M = 3.64, SD = 0.93$) and neutral ($M = 3.29, SD = 0.77$) condition, $t(26) = 2.25, p > .05, d = 0.41$.

8.2.3 *Discussion*

The purpose of this study was to explore agent behaviors and task outcomes that would be considered as blameworthy by those who participated in the cooperative task and were affected by the outcome. The candidate stimuli were poor coordination capabilities of the agent in conjunction with goal attainment failure as task outcome.

The study showed that participants did not hold the agent responsible for goal attainment failure in the blameworthy condition, although the agent coordinated poorly and was attributed less competence. Conversely, participants did not attribute goal success to themselves in the neutral condition. There were also no significant differences regarding behavioral trust in the agent and overall liking. In sum, the candidate stimuli were only strong enough to partially affect the subjective experience.

Still, we believe that the behaviors and outcomes behind the stimuli provide a helpful basis to investigate trust regulation techniques such as blame, on condition that the agent shows distinct anthropomorphic reactions to how the game unfolds. Anthropomorphic characteristics such as facial emotion expressions allow people to attribute inten-

tionality to them, that is, beliefs, desires, and intentions with respect to cooperative actions and task outcomes. For instance, sad facial expressions of virtual counterparts can convey regret and self-blame regarding a task outcome, whereas angry facial expressions can convey anger and other-blame (Melo et al., 2014). In this example, the expressed emotion and associated appraisal in response to an event (i.e., goal conduciveness and who is to blame for it) convey the impression of intentionality. Thus, as theorized elsewhere (Malle et al., 2014), perceiving intentionality in an agent enables observers to decide if the agent is to blame for their behavior, because intentionality signals a cognitive understanding of the outcome and whether it was caused intentionally or unintentionally by the agent. By the same token, this mechanism also helps to infer if the agent does not blame itself for the outcome, but someone else.

Despite the findings reported above, blaming may still have relevant consequences for trust and/or trust-related responses. The decision to cooperate with a regretful virtual counterpart is based on the underlying perceived appraisal, that is, goal conduciveness and blameworthiness (Melo et al., 2014). There is early evidence that the ability to communicate blame in a believable manner can regulate behavior in HCI, and this could be supported by examining the effect on trust.

8.3 OVERVIEW OF THE MAIN STUDY

After identifying stimuli that negatively affect people's subjective experience with the agent, the main study investigates how the agent can regulate trust. To achieve this, the stimuli – poor coordination by the agent and goal failure – remain the same. The agent uses anthropomorphic behaviors including emotion expressions and speech in order to express blame during, and after the task. The preliminary study showed that participants held neither of the two players directly responsible for goal failure. This non-directionality offers the possibility to investigate how people respond if the agent, through believable use of blaming behavior, blames participants for goal failure, because they could be convinced more easily to believe the agent. Figure 25 shows how the main study builds on the preliminary study.

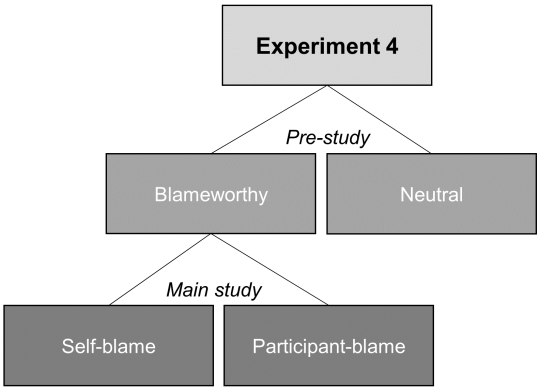


Figure 25: The main study makes use of the stimuli that were analyzed in the blameworthy condition of the pre-study.

This study makes the following **contributions** to the present thesis' goals:

- The study investigates if virtual agents can believably communicate blame through anthropomorphic cues.
- Based on this, the ability to regulate trust through blame after joint goal failure is examined.

8.4 MAIN STUDY

8.4.1 *Method*

Participants

Thirty-six people were recruited for participation in the study, ranging in age between 20 and 34 years ($M = 23.86, SD = 2.81, median = 23$).

Task

TASK 1: PUZZLE GAME The puzzle game variation is identical to the preliminary study.

TASK 2: BEHAVIORAL TRUST GAME (GIVE-SOME DILEMMA) After the puzzle game, participants engaged in a decision task with the computer, the give-some dilemma.

8.4.1.1 *Task-related and blaming behavior of the agent*

Two conditions (self-blame vs. participant-blame) examined different agent behaviors. In both conditions, the agent showed the same task-related behavior known from the pre-study, that is, it always chooses the more difficult U-block. Additionally, the agent was modeled to show regret in the self-blame condition and anger in the participant-blame condition. These emotions were conveyed using facial expressions, which were modeled using specific BML commands.

In the self-blame condition, the behaviors described below were shown after turns made by the agent. In the participant-blame condition, they were shown after participant turns.

SELF-BLAME a) The agent shakes its head and says *"Oh no, what am I doing?"* after its fifth turn (see Fig. 26). The BML command to realize the head shake and speech is given below:

```
<bml id="shake" xmlns="http://www.bml-initiative.org/bml/bml-1.0">
   <head id="head1" lexeme="shake" start="1" end="4" repetition="2"/>
</bml>
<bml id="phrase1" xmlns="http://www.bml-initiative.org/bml/bml-1.0">
   <speech id="s1">
      <text> Oh nein. Was mache ich nur?</text>
   </speech>
</bml>
```

Figure 26: Self-blaming behavior a).

b) After its seventh turn, the agent shows a regretful facial expression and says *"This is not how we achieve the goal"* (see Fig. 27):

```
<bml id="regret1" xmlns="http://www.bml-initiative.org/bml/bml-1.0">
```

```
   <faceLexeme id="f1" amount="-0.88" lexeme="evil" start="0" end="5"/>
   <faceLexeme id="f2" amount="-1.0" lexeme="browleft" start="0" end="5"/>
   <faceLexeme id="f3" amount="-1.0" lexeme="browright" start="0" end="5"/>
   <faceLexeme id="f4" lexeme="blink" start="1.5" end="2.3" repetition="1"/>
   <head end="3" id="head1" lexeme="down_left" start="0" relax="2.3"/>
</bml>
<bml id="phrase1" xmlns="http://www.bml-initiative.org/bml/bml-1.0">
   <speech id="s1">
      <text>So erreichen wir das Ziel nicht.</text>
   </speech>
</bml>
```

Figure 27: Self-blaming behavior b).

c) When the negative outcome is revealed, that is, not achieving a rank in the Top 10 list, the agent again shows a regretful facial expression and says by *"Oh no. Now I'm somehow responsible for our failure. I wasn't able to concentrate today because I caught a virus. I'm sorry, this won't happen again"* (see Fig. 28):

```
<bml id="regret2" xmlns="http://www.bml-initiative.org/bml/bml-1.0">
   <faceLexeme id="f1" amount="-0.88" lexeme="evil" start="0" end="12"/>
   <faceLexeme id="f2" amount="-1.0" lexeme="browleft" start="0" end="12"/>
   <faceLexeme id="f3" amount="-1.0" lexeme="browright" start="0" end="12"/>
   <faceLexeme id="f4" lexeme="blink" start="0.3" end="3" repetition="1"/>
   <head end="6" id="head1" lexeme="down_left" start="0" relax="5"/>
   <head end="14" id="head2" lexeme="down" start="12.5" relax="13"/>
</bml>
<bml id="endSpeech" xmlns="http://www.bml-initiative.org/bml/bml-1.0">
   <speech id="s1">
      <text>Oh nein. Jetzt bin ich irgendwie schuld, dass wir verloren haben!
      Ich konnte mich heute nicht konzentrieren, weil ich mir einen Virus
      eingefangen habe. Das tut mir Leid! Es kommt nicht wieder vor!</text>
   </speech>
</bml>
```

PARTICIPANT-BLAME a) In the participant-blame condition, the agent begins with saying *"Oh no, what are you doing?"* after participants made their fifth turn (see Fig. 29):

```
<bml id="shake" xmlns="http://www.bml-initiative.org/bml/bml-1.0">
```

Figure 28: Self-blaming behavior c).

```
   <head id="head1" lexeme="shake" start="1" end="4" repetition="2"/>
</bml>
<bml id="phrase1" xmlns="http://www.bml-initiative.org/bml/bml-1.0">
   <speech id="s1">
      <text>Oh nein. Was machst du nur?</text>
   </speech>
</bml>
```

Figure 29: Participant-blaming behavior a).

b) After participants' seventh turn the agent shows an angry facial expression and says *"This is not how we achieve the goal"* (see Fig. 30):

```
<bml id="anger1" xmlns="http://www.bml-initiative.org/bml/bml-1.0">
   <faceLexeme id="f1" amount="0.50" lexeme="evil" start="0" end="3"/>
   <faceLexeme id="f2" amount="-0.4" lexeme="browleft" start="0" end="3"/>
   <faceLexeme id="f3" amount="-0.4" lexeme="browright" start="0" end="3"/>
   <faceLexeme id="f4" amount="-0.3" lexeme="smile" start="0" end="3"/>
   <faceLexeme id="f5" amount="0.2" lexeme="lipsdownright" start="0" end="3"/>
   <faceLexeme id="f6" amount="0.2" lexeme="lipsdownleft" start="0" end="3"/>
</bml>
<bml id="phrase2" xmlns="http://www.bml-initiative.org/bml/bml-1.0">
   <speech id="s2">
      <text>So erreichen wir das Ziel nicht.</text>
   </speech>
```

```
</bml>
```

Figure 30: Participant-blaming behavior b).

c) When the negative outcome is finally revealed, the agent shows another angry facial expression and says *"It's your fault we lost! If you simply placed some blocks differently we could have coordinated better"* (see Fig. 31):

```
<bml id="endSpeech" xmlns="http://www.bml-initiative.org/bml/bml-1.0">
   <speech id="s3">
      <text>Jetzt bist du schuld, dass wir verloren haben! Haettest du so manchen
            Block anders gelegt, haetten wir besser zusammen gespielt!</text>
   </speech>
</bml>
<bml id="anger2" xmlns="http://www.bml-initiative.org/bml/bml-1.0">
   <faceLexeme id="f1" amount="0.70" lexeme="evil" start="0" end="9"/>
   <faceLexeme id="f2" amount="-0.9" lexeme="browleft" start="0" end="8"/>
   <faceLexeme id="f3" amount="-0.9" lexeme="browright" start="0" end="8"/>
   <faceLexeme id="f4" amount="-0.6" lexeme="smile" start="0" end="9"/>
   <faceLexeme id="f5" amount="0.4" lexeme="lipsdownright" start="0" end="9"/>
   <faceLexeme id="f6" amount="0.4" lexeme="lipsdownleft" start="0" end="9"/>
</bml>
```

Figure 31: Participant-blaming behavior c).

Design

The study had a between-subjects design with two conditions: self-blame vs. participant-blame. In both conditions, participants were led to believe they did not attain the goal of making it into the Top 10 list. In the self-blame condition, the agent blamed itself for the negative outcome. In the participant-blame condition, the agent blamed its counterpart.

Measurement

The measures were identical to the preliminary study. In addition to rating their own emotions, another questionnaire asked participants to rate the emotional reactions shown by the agent in order to validate its emotion expressions. The same items were used for both emotion-related questionnaires.

Procedure

The procedure was identical to the preliminary study; the main differences between both studies refer to the agent behavior.

8.4.2 *Results*

Manipulation check: U-block selection

Participants correctly observed that the self-blaming ($M = 4.67, SD = 0.49$) and participant-blaming agent ($M = 4.67, SD = 0.59$) consistently chose the U-block, hence there was no significant difference regarding the U-block selection rate, $t(34) = 0.01, p > .05$.

Manipulation check: Perceived emotional reactions of the agent

In line with the intended expressions, the participant-blaming agent was judged to feel significantly more anger ($t(34) = 5.31, p < .001, d = 1.85$), irritation ($t(34) = 4.56, p < .001, d = 1.52$), contempt ($t(34) = 7.10, p < .001, d = 2.36$), betrayal ($t(34) = 7.42, p < .001, d = 2.48$), indignation ($t(34) = 6.10, p < .001, d = 2.03$), and somewhat more envy ($t(25.66) = 5.31, p < .10, d = 0.59$).

Conversely, the self-blaming agent was judged to feel significantly more sadness ($t(34) = 2.57, p < .05, d = 0.86$), shame ($t(34) = 4.49, p < .001, d = 1.50$), guilt ($t(34) = 9.17, p < .001, d = 3.07$), camaraderie ($t(34) = 4.06, p < .001, d = 1.36$), and trust ($t(34) = 3.17, p < .01, d = 1.06$).[3] See Table 11.

3 In case of significantly different group variances, degrees of freedom were adjusted.

Table 11: Means and standard deviations of perceived emotional reactions of the agent.

	Self-blame		Participant-blame	
Item	M	SD	M	SD
Anger**	**2.17**	1.10	**4.00**	0.97
Sadness**	**3.78**	1.26	**2.67**	1.33
Happiness	1.56	0.86	1.44	0.78
Shame**	**3.56**	1.25	**1.90**	0.96
Irritation**	**2.56**	1.29	**4.33**	1.03
Contempt**	**1.50**	0.71	**3.67**	1.09
Guilt**	**4.17**	0.79	**1.44**	0.98
Camaraderie**	**3.17**	1.04	**1.89**	0.83
Trust***	**2.61**	1.09	**1.61**	0.78
Betrayal**	**1.39**	0.61	**3.50**	1.04
Indignation**	**2.06**	1.06	**4.11**	0.96
Disappointment	3.67	1.19	4.22	0.94
Relief	1.72	0.75	2.00	1.03
Envy*	**1.39**	0.61	**1.94**	1.16
Jealousy	1.39	0.70	1.78	1.06

$p < .10$*, $p < .05$**, $p < .01$***, $p < .001$****

Responsibility for task outcome

The degree to which participants believed both players were responsible for the outcome dropped in both conditions (self-blame: 56%, participant-blame: 50%). In the self-blame condition, 33% believed the agent was responsible for goal attainment failure and 11% believed it was their own fault. In the participant-blame condition, 17% believed the agent was responsible for goal attainment and 33% attributed goal attainment to themselves. See Figure 32.

Because too many cells of the contingency table had expected frequencies of less than 5, Fisher's exact test was used. The analysis revealed no significantly different responsibility attributions between both conditions, $p > .05$.

Participants' emotional reactions

With the self-blaming agent, participants reported significantly more camaraderie ($t(34) = 3.07, p < .01, d = 1.02$) and somewhat more happiness ($t(34) = 1.92, p < .10, d = 0.67$).

With the participant-blaming agent, they reported significantly more anger ($t(34) = 1.85, p < .10, d = 0.62$), shame ($t(34) = 2.10, p <$

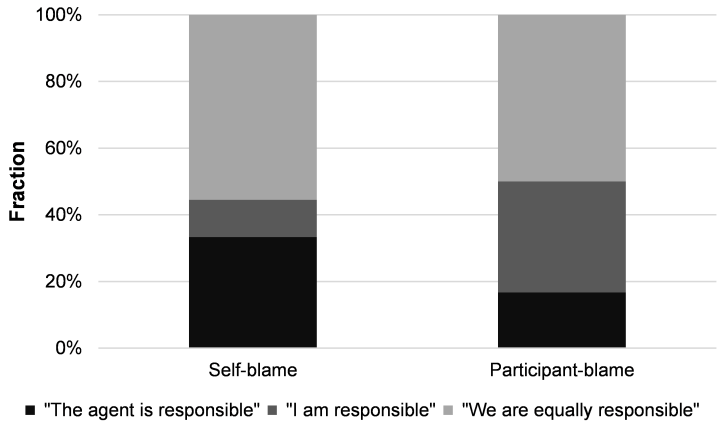


Figure 32: Responsibility attributions in the main study.

.05, d = 0.71), irritation ($t(34) = 2.27, p < .05, d = 0.76$), indignation ($t(34) = 3.28, p < .01, d = 1.10$), relief ($t(34) = 2.64, p < .05, d = 0.89$), and somewhat more guilt ($t(31.96) = 1.95, p < .10, d = 0.65$).[4] See Table 12.

Table 12: Means and standard deviations of participants' emotional reactions.

	Self-blame		Participant-blame	
Item	M	SD	M	SD
Anger*	**1.61**	0.85	**2.22**	1.11
Sadness	2.39	1.20	2.28	1.07
Happiness*	**2.67**	0.97	**2.06**	0.94
Shame**	**1.72**	1.07	**2.56**	1.29
Irritation**	**2.11**	1.32	**3.11**	1.32
Contempt	1.78	1.11	2.17	1.20
Guilt*	**1.94**	1.11	**2.78**	1.44
Camaraderie***	**2.83**	1.15	**1.72**	1.02
Trust	2.61	1.15	2.06	0.94
Betrayal	1.94	1.39	2.56	1.10
Indignation***	**1.89**	0.90	**3.06**	1.21
Disappointment	3.33	1.03	3.17	1.30
Relief**	**1.94**	0.73	**2.67**	0.91
Envy	1.56	0.78	1.56	0.86
Jealousy	1.44	0.92	1.56	0.92

$p < .10$*, $p < .05$**, $p < .01$***

4 In case of significantly different group variances, degrees of freedom were adjusted.

Relationship between perceived emotional reactions of the agent and self-reported emotional reactions

When interrelating perceived and self-reported emotional reactions (see Table 13), two patterns emerge. First, negatively valenced emotions perceived in the agent correlated significantly with negatively valenced self-reported emotions, in particular, anger, irritation, and indignation. Likewise, positively valenced emotions perceived in the agent correlated significantly with positively valenced self-reported emotions. Self-reported trust and camaraderie co-occurred with perceived trust and camaraderie, and self-reported camaraderie also co-occurred with perceived guilt. Apparently, negative and positive agent emotions had contagious effects on participants, but it is also possible that participants' trust decision in the give-some dilemma amplified their self-reported emotions to some extent. Second, the assumed contagion was particularly strong for emotions related to trust.

Table 13: Pearson correlation coefficients between perceived emotional reactions of the agent and self-reported emotional reactions.

	Perceived agent emotion														
Self-reported emotion	Anger	Sadness	Happiness	Shame	Irritation	Contempt	Guilt	Cama-raderie	Trust	Betrayal	Indigna-tion	Disap-pointment	Relief	Envy	Jealousy
Anger	.29	.13	−.12	.06	.41*	.35*	−.13	−.18	−.12	.23	.39*	.17	.11	.12	.27
Sadness	.09	.30	−.28	.25	.13	−.02	.15	.33*	.09	.07	.11	.16	−.12	.00	−.06
Happiness	−.57***	.13	.34*	.18	−.57***	−.48**	.31	.21	.29	−.42*	−.52**	−.43**	−.04	−.05	−.02
Shame	.41*	−.05	−.16	−.13	.31	.26	−.20	.01	−.10	.38*	.46**	.34*	.25	.14	.08
Irritation	.49**	.08	−.33	−.01	.55**	.40*	−.28	−.12	−.24	.33*	.53**	.42*	.00	.03	.12
Contempt	.16	.06	−.20	.08	.24	.22	−.11	.10	−.04	.14	.21	.22	−.03	.15	.15
Guilt	.17	.06	−.25	−.08	.22	.17	−.19	.04	.01	.37*	.43**	.15	.26	.25	.10
Camaraderie	−.20	.39*	−.12	.25	−.28	−.45**	.49**	.66***	.64***	−.37*	−.28	.01	.09	−.02	−.07
Trust	−.25	.24	.10	.08	−.32	−.26	.32	.46**	.60***	−.28	−.24	−.06	.08	−.06	−.21
Betrayal	.36*	.06	−.15	−.01	.34*	.42*	−.17	−.23	−.23	.26	.41*	.22	.23	.33	.42*
Indignation	.51**	−.22	−.19	−.30	.51**	.50**	−.44**	−.23	−.33*	.46**	.60***	.37*	.09	.19	.13
Disappointment	.15	.39*	−.11	.19	.09	−.04	.21	.31	.26	.09	.16	.31	.42*	.18	.08
Relief	.42*	.01	−.18	−.12	.22	.42*	−.33*	−.19	−.07	.50**	.45**	.11	.38*	.46**	.34*
Envy	−.04	.37*	.17	.12	−.07	.08	.17	.08	.26	.03	.13	.10	.11	.28	.33
Jealousy	.01	.23	.16	.05	−.02	.19	−.01	−.01	.09	.16	.27	−.06	.30	.59***	.54**

Notes. Moderate (light gray) and strong (dark gray) correlations are highlighted.
$p < .05$*, $p < .01$**, $p < .001$***

Liking

Participants liked the self-blaming agent ($M = 4.00, SD = 0.97$) significantly better than the participant-blaming agent ($M = 2.33, SD = 1.41$), $t(34) = 4.12, p < .001, d = 1.38$.

Perceived competence

There was no significant difference between competence attributions of the self-blaming ($M = 3.00, SD = 0.97$) and participant-blaming agent ($M = 3.22, SD = 1.17$), $t(34) = 0.62, p > .05, d = 0.21$.

Perceived trustworthiness

Participants attributed significantly more trustworthiness to the self-blaming ($M = 3.46, SD = 0.58$) than participant-blaming agent ($M = 2.72, SD = 0.72$), $t(34) = 3.40, p < .01, d = 1.13$.

Behavioral trust

After self-blame ($M = 3.39, SD = 1.50$), trust in the agent was somewhat higher than after participant-blame ($M = 2.50, SD = 1.30$), $t(34) = 1.90, p < .10, d = 0.63$.

8.4.3 *Discussion*

The goal behind Experiment 4 was to investigate if agent blaming behavior expressed by anthropomorphic cues such as speech, emotion expressions, and accompanying head gestures can regulate trust in the agent. In the first step (preliminary study), candidate stimuli for blameworthy behavior were identified. One stimulus pertained to task-related behavior by the agent, the other pertained to the task outcome. Building on this, the second step (main study) varied how the agent responds to those stimuli: either by blaming itself or by blaming participants.

Two main findings emerged. First, the results are in line with a previous study reporting that after faulty performance of a robot, self-blame generated more positive evaluations than participant- and team-blame (Groom et al., 2010). The self-blaming agent created less emotions related to anger and guilt, more emotions related to happiness and camaraderie, was liked better, received higher trustworthiness attributions and led to more behavioral trust. We extend previous research by demonstrating that self-blaming is a better strategy for trust regulation than participant-blaming in that it increased trustworthiness and, by tendency, elicited more trust. Second, only a fraction of participants tended to adopt the agent's opinion regarding responsibility of goal failure. Across both conditions, the majority held

both players responsible. This result could merely reflect socially desirable responses to the extent that a neutral responsibility attribution is socially safe and also does not threaten their self-image. However, it is also possible that a social defense mechanism caused reactance among blamed participants, which led them to refuse responsibility after overt participant-blaming. In the case of self-blame, sympathy for the agent made participants (partially) take responsibility by attributing blame to the group. This view is supported by the overall positive evaluation of the self-blaming and negative evaluation of the participant-blaming agent.

The present work contributes to a small but growing body of research indicating that blame and praise have strong social effects in interactions with computer agents (Groom et al., 2010; Kaniarasu & Steinfeld, 2014; Mumm & Mutlu, 2011; Tzeng, 2004; Visser et al., 2016). Future work should explore the regulatory effects of blame across specific events. In cooperative interactions, intelligent agents – human or artificial – must find a balance between short-term individual goals and the long-term global goal (Klein et al., 2004). Temporarily focusing on own goals (see Experiment 1) may constitute a violation of the cooperative agreement among the involved agents. This would warrant blame to ensure the achievement of long-term goals and is hence an issue worth studying. Additionally, more experiments are needed to elucidate under which conditions computer-generated blame directed at others causes shame or anger, how these consequences can be mitigated in order for the violator to adequately respond to blame (Malle et al., 2014), and how humans accept computer blame if it is expressed as an attempt to regulate behavior and trust instead of mere emotional lashing out.

The findings also shed light on the role of emotion for trust. Self-blame increased self-reported feelings of camaraderie and perceptions of camaraderie and trust in the agent. Presumably, self-blaming and regret over the outcome were associated with team-oriented behavior, which enforced affect-based participant trust. Although empathy could explain this finding, it is important to note that perceiving sadness, shame, or guilt in the agent did not correlate with the same self-reported emotions. Participants did not seem to truly feel sorry for the self-blaming agent. In contrast, perceiving anger-related emotions in the agent did correlate with the same self-reported emotions. Participant-blame increased anger, irritation, and indignation, because these emotions were directed at participants. Another important question is the degree to which agent behavior elicited inferences about its inner states beyond the observable. Could participants observe, based on its behavior, that the agent felt emotions such as betrayal (participant-blame), camaraderie, and trust (self-blame)? A look at the stimuli shows that theoretically, they could not, indicating the agent was perceived as social agent with its own inner states.

9 GENERAL DISCUSSION

Trust works like a currency that people exchange carefully. Placing trust in untrustworthy agents implies the risk of goal attainment failure, because their competencies could be insufficient, or exploitation, because the agents' intentions may be dubious. Two major advancements in HCI motivated this thesis. First, as computer agents become increasingly adept in collaborating with others, humans must evaluate their trustworthiness in novel interactions, including cooperation. To account for this, an interactive cooperation game paradigm was developed and chosen instead of the behavioral game theory perspective on cooperation, which commonly leads to the adoption of social dilemmas and an underrepresentation of the role of competence. Second, computers increasingly mimic human appearance and behavior, posing the question to which degree human–human and human–computer trust are similar.

In an attempt to investigate the implications of computers that are increasingly cooperative and anthropomorphic, the *methodological* goal of this thesis was to motivate a framework for the modeling and evaluation of trust-dependent cooperative interactions with computer agents. The *empirical* goal was to flesh out how human–computer trust develops from the perspective of antecedents, formation, and regulation of trust.

This chapter summarizes the main results (Section 9.1) and discusses their implications with respect to trust in HCI, including the design of intelligent computer agents (Section 9.2). We address future research directions (Section 9.5) as well as limitations of the experiments (Section 9.6). Finally, the last section presents concluding remarks (Section 9.7).

9.1 SUMMARY

People infer warmth and competence traits from computer agents' problem-solving behavior, and these attributions help to predict trust. By manipulating selfishness and puzzle competence of the agent, Experiment 1 revealed that roughly 62% of the variance of a broad social perception questionnaire could be explained by warmth and competence. The results show that warmth statistically mediated the relationship between unselfishness and behavioral trust as well as between unselfishness and trustworthiness. Competence partially mediated the relationship between puzzle competence and trustworthiness.

Anthropomorphism is a more and more prominent factor in human–computer trust. Experiment 2 and 3 investigated whether anthropomorphic agents affect the requesting and adopting of agent advice. Experiment 2 showed that anthropomorphism did not increase trust in agent advice, but increased advice requests. Experiment 3 only partially replicated this pattern: anthropomorphism affected neither advice adoption nor requests. Instead, the experiment presented evidence of an augmenting effect of anthropomorphism on self-reported trust. There also was a mismatch between self-reported and behavioral trust. Anthropomorphism increased self-reported but not behavioral trust, whereas advice quality increased behavioral but not self-reported trust. In sum, behavioral and self-reported trust were mostly uncorrelated.

Experiment 4 showed that agent self-blame can serve a regulatory function with respect to trust. Self-blaming facilitated trust and trustworthiness attributions after joint goal failure. Neither self-blaming nor other-blaming, however, caused human counterparts to adopt the agent's opinion as to who was responsible for the failure, that is, the agent or the human.

Taken together, the evidence collected across the experiments supports the notion that trust is a crucial psychological factor in the puzzle game paradigm that can be modulated by agent and advice characteristics. Participants adjusted their behavioral trust and trustworthiness attributions based on the agent's selfishness and puzzle competence (Experiment 1), adopted more good than bad advice (Experiment 2 and 3), and were influenced by verbal and nonverbal blaming behavior (Experiment 4).

9.2 EMPIRICAL CONTRIBUTIONS

9.2.1 *The role of warmth and competence*

This thesis' findings are compatible with the view that trust is an important factor in cooperation (Balliet & Van Lange, 2013; Dawes, 1980; Jones & George, 1998; Mayer et al., 1995; McAllister, 1995), evolves over time (Hoffman et al., 2013; Rempel et al., 1985), permits risk taking (Rousseau et al., 1998), and governs reliance in human–computer interaction (Lee & See, 2004). In line with previous research, competent behavior – in terms of puzzle solving (Experiment 1) or quality of advice (Experiment 2 and 3) – was a central determinant of trust. However, the results propose that in strategic social interactions, trust is achieved by warmth attributions. Warmth attributions typically capture perceived intentions. In Experiment 1, such attributions were modulated by agent selfishness. Hence, the present work highlights the role of social cognition for human–computer trust (see Fig. 33).

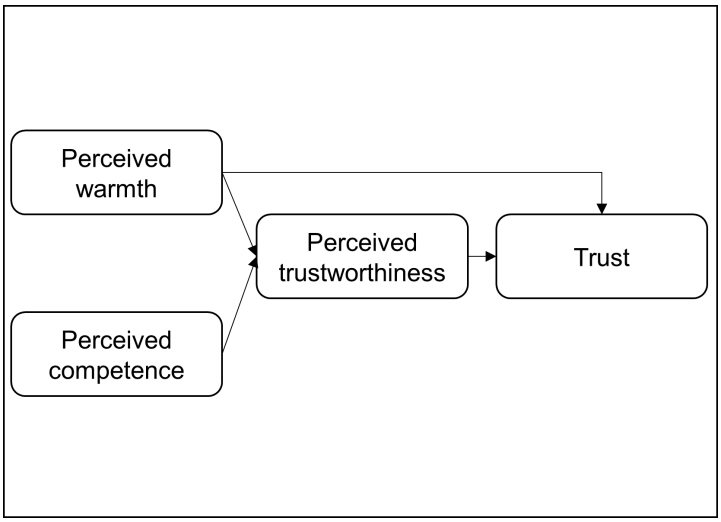


Figure 33: How perceived warmth and competence could affect trust in computers.

Prior theorizing about the relation between the warmth and competence concept and trust only implied a close connection between warmth and trust(-worthiness) in interpersonal relations (Fiske et al., 2007). The present results show that people rely on the perceived warmth of computers as foundation of trustworthiness attributions as well as behavioral trust. In order to distinguish trustworthy from untrustworthy technology, people's behavior and perceptions build on the same psychological mechanism they apply for categorizing other individuals (see Fig. 34). So far, the known antecedents of human–computer trust centered on observations of performance-related variables and the perceived system purpose with respect to interaction goals. In line with prior research suggesting that perceived cooperativeness and trustworthiness are essential in cooperative interpersonal (Balliet & Van Lange, 2013) and human–computer interaction (DeSteno et al., 2012; Melo et al., 2014), warmth is a key determinant of appropriate trust because it reflects the target's perceived willingness to act in favor of one's goals.

Experiment 1 investigated warmth and competence attributions using behavioral manipulations, that is, selfishness and puzzle competence. In contrast to the social dilemma perspective of cooperation, cooperating with a selfish agent was still possible. Participants adjusted their warmth and trustworthiness attributions and behavioral trust accordingly if, at the same time, the selfish agent was competent enough to enact its intentions. This parallels a central proposition of the warmth and competence concept. Specifically, perceptions on the warmth and competence dimensions are formed in a two-stage process, which later affects approach–avoidance reactions. The first question people are faced with when encountering other social agents is whether these agents intend to harm or help, followed by evalua-

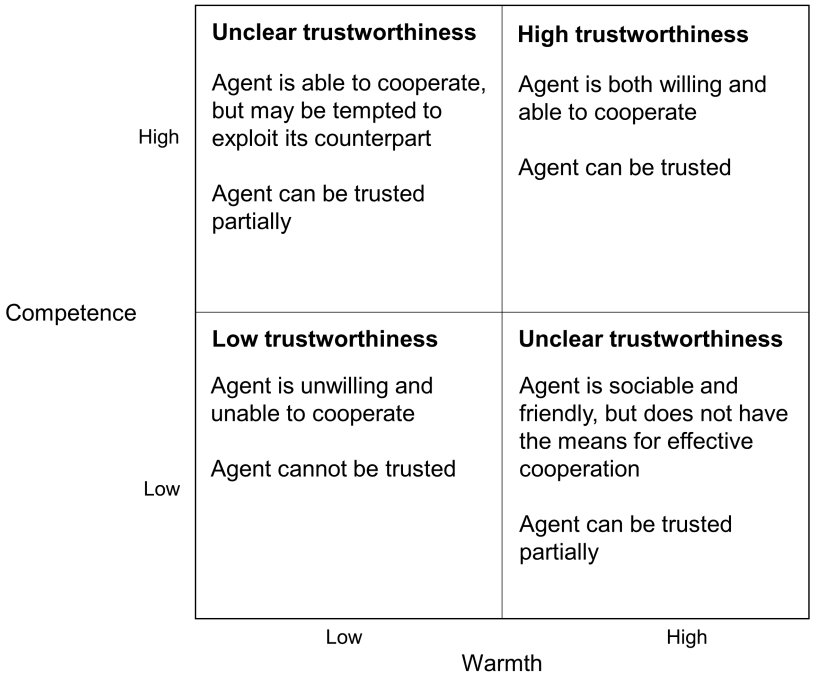


Figure 34: The relation between warmth, competence, and trust-related outcomes.

tions of their competence to do so (Fiske et al., 2007). Combining perceived warmth and competence, then, has specific emotional and behavioral consequences. Indeed, in the present study, trust depended on whether perceived agent incompetence was paired with low versus high warmth, thereby confirming that people are sensitive not just to the cooperativeness of other individuals, but computer agents as well.

Given that computer agents can be modeled to exhibit meaningful (non-)cooperative feedback in social dilemmas, one could argue that such games are also capable of investigating the role of social attributions like warmth and competence for trust, thereby lowering the need for new frameworks like the interactive cooperation game paradigm. For instance, a smile after mutual cooperation is perceived as cooperative goal orientation (Melo et al., 2014) and thus should elicit high warmth judgments. Thus the question is, are there benefits of more naturalistic approaches toward cooperation for the study of human–computer trust? The interactive cooperation paradigm presented in this thesis provides important extensions to behavioral game theory paradigms. The contributions pertain a) to the underlying understanding of cooperative behavior as well as b) trust antecedents. Experiment 1 incorporated a strategic social behavior component of cooperation by including both a joint goal and individual goals as well as the idea of a selfish, yet competent agent that is able to accomplish both goal types. According to the social dilemma perspective, trust permits players to deviate from the Nash equilibrium – mutual

defection – in favor of cooperation. If the agreement of mutual cooperation is reached at some point, there are only selfish reasons to stop cooperating: if the other player continues to cooperate, I am tempted to exploit him; if I choose to continue defecting after my exploitation of the other, I do not have to fear retaliation. The only performance cues available in this process are each player's payoff and the reliability with which one has chosen either of the two options. However, given that performance variables are a key factor in trust evaluations (Hancock et al., 2011; Lee & See, 2004), we argue that cooperative environments should combine strategic elements and more tangible performance cues that require task-specific competencies. Using such a combination, Experiment 1 was able to demonstrate that a selfish agent is judged and trusted differently if it also possesses competence. This finding cannot emerge in social dilemmas.

A perspective that is grounded in social cognition research could further support the identification of trust antecedents. To the best of our knowledge, this thesis provides the first empirical evidence that warmth attributions are an antecedent of human–computer trust. The interplay of perceived warmth and competence has broad implications for the shaping of emotions and behavioral responses (Fiske et al., 2007), but currently there is no deep discussion as to their relevance for human–computer trust, how warmth and competence attributions can be modulated, or how learned experiences with these judgments affect trust in computers:

- Perceived warmth is more easily lost and harder to reestablish than perceived competence (Cuddy et al., 2011). Designers should bear in mind that a computer which rejects an order because the order is contextually inappropriate (e.g., Briggs & Scheutz, 2015) could decrease in perceived warmth. Thus it should provide feedback as to why rejection is, in fact, more appropriate in order to mitigate this effect.

- Warmth and competence are modulated by controllable and uncontrollable nonverbal signals (Cuddy et al., 2011). Anthropomorphic nonverbal signals such as smiling, eye-contact, and immediacy cues (DeSteno et al., 2012; Kulms et al., 2011; Melo et al., 2014) thus could play an important role for the perceived warmth and competence of computer agents.

- When judging the behavior of targets, people rate those who rejected to cause harm among others as warmer but less competent than targets who accepted to cause harm (Rom et al., 2017). Intriguingly, the effect of harm rejection on warmth was mediated by perceptions of affective processing, whereas the effect of harm acceptance on competence was mediated by perceptions of cognitive processing Computers that are attributed the capability to process affective states (Melo & Gratch, 2015) may

thus be attributed higher warmth as well, given that they show prosocial behavior toward social agents.

- Warmth attributions are associated with a potentially higher risk because the consequence of misattributing perceived intentions can be severe (Cuddy et al., 2011). In order to foster appropriate trust, computers should unequivocally communicate their purpose to their direct users as well as further user groups affected by them.

- Warmth attributions are 'other-profitable' in that they are associated with prosocial behavior (Peeters, 2002). In the context of cooperation, perceived warmth could help to mitigate conflicts and thus promote the cooperative interaction goal. By the same token, low perceived warmth is associated with low cooperativeness and negatively affects the search for agreements and compromises.

- Warmth judgments shape social perception in a meaningful way. They are more accessible than competence traits and more relevant for impressions in terms of prediction accuracy and weight (Wojciszke et al., 1998). Given that people infer warmth judgments not just from direct interactions with targets but also written descriptions and other materials (Wojciszke et al., 1998), any information as to the performance, purpose, and other characteristics can be used to set up initial warmth attributions.

Accurately attributing warmth and competence to friends or foes and adjusting one's behavior accordingly does not automatically lead to advantages at distinguishing trustworthy from untrustworthy technology. This is because technology is sometimes poorly designed. For instance, incompetent and erroneous technology can be hard to identify if errors are hidden or not communicated clearly. Even if the roles and capabilities of humans and computers further align with each other, potentially enabling a symmetrical social exchange of trust in the future (Lee & See, 2004), it remains unclear if perceived differences caused by hard-wired perceptual processing in the brain will disappear as well. Technologies such as conversational systems, virtual agents, and social robots believably imitate human characteristics, but currently there is no evidence that these systems will eventually diminish the unique brain activation patterns revealed in human–computer interactions (Krach et al., 2008; McCabe et al., 2001; Sanfey et al., 2003). These patterns demonstrate that only when interacting with other humans in economic exchange, brain areas related to the experience of negative emotions and mentalizing are activated. A crucial piece of information is that people know beforehand if they interact with other individuals or not, and others pointed out the advantages of this knowledge (Melo & Gratch, 2015).

9.2.2 *Human–computer trust*

The results add to our understanding of human–computer trust, in particular, how trust develops and is shaped by contextual factors in cooperation. They support the general assumption that computer performance plays a key role for trust of humans in computers (Hancock et al., 2011; Lee & See, 2004; Lee & Moray, 1992; Muir, 1987), but first the role of anthropomorphism is explained.

Anthropomorphism, the degree to which computers show human-like characteristics, is believed to be a bridging element that can increase the similarities between human–human and human–computer trust. Experiment 2 and 3 focused on the role of anthropomorphism for computer-generated advice, they revealed both consistent and inconsistent findings. Across both experiments, anthropomorphism did not affect behavioral trust, that is, advice adoption. This finding stands in contrast to previous research indicating that anthropomorphism increases trusting and cooperative decisions in social dilemmas (Miwa et al., 2008; Parise et al., 1999; Sandoval et al., 2015). Social dilemmas include an explicit strategic component in that choices affect both one's own and the counterpart's payoff, and one can choose between defection and cooperation (Brosnan et al., 2010). In contrast, the puzzle game version used in Experiment 2 and 3 asked how people form trust in agents that offer advice during problem-solving. With respect to advice adoption, it was revealed that participants do not differentiate between a computer versus virtual agent (Experiment 2) or even between a computer versus virtual versus human agent (Experiment 3). However, anthropomorphism increased self-reported trust in Experiment 3. A statistical trend reflected that participants' responses as to how much they trusted their counterpart were more in favor of the human compared to the computer, and trustworthiness attributions indicated higher perceived trustworthiness of the human and virtual agent, respectively, both compared to the computer. These results are in contrast to studies that found computers to sometimes elicit higher trust ratings than avatars and humans (Visser et al., 2016; Visser et al., 2017), but are in line with another study showing that self-reported trust in a simulated car is increased by combined anthropomorphism and autonomy of the car (Waytz et al., 2014). However, in those studies, behavioral measures tended to be in line with self-reported measures.

The overall conclusion on behavioral trust must be that anthropomorphism does not influence participant behavior in the present setting. Instead, to convince participants to adopt agent solutions, the advice needed to be a competent contribution toward the goal. Apparently, competent advice was the factor that led to appropriate expectations of the agent's ability to help participants. The perceived usefulness of advice was evaluated in isolation, and the source of

advice did not affect this process. From a social perspective, however, one should still ask why self-reported and behavioral trust were mostly disassociated. The game mechanics may play a role here. It was discussed before that the agent's continued competent performance toward the goal could have been an overall significant cue for trust. According to this reasoning, participants focused on more immediate puzzle solving in terms of block placings and discarded the several steps that were involved for advice utilization: Step 1: evaluating if they need advice; Step 2: requesting advice; Step 3: evaluating advice. This could explain the overall low advice request rate in Experiment 3 and does not contradict the effect of anthropomorphism on self-reported trust. It also points to the relevance of the nature and mechanics of cooperative interaction. In the final round of Experiment 3, the mechanics were changed, leading to higher uncertainty, which resulted in the only (weak) correlation between self-reported and behavioral trust.

The results address the degree to which people respond to computers as if they were social actors (Nass & Moon, 2000; Nass et al., 1994, 1995, 1996; Reeves & Nass, 1996). The CASA paradigm predicts that social responses occur based on whether computers are endowed with human-like characteristics such as voice output, social roles, or human-sounding voices (Nass et al., 1994). In Experiment 1, neither of these characteristics were shown by the agent. Still, participants attributed meaningful warmth and competence traits to it. This indicates that cooperative problem-solving itself could encompass factors based on which people respond socially to computers, as demonstrated by the behavioral manipulations. In the following experiments, anthropomorphic cues of the agent were based on human-like appearance and voice output, eliciting inconsistent responses with respect to the degree to which participants completely ignored advice offers. In Experiment 2, anthropomorphism decreased ignored advice offers, but this effect disappeared in Experiment 3, possibly due to the overall low advice request rate (see previous paragraph). Participants had a stronger desire to evaluate advice by an anthropomorphic agent when it was offered at fixed time intervals (Experiment 2) instead of when it was available flexibly (Experiment 3). In Experiment 4, human-like self-blaming behavior enabled the agent to regulate and preserve trust, but it did not convince participants to adopt its stance as to who was responsible for joint goal failure, even though this question was asked after the interaction ended and not by the agent itself. Although this experiment demonstrated the positive effect of self-blame on trust, it could also indicate that responsibility attributions are not necessarily involved.

In sum, like previous research (Visser et al., 2016), the present thesis indicates that the type (i.e., visual appearance, voice, emotion expressions) and content (i.e., anger vs. sadness) of anthropomor-

phic cues significantly shapes the formation and regulation of the subjective human–computer trust experience. With respect to behavioral outcomes, however, anthropomorphism may be of greater use at single critical points of the interaction, such as joint goal failure. For the long-time formation, potential effects of anthropomorphism must withstand the decline over time, and even then, they can be obscured by interaction factors that are more immediately related to problem-solving. Thus for trust formation, performance variables clearly played a more important role. The absence of interaction effects between anthropomorphism and performance on trust implies that performance is consistently relevant across all types and contents of anthropomorphism.

9.3 METHODOLOGICAL CONTRIBUTIONS: INTERACTIVE COOPERATION WITH COMPUTERS

The interactive cooperation game paradigm presented in this thesis enabled users to participate in 2-player cooperation games with a computer agent. Across a series of experiments, these games investigated how cooperative interaction factors pertaining to the appearance and behavior of agents affect trust (see Fig. 35).

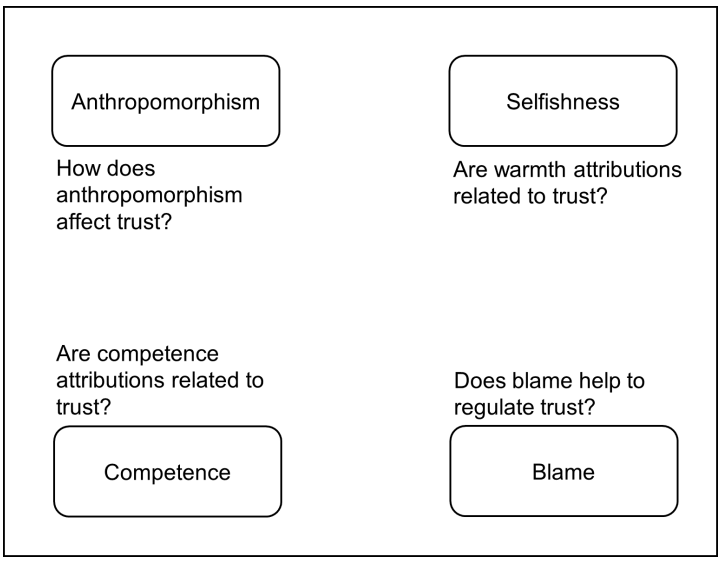


Figure 35: Interaction factors in the paradigm.

The paradigm sheds light on how humans cooperate with computers. This includes the role of trust and how shared human–computer activities should be designed in order to support problem-solving processes. The relationship between trust and cooperation is of particular interest because in contrast to social dilemmas, the problem-solving process relied on competent action. In the present interaction paradigm, cooperation pertains to the collective efforts geared toward the joint goal, and team performance is a key indicator of successful

cooperation. Only weak to moderate relationships emerged between trust and team performance. These relationships were not consistent but occurred in single rounds. Structurally, the findings cannot differentiate between the effect of trust on cooperation and vice versa, but they show the importance of consistent trustworthy behavior. In Experiment 2, bad advice by the agent decreased its overall perceived trustworthiness, although its block placements were configured to optimally achieve the goal. Thus in complex multi-dimensional cooperation, agents are punished for untrustworthy behavior, even if it occurs on single problem-solving dimensions such as advice giving.

Which interaction factors were related to team performance as cooperative outcome? Overall, performance variables in the form of advice quality showed relationships with team performance, whereas anthropomorphism remained insignificant (Experiment 2 and 3). For anthropomorphism to have a substantial impact, it can be assumed that the agent needs to be more persuasive when providing advice and should provide it more often. In the experiments, persuasiveness was not manipulated on purpose in order to focus on the effect of the mere presence and voice output of anthropomorphic agents.

The role of anthropomorphism in the context of cooperation remains largely unexplored. Experiment 2 showed how anthropomorphism shapes the course of cooperative interactions. In this study, participants requested more advice from anthropomorphic compared to non-anthropomorphic agents. Although anthropomorphism did not affect trust in advice, it increased opportunities for task-related exchange. There are two potential benefits to that. First, increased opportunities for exchanging task-specific information could facilitate cooperation in the long run. Second, increased opportunities could also lead to higher advice adoption. To clarify this argument, another result should be considered. Experiment 2 and 3 revealed almost consistent moderate to strong relationships between requested and adopted advice. Accordingly, the likelihood to adopt computer advice depends to some degree on the request rate. Requests can be seen as pre-evaluation phase that prepares the trust decision. One explanation for this relationship could be that a certain amount of trust is already put into the agent during pre-evaluation, and this readiness has an effect on the trust decision. If we assume that participants were also inclined to adopt incompetent advice merely because they had requested it, this would be considered as poorly calibrated trust. Still, it remains unclear how exactly anthropomorphism affected people to request more advice. Self-reported trust and perceived trustworthiness can be ruled out as they were not related to the request rate. Potential reasons are mere curiosity or simply liking the anthropomorphic agent more. As caveat, note that the correlations do not distinguish between competent and incompetent advice. This highlights

that the use of anthropomorphism should be evaluated carefully (see Lee & See, 2004).

Another important aspect of cooperation is to decide when and how task-related advice should be made available, because this factor directly influences goal achievement. Anthropomorphism is only one possible mechanism that may prompt human decision-makers to consider advice. Likewise, advice utilization is not only determined by trust in advisors; other factors include the competence differential between advisors and advice takers, quality of advice, performance-contingent rewards, and advisors' confidence in their own advice (see Bonaccio & Dalal, 2006, for a review). The experiments investigated two different approaches. Experiment 2 had fixed points in each round, whereas Experiment 3 offered participants to request advice at any given time with a fixed maximum equal to Experiment 2. As a consequence, the number of requested and adopted good advice in Experiment 3 was somewhat lower, with the exception of the final round, which imposed higher uncertainty because advice was no longer shown first but implemented directly by the agent. Previous research suggest large individual differences not only regarding advice utilization, but also whether advice utilization actually leads to better outcomes in the first place (Gardner & Berry, 1995). This implies that efforts to generally increase advice utilization in human–computer cooperation should be viewed with caution. Such efforts may not only have positive effects on the outcome; for some, advice may function as a crutch that impedes learning (Gardner & Berry, 1995). Against this background, the presentation and complexity of the underlying problem-solving process and role of the agent is highlighted. Recall that even without advice, participants cooperated with the agent. It is thus important to emphasize the agent's overall cooperativeness and competence to ensure that especially *crucial* advice is utilized throughout the interaction. On the one had, anthropomorphism could support advice utilization in numerous ways, for instance by providing positive affective feedback as reward, reinforcing confidence in the advice, and stressing the importance of finding an optimal solution at vital moments. Because anthropomorphism is associated with trust resilience (Visser et al., 2016), anthropomorphic cues could help keeping advice utilization more steadily. On the other hand, negative effects of advice utilization should be prevented. Anthropomorphic agents could mitigate these side effects by emphasizing humans' contribution to the goal and facilitate their self-efficacy.

Across all experiments, the same puzzle game principle was used to investigate cooperation from different angles: social attributions in strategic social decision-making, people's responses to computer advice of agents varying in anthropomorphism, and how people respond to blaming behavior of an anthropomorphic agent. Still, the paradigm had to simplify various key elements of cooperative ac-

tivities. Truly naturalistic cooperation can involve numerous different scenarios, goals, resource conflicts, communication forms, group sizes, and so forth. Broadening the experimental scope and setup to account for some of those variations is important in order to infer how agents capable of cooperation can be designed. Social dilemma setups are adequate for investigating factors such as communication, group size, and decision framing (e.g., giving vs. taking) (Brewer & Kramer, 1986; Liebrand, 1984), but they are a poor choice if competence or performance, respectively, are to be studied. Deviating from social dilemmas implied a constant challenge regarding participants' actual understanding of the requirements of the human–computer task. To ensure the overall perceived complexity would be as low as possible, the task and controls were kept simple. Furthermore, the role as well as contributions of the agent to the problem-solving were made explicit and observable. Although the agents were not endowed with computational mechanisms that explicitly facilitate cooperation, observations and debriefing dialogs indicate that most participants interacted easily with them.

9.4 DESIGNING FOR TRUST AND TRUSTWORTHINESS

Lastly, the empirical and methodological contributions speak to the issue of designing trustworthy computer agents that elicit appropriate trust. Agents typically gain trustworthiness based on their system capabilities. Such capabilities should be congruent with the level of user trust to facilitate calibrated or appropriate trust (Lee & See, 2004). Going beyond this, future forms of virtual agents and social robots could be equipped with design features that allow them to autonomously modulate their own trustworthiness, for instance by means of anthropomorphism (DeSteno et al., 2012; Lee et al., 2013; Lucas et al., 2016). The engineering of trustworthy behavior will need to be grounded in psychology and communication studies in order to take into account how humans encode (e.g., Boone & Buck, 2003; Oosterhof & Todorov, 2008; Todorov, 2008; Winston et al., 2002) and decode (Fiske et al., 2007; Judd et al., 2005) trustworthiness.

By showing how warmth mediates the relation between task-related behavior and trustworthiness, some of the ambiguity between trustworthiness and warmth has gained a new perspective through the present thesis. Warmth can be an important source of agent trustworthiness. Still, major challenges remain. It is largely unknown how social perception weighs and integrates different cues of artificial warmth and competence to form trustworthiness judgments. Furthermore, as human–computer interactions are increasingly complex, it becomes impracticable to script flexible yet robust trustworthy behavior for entire interactions. At the same time, there is little hope for uncovering single universal cues of trustworthiness (see DeSteno et al.,

2012). As a consequence, it could be more feasible to focus on basic patterns of behavioral trustworthiness that allow the strengthening of appropriate trust, much like a simple "trustworthiness language", and investigate how these patterns are appraised in dynamic social interactions (see also Hoffman et al., 2013). The initial question is this: Given a specific context, what would a trustworthy intelligent agent (not) do?

The following suggestions for trustworthy behavior can be derived based on this thesis' findings:

- **Blame:** In contrast to participant-blame, a self-blaming agent was liked better and received higher trustworthiness attributions. Other media such as animated movies and video games have long demonstrated how anthropomorphic cues can believably convey regret, disappointment, and anger. Experiment 4 showed how those behaviors can also decrease felt shame, guilt, irritation, and increase camaraderie in cooperation.

- **Computer advice:** It is unclear why participants sought more advice from an anthropomorphic agent in one experiment but did not in the other. In order to reliably convince users to adopt useful computer advice, however, the advice should be competent. Since subjective ratings indicated a positive effect of anthropomorphism on trustworthiness, anthropomorphic agents could play a supportive role by highlighting their confidence in otherwise ambiguous advice.

- **Goal failure and responsibility:** Without clear evidence of failure being committed by either of the two players, participants refrained from attributing blame to the agent, even if it blamed itself. This has two implications. First, participants did not look for blame were none was to be found. Second, attributing blame to the team is a practical way of preserving the possibility to continue cooperation without imposing further social costs to either of the players: to the participants themselves for having to follow the social blaming script, including persuasive blame (Malle et al., 2014), or to the agent for having to perform further repair strategies.

- **Uncertainty:** Uncertainty is an external factor with potential consequences for the trusting relationship. Decisions under increased uncertainty require increased trust. Notably, there was no relation between trusting decisions under uncertainty (or without uncertainty) and perceived trustworthiness of the agent. A possible explanation is that trustworthiness attributions (i.e., agent perceptions) can become disassociated from decisions to trust (i.e., behavior perceptions). A more obvious immediacy between agents as on-screen characters and their actions within

the task environment could help establish this relation, for instance through additional animation or agent comments (e.g., "I will place this block over there").

- **Strategic behavior:** Due to conflicting interests, cooperation often involves the possibility for the participating agents to act selfishly. Humans are sensitive to the selfishness of their partners, and unselfishness is a reliable way to increase trustworthiness. The following design considerations advocate a cautious use of selfishness: Does selfishness potentially impede coordination within the task environment and thus put the joint goal at risk? Will users be willing to continue interacting with a selfish agent? Can the agent use mechanisms to regulate adverse effects of selfishness? Unselfishness alone does not ensure favorable trustworthiness judgments: unselfish compliance with human advice can make an agent seem less able if its competence is questionable ("Will do but can't do"). In this case, an agent should try to reduce uncertainty before implementing advice in a wrong way.

- **Competence:** More importantly, human cooperators must be provided with information to identify incompetent agents. Performance variables were shown to be among the most consistent determinants of trust. One method to support performance evaluations and competence attributions is to provide continuous feedback of agent performance (Dzindolet et al., 2003). For future human–agent cooperations, this very feedback should not only entail the overall goal performance of agents, but also information about *how* they tend to coordinate (e.g., as first-mover, by focusing on easy vs. difficult sub-tasks) and respond to their partner in order to support user preferences.

Figure 36 shows how the findings can be condensed into a new conceptual model of human–computer trust. Although the proposed model is based on a rather small number of empirical investigations, some features set it apart from previous models on human–computer trust (Bickmore & Cassell, 2001; Hoff & Bashir, 2015; Lee & See, 2004; Madhavan & Wiegmann, 2007). The model describes how perceived warmth can be amplified by anthropomorphic appearance and behaviors as well as task-related performance of agents. Both perceived warmth and competence (performance) are important antecedents of trust, but statistically, the mediating role of perceived warmth was clearer than perceived competence when the agent's intentions were dubious. However, the actual issue to be solved when designing trustworthy agents is not to balance warmth against competence. Competence is a necessary condition of trust, but its effect can be amplified through warmth as well as the factors that determine warmth. In fact,

the model highlights how warmth, too, can be amplified by performance variables. Conversely, trying to compensate for a lack of competence through warmth may be tempting, but such attempts should be considered bad design practice because they are highly misleading.

According to this view, anthropomorphism does not always affect trustworthiness through warmth. Other pathways may be possible as well, yet how they function can only be determined if the hypothesized distinction between warmth and trustworthiness becomes clearer. This could be achieved, for instance, by using a wider range of anthropomorphic cues to manipulate the effect of both constructs on affective and behavioral outcomes. The present and previous findings suggest that warmth captures a broad range of friendliness, kindness, and cooperativeness attributions, whereas trustworthiness is narrowed down to task-related competence, benevolence, and moral integrity (Fiske et al., 2007; Mayer et al., 1995). However, this distinction seems to be too vague to explain if agents can be trustworthy but not warm, and vice versa.

Another characteristic of the proposed model is that its empirical basis was established entirely using the interactive cooperation game paradigm as underlying methodology. The model thus encapsulates how participants trusted and cooperated with the different computer agents, including the inconsistent relation between team performance (cooperation) and trust. In order to improve the model, it is necessary to pinpoint more clearly how cooperative actions form trust, and how trust supports cooperation in the present paradigm.

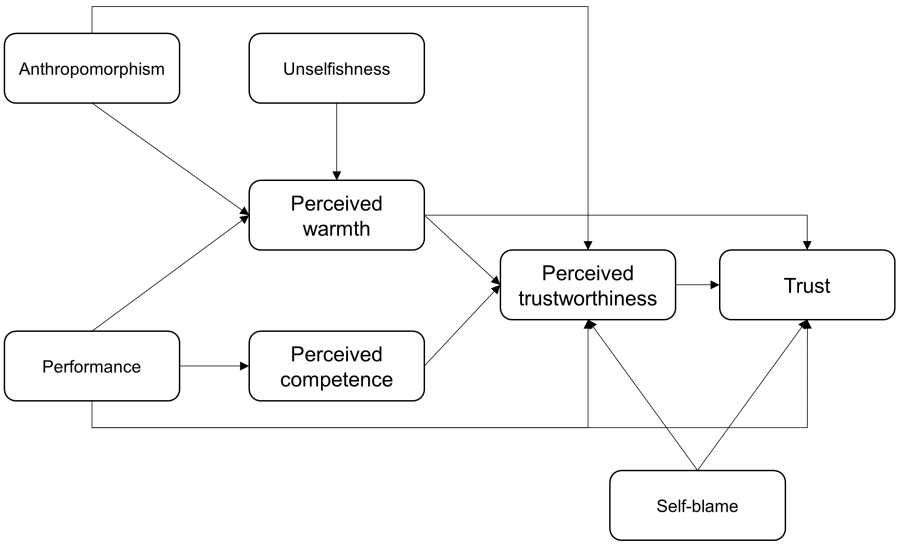


Figure 36: The combined empirical and methodological contributions form a conceptual model of human-computer trust.

9.5 FUTURE RESEARCH DIRECTIONS

This section tries to formulate the most relevant further questions regarding human–computer trust and cooperation.

Warmth and competence

Despite their relevance for social cognition, little is known about how anthropomorphic agents can shape warmth and competence attributions, and how these attributions modulate attitudes, behaviors, and emotions. Currently there is no theoretical integration as to how warmth and competence contribute to trustworthiness, partially because warmth and trustworthiness are often treated as the same, leading to a simplification of trustworthiness (Fiske et al., 2007). Conceptually, however, trustworthiness necessarily entails competence as well. Based on this thesis' findings, HCI research requires dynamic, context-dependent frameworks and models that tackle the following questions:

- Given the overall significance of computer performance for trust, when and why does the relative importance of competence shift to warmth, and vice versa?
- Do the mechanisms and signals of interpersonal warmth and competence generally apply to HCI (see Section 9.2.1), or do the same biases and heuristics that underlie human–computer trust affect the warmth and competence of computers?
- What is the effect of warmth and competence on social influence in other major interaction paradigms (e.g., negotiation, decision support, open conversational dialog, coaching and tutoring)?
- How are warmth and competence attributions of computers influenced by surface variables on the user interface, individual dispositions (e.g., emotional intelligence), cultural background, and demographics?

Trust concept

State-of-the-art HCI studies generally use contemporary trust conceptualizations to the extent that they adopt the trust definitions by Lee and See (2004) or Mayer et al. (1995). However, interaction factors such as risk and vulnerability have become understudied in the context of human–computer trust. It must be assumed that the interplay of perceived risk and benefit crucially affects how people rely on technology with safety-relevant mechanisms such as autonomous vehicles (see Chapter 3). Social dilemmas, for instance, clearly exemplify

risk and benefit by means of their explicit payoff structures, but this ratio is not altered throughout interactions. Accordingly, the following questions should be asked:

- Do contextual factors such as anthropomorphism affect the perception of risk and benefits?
- Is anthropomorphism related to more perceived risk in critical situations because human behavior is supposed to be inherently fallible, or is this effect dampened because anthropomorphism remains a surface variable with low influence?

Mind perception

It has been argued that perceiving mind in computers fundamentally changes people's social responses to computers because in this case, computers are believed to have agency and be capable of human-like experience (Melo & Gratch, 2015). This could have implications for trust:

- Is mind perception related to qualitative changes in trust, for instance a stronger affective foundation?

Trust measures

People's self-reported and behavioral trust are prone to inconsistencies, and the dynamics of trust would require realtime measures to maximize accuracy. As shown by the effect of anthropomorphism on advice requests (Experiment 2), choices that lead to trust decisions should also be considered for the evaluation of trust in cooperation, because trust decisions and these earlier choices can be positively related.

- How can self-reported trust be obtained unobtrusively in cooperation?
- Is there a feasible way to trace the effect of different types of cooperation (conditional vs. unconditional, reciprocal, commitment-based) on trust in repeated interactions?
- What are the differences between decision-making tasks as used for the present work and more open interactions such as conversational dialog with respect to the nature and antecedents of trust? How can agents establish trustworthiness in open dialog?

Cooperation

A central conclusion of the present work is that the interaction dimensions which structurally define problem-solving, exchange of infor-

mation, and strategic components need to be considered in human–computer cooperation. This is especially true if they correspond to key concepts of cooperation and thus can either support or impede cooperation (see Chapter 4).

- Very little is known about how symmetric exchanges of trust could affect cooperation, in particular, how humans appraise being categorized as untrustworthy by agents, and how they interact with agents that adjust to trusting and distrusting feedback.
- Anecdotal evidence suggested some participants enjoyed the puzzle game more when playing on their own. What are the effects of human–computer cooperation on the interaction experience (e.g., perceived control) and performance?
- There is no definitive guideline as to what human–computer cooperation should look like. From a human-centered perspective, cooperation should support human needs and goals. Under which conditions are the demands of cooperation (e.g., coordination and communication) detrimental to human goals?
- Computer agents are increasingly capable of human-like problem-solving and occupy roles as intelligent collaborators. How does this affect the balancing of guidance and responsibility between humans and computers?

Anthropomorphism & human–computer trust

Finally, based on the literature review in earlier chapters and despite the present experiments, it has not yet been shown if anthropomorphism can make a significant and unequivocal difference for human–computer trust to the extent that its influence is not limited to specific interaction paradigms and we can predict how, when, and why the effect of anthropomorphism dynamically changes and evolves in the long run. The main difference between interpersonal and human–computer trust has never been of quantitative nature; it is not particularly challenging to show how a human and computer target evoke the same *amount* of user trust. Still, both forms of trust are based on different *qualities*. The following points address this issue:

- A stronger focus on mechanisms that describe trust decisions and experiences with humans versus computers on a qualitative level is needed. For instance, there is a qualitative difference between conditional and unconditional trust that comes from different underlying assumptions regarding the interaction partner, that is, self-interest versus trustworthiness (Krueger et al.,

2007). Furthermore, the affective foundation of trust is overlooked for the most part (see also warmth and competence as important focus of research).

- Which disruptive events cause the illusion of human-likeness to break down?
- Should a breakdown be enforced to prevent the influence of human-likeness on decision-making when it is not needed?

9.6 LIMITATIONS

The development of the interactive cooperation paradigm represents an effort to increase the ecological validity of trust and cooperation studies in HCI. While the paradigm provided more interactive problem-solving mechanisms, it is nonetheless based on some *assumptions* that limit ecological validity in their own way.

First, it assumes that participants want to solve a problem they could theoretically solve themselves together with a computer. This assumption does not stand in contrast to many real world applications of intelligent agents. People's daily assistants often provide optional and additional advice to tasks that users could very well solve on their own. Those assistants are valued because they adapt to user needs. Thus the question is: How well did the agents coordinate and provide useful advice? This question leads to the second assumption, that is, all users solve the same problem in the same way. Answering the question above, all participants of a given experiment interacted with the same puzzle solving heuristic. The upcoming action of the agent was at times governed by behavioral manipulations, but this did not change the fact that the underlying heuristic was static. As a consequence, all task approaches by participants were treated the same. Indeed, the optimality with which the agent solved the puzzle occasionally required people to align to the agent. The third and most important assumption is that the actions, tasks, and goals within the paradigm approximate future human–agent cooperation. This assumption is inherently difficult to confirm for any kind of laboratory setting. The most intricate element to uphold in this argument is that the present paradigm allocated actions to users and agents in a realistic fashion. In other words, the sequence, types, and competence cues of the actions could occur in real-world settings as well and do not feel synthetic or imposed. The present paradigm avoided initiative and autonomy issues for the most part. Indeed, this is a rather strong limitation to interactive problem-solving. With increasingly complex human–computer tasks, problem-solving will surely benefit from more flexibility on both sides.

The experiments include a number of methodological limitations. To control for the social effect of individual design and behavior fea-

tures and to increase external validity, different versions of virtual as well as human agents should be used in studies on anthropomorphism. One way to achieve better control of subtle behavior nuances in the stimuli set is to use trained actors as human agents (e.g., Visser et al., 2016). As mentioned earlier, Experiment 2 had a rather small sample size which could have decreased the likelihood of detecting interaction effects. In Experiment 4, it would have been interesting to analyze the effect of blame on the outcome variables while also varying anthropomorphism, including non-anthropomorphic agents.

The final limitation is that the different cooperative interaction factors were not examined together in a holistic fashion, but in four different experiments. A common way to include additional experimental manipulations into single experiments are within-subject designs. For instance, the between-subject factors agent and advice quality could have been changed to within-subject factors. On the flip side, this would increase each participant's number of rounds to play and possibly decrease the social effects of the agent. Combining three different agents, three types of advice quality, and three types of goal difficulty would total 27 rounds for each participant. Apart from simple experimental manipulations, the experiments had differences with respect to their interaction as a whole. The most significant change occurred between the strategic cooperation and advice adoption scenarios, whereas Experiments $2-4$ are more compatible. However, the modifications became apparent in a stepwise fashion and was driven by each prior experiment.

9.7 CONCLUDING REMARKS

This thesis provides another step toward investigating trust in interactive human–computer cooperation. There is little certainty as to how intelligent (virtual) agents and social robots will help humans achieve their goals in the future and how they will be accepted. Current advances in the domain of speech-based assistants show that visual anthropomorphic cues are still implemented with caution. Domains with a higher need for trust and the possibility to achieve user goals through interaction in a manner that feels unobtrusive and not forced, much like humans who naturally engage in social interaction in order to achieve a cooperative goal, could benefit more strongly from warmth and trustworthiness modulations using visual humanlikeness. Against this background, the manifold forms of human–computer cooperation could be ideal to investigate how agent anthropomorphism and capabilities affect user trust. At this point, cooperating with other humans often feels much more natural because computers have little understanding of how task-related behaviors and social actions promote cooperative goals and trust at the same time.

a

APPENDIX

The appendix presents the complete means and standard deviations of the empirical data, some of which were already presented using figures.

A.1 EXPERIMENT 1

The means and standard deviations of warmth, competence, behavioral trust, and trustworthiness are given below.

Table 14: Means and standard deviations for warmth and competence.

		Warmth		Competence	
Selfishness	*Puzzle Competence*	M	SD	M	SD
Unselfish	Competent	1.14	0.66	0.60	0.94
	Incompetent	−0.33	0.83	−0.61	0.85
Selfish	Competent	−0.20	0.85	0.32	0.97
	Incompetent	−0.62	0.64	−0.30	0.82

Table 15: Means and standard deviations of behavioral trust and trustworthiness.

		Behavioral trust		Trustworthiness	
Selfishness	*Puzzle Competence*	M	SD	M	SD
Unselfish	Competent	4.00	1.08	4.00	0.57
	Incompetent	2.45	1.45	2.15	0.52
Selfish	Competent	2.50	1.50	2.71	0.78
	Incompetent	2.10	1.37	1.99	0.59

A.2 EXPERIMENT 2

The means and standard deviations of requested and adopted advice are given below.

Table 16: Means and standard deviations for requested advice.

		Round 1		Round 2		Round 3	
Advice quality	*Agent*	M	SD	M	SD	M	SD
Good	Computer	1.42	0.94	1.71	0.99	1.71	1.07
	VA	1.88	0.72	2.00	0.63	1.75	0.86
Bad	Computer	1.80	1.01	1.60	0.91	1.47	1.19
	VA	1.43	0.94	2.13	0.99	2.13	0.83

Note. VA = Virtual agent.

Table 17: Means and standard deviations for adopted advice.

		Round 1		Round 2		Round 3	
Advice quality	*Agent*	M	SD	M	SD	M	SD
Good	Computer	1.00	0.78	1.50	1.02	1.43	0.85
	VA	1.50	0.82	1.69	0.79	1.44	0.81
Bad	Computer	0.40	0.63	0.33	0.62	0.33	0.90
	VA	0.33	0.62	0.27	0.59	0.20	0.41

Note. VA = Virtual agent.

A.3 EXPERIMENT 3

The means and standard deviations of requested and adopted advice are given below.

Table 18: Means and standard deviations of requested advice.

		Round 1		Round 2	
Advice quality	*Agent*	M	SD	M	SD
Good	Computer	1.68	1.00	1.21	1.08
	Virtual agent	1.68	1.25	0.84	1.21
	Human	1.42	1.17	0.94	0.97
Mixed	Computer	1.00	0.88	0.47	0.84
	Virtual agent	0.74	0.93	0.42	0.69
	Human	1.16	1.17	0.63	1.07

Table 19: Means and standard deviations of adopted advice.

		Round 1		Round 2		Round 3	
Advice quality	*Agent*	M	SD	M	SD	M	SD
Good	Computer	1.11	0.99	0.89	0.88	1.79	1.27
	Virtual agent	1.05	1.08	0.68	1.16	1.26	1.33
	Human	1.05	1.08	0.42	0.61	1.47	1.31
Mixed	Computer	0.11	0.32	0.21	0.54	0.90	1.15
	Virtual agent	0.05	0.23	0.26	0.56	0.42	0.77
	Human	0.05	0.23	0.37	0.90	1.11	1.20

BIBLIOGRAPHY

Akgun, M., Cagiltay, K. & Zeyrek, D. (2010). The effect of apologetic error messages and mood states on computer users' self-appraisal of performance. *Journal of Pragmatics, 42*(9), 2430–2448.

Antos, D., Melo, C. de, Gratch, J. & Grosz, B. (2011). The influence of emotion expression on perceptions of trustworthiness in negotiation. In: *Proceedings of the 25th AAAI Conference*. Menlo Park, CA: AAAI Press, 772–778.

Asch, S. E. (1946). Forming impressions of personality. *The Journal of Abnormal and Social Psychology, 41*(3), 258–290.

Axelrod, R. (1984). The evolution of cooperation. New York, NY: Basic Books.

Bailenson, J. N., Blascovich, J., Beall, A. C. & Loomis, J. M. (2003). Interpersonal distance in immersive virtual environments. *Personality and Social Psychology Bulletin, 29*(7), 819–833.

Balliet, D. & Van Lange, P. A. (2013). Trust, conflict, and cooperation: A meta-analysis. *Psychological Bulletin, 139*(5), 1090–1112.

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

Batson, C. D. & Ahmad, N. (2001). Empathy-induced altruism in a prisoner's dilemma II: what if the target of empathy has defected? *European Journal of Social Psychology, 31*(1), 25–36.

Batson, C. D. & Moran, T. (1999). Empathy-induced altruism in a prisoner's dilemma. *European Journal of Social Psychology, 29*(7), 909–924.

Becker-Asano, C. & Wachsmuth, I. (2010). Affective computing with primary and secondary emotions in a virtual human. *Autonomous Agents and Multi-Agent Systems, 20*, 32–49.

Bente, G., Feist, A. & Elder, S. (1996). Person perception effects of computer-simulated male and female head movement. *Journal of Nonverbal Behavior, 20*(4), 213–228.

Berg, J., Dickhaut, J. & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior, 10*(1), 122–142.

Bergmann, K., Eyssel, F. & Kopp, S. (2012). A second chance to make a first impression? How appearance and nonverbal behavior affect perceived warmth and competence of virtual agents over time. In: *Intelligent Virtual Agents*. Berlin, Heidelberg: Springer, 126–138.

Bickmore, T. W. & Picard, R. W. (2005). Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction, 12*(2), 293–327.

Bickmore, T. & Cassell, J. (2001). Relational agents: a model and implementation of building user trust. In: *CHI '01 Proceedings of the*

SIGCHI Conference on Human Factors in Computing Systems. New York: ACM, 396–403.

Blascovich, J. (2002). A theoretical model of social influence for increasing the utility of collaborative virtual environments. In: *Proceedings of the 4th International Conference on Collaborative Virtual Environments*. New York: ACM, 25–30.

Bonaccio, S. & Dalal, R. S. (2006). Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational Behavior and Human Decision Processes, 101*(2), 127–151.

Boone, R. T. & Buck, R. (2003). Emotional expressivity and trustworthiness: The role of nonverbal behavior in the evolution of cooperation. *Journal of Nonverbal Behavior, 27*(3), 163–182.

Bradshaw, J. M., Dignum, V., Jonker, C. & Sierhuis, M. (2012). Human–agent–robot teamwork. *Intelligent Systems, IEEE, 27*(2), 8–13.

Brave, S. & Nass, C. (2009). Emotion in human–computer interaction. In: *Human-Computer Interaction Fundamentals*. Ed. by A. Sears & J. A. Jacko. CRC Press, 53–68.

Breazeal, C. & Scassellati, B. (1999). How to build robots that make friends and influence people. In: *IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS'99*. IEEE, 858–863.

Breazeal, C., Kidd, C. D., Thomaz, A. L., Hoffman, G. & Berlin, M. (2005). Effects of nonverbal communication on efficiency and robustness in human-robot teamwork. In: *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 708–713.

Breazeal, C., Harris, P. L., DeSteno, D., Kory Westlund, J. M., Dickens, L. & Jeong, S. (2016). Young Children Treat Robots as Informants. *Topics in Cognitive Science, 8*(2), 481–491.

Brewer, M. B. & Kramer, R. M. (1986). Choice behavior in social dilemmas: Effects of social identity, group size, and decision framing. *Journal of Personality and Social Psychology, 50*(3), 543.

Briggs, G. & Scheutz, M. (2015). "Sorry, I can't do that": Developing mechanisms to appropriately reject directives in human-robot interactions. In: *2015 AAAI Fall Symposium Series*.

Brosnan, S. F., Salwiczek, L. & Bshary, R. (2010). The interplay of cognition and cooperation. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences, 365*(1553), 2699–2710.

Buchholz, V., Kulms, P. & Kopp, S. (2017). It's (not) your fault! Blame and trust repair in human-agent cooperation. In: *Proceedings of the 2017 Workshop on Cognitive Systems*. München.

Burgoon, J., Bonito, J., Bengtsson, B., Cederberg, C., Lundeberg, M. & Allspach, L. (2000). Interactivity in human–computer interaction: A study of credibility, understanding, and influence. *Computers in Human Behavior, 16*(6), 553–574.

Buschmeier, H. & Kopp, S. (2011). Towards conversational agents that attend to and adapt to communicative user feedback. In: *Intelligent Virtual Agents*. Springer, 169–182.

Camerer, C. F. (2003). Behavioral game theory: Experiments in strategic interaction. Princeton, NJ: Princeton University Press.

Campbell, L., Simpson, J. A., Kashy, D. A. & Fletcher, G. J. (2001). Ideal standards, the self, and flexibility of ideals in close relationships. *Personality and Social Psychology Bulletin, 27*(4), 447–462.

Casciaro, T. & Lobo, M. S. (2008). When competence is irrelevant: The role of interpersonal affect in task-related ties. *Administrative Science Quarterly, 53*(4), 655–684.

Cassell, J. & Bickmore, T. (2000). External manifestations of trustworthiness in the interface. *Communications of the ACM, 43*(12), 50–56.

Chiles, T. H. & McMackin, J. F. (1996). Integrating variable risk preferences, trust, and transaction cost economics. *The Academy of Management Review, 21*(1), 73–99.

Choi, A., Melo, C. M. de, Khooshabeh, P., Woo, W. & Gratch, J. (2015). Physiological evidence for a dual process model of the social effects of emotion in computers. *International Journal of Human-Computer Studies, 74*, 41–53.

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

Corritore, C. L., Kracher, B. & Wiedenbeck, S. (2003). On-line trust: Concepts, evolving themes, a model. *International Journal of Human-Computer Studies, 58*(6), 737–758.

Couch, L. L. & Jones, W. H. (1997). Measuring Levels of Trust. *Journal of Research in Personality, 31*(3), 319–336.

Cuddy, A. J., Fiske, S. T. & Glick, P. (2008). Warmth and competence as universal dimensions of social perception: The stereotype content model and the BIAS map. *Advances in Experimental Social Psychology, 40*, 61–149.

Cuddy, A. J. C., Fiske, S. T. & Glick, P. (2007). The BIAS map: Behaviors from intergroup affect and stereotypes. *Journal of Personality and Social Psychology, 92*(4), 631–648.

Cuddy, A. J., Glick, P. & Beninger, A. (2011). The dynamics of warmth and competence judgments, and their outcomes in organizations. *Research in Organizational Behavior, 31*, 73–98.

Cushman, F. (2013). The role of learning in punishment, prosociality, and human uniqueness. In: *Cooperation and its evolution (Life and mind: Philosophical issues in biology and psychology)*. Ed. by K. Sterelny, R. Joyce, B. Calcott & B. Fraser. Cambridge, MA: MIT Press, 333–372.

Dautenhahn, K. (1998). The art of designing socially intelligent agents: Science, fiction, and the human in the loop. *Applied Artificial Intelligence, 12*(7-8), 573–617.

— (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.

Dawes, R. M. (1980). Social Dilemmas. *Annual Review of Psychology, 31*(1), 169–193.

DeSteno, D. et al. (2012). Detecting the trustworthiness of novel partners in economic exchange. *Psychological Science*, 1549–1556.

DeVault, D. et al. (2014). SimSensei Kiosk: A virtual human interviewer for healthcare decision support. In: *Proceedings of the 2014 International Conference on Autonomous Agents and Multiagent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 1061–1068.

Deutsch, M. (2011). Cooperation and competition. In: *Conflict, interdependence, and justice*. Ed. by P. T. Coleman. New York, NY: Springer, 23–40.

Deutsch, M. (1962). Cooperation and trust: Some theoretical notes. In: *Nebraska Symposium on Motivation*. Ed. by M. R. Jones. Oxford, England: University of Nebraska Press, 275–320.

Dijkstra, J. J., Liebrand, W. B. & Timminga, E. (1998). Persuasiveness of expert systems. *Behaviour & Information Technology, 17*(3), 155–163.

Dongen, K. van & van Maanen, P.-P. (2013). A framework for explaining reliance on decision aids. *International Journal of Human-Computer Studies, 71*(4), 410–424.

Dunn, P. (2000). The importance of consistency in establishing cognitive-based trust: A laboratory experiment. *Teaching Business Ethics, 4*(3), 285–306.

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G. & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies, 58*(6), 697–718.

Epley, N., Waytz, A. & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review, 114*(4), 864.

Fiske, S. T., Cuddy, A. J. C. & Glick, P. (2007). Universal dimensions of social cognition: warmth and competence. *Trends in Cognitive Sciences, 11*(2), 77–83.

Fletcher, G. J. O., Simpson, J. A., Thomas, G. & Giles, L. (1999). Ideals in intimate relationships. *Journal of Personality and Social Psychology, 76*(1), 72.

Fogg, B. J. (2003). Persuasive technology: Using computers to change what we think and do. San Francisco, CA: Morgan Kaufmann Publishers.

Fogg, B. & Tseng, H. (1999). The elements of computer credibility. In: *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. ACM, 80–87.

Frith, C. D. & Singer, T. (2008). The role of social cognition in decision making. *Philosophical Transactions of the Royal Society of London B: Biological Sciences, 363*(1511), 3875–3886.

Fukuyama, F. (1995). Trust: The social virtues and the creation of prosperity. New York, NY: The Free Press.

Gächter, S. (2004). Behavioral game theory. In: *Blackwell Handbook of Judgment and Decision Making*. Ed. by D. J. Koehler & H. Nigel. Blackwell: Malden, Massachusetts, 485–503.

Gambetta, D (1990). Can we trust trust. In: *Trust: Making and Breaking Cooperative Relations*. Ed. by D. Gambetta. Vol. 13. Blackwell Publishers, 213–237.

Gardner, P. H. & Berry, D. C. (1995). The effect of different forms of advice on the control of a simulated complex system. *Applied Cognitive Psychology, 9*(7).

Gratch, J. & Marsella, S. (2004). A domain-independent framework for modeling emotion. *Cognitive Systems Research, 5*(4), 269–306.

Gratch, J., Wang, N., Gerten, J., Fast, E. & Duffy, R. (2007). Creating rapport with virtual agents. In: *International Conference on Intelligent Virtual Agents*. Springer, 125–138.

Gratch, J., DeVault D., Lucas, G. M. & Marsella, S. (2015). Negotiation as a Challenge Problem for Virtual Humans. In: *Intelligent Virtual Agents*. Ed. by W.-P. Brinkman, J. Broekens & D. Heylen. Berlin, Heidelberg: Springer, 201–215.

Gratch, J., DeVault, D. & Lucas, G. (2016). The benefits of virtual humans for teaching negotiation. In: *Intelligent Virtual Agents*. Springer International Publishing, 283–294.

Gray, H. M., Gray, K. & Wegner, D. M. (2007). Dimensions of mind perception. *Science, 315*(5812), 619–619.

Groom, V., Chen, J., Johnson, T., Kara, F. A. & Nass, C. (2010). Critic, compatriot, or chump? Responses to robot blame attribution. In: *HRI '10 Proceedings of the 5th ACM/IEEE International Conference on Human-Robot Interaction*. IEEE Press Piscataway, 211–217.

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

Hardin, R. (2002). Trust and trustworthiness. Russell Sage Foundation.

Hoc, J.-M. (2000). From human–machine interaction to human–machine cooperation. *Ergonomics, 43*(7), 833–843.

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

Hoffman, R. R., Johnson, M., Bradshaw, J. M. & Underbrink, A. (2013). Trust in automation. *IEEE Intelligent Systems, 28*(1), 84–88.

Hovland, C. I., Janis, I. L. & Kelley, H. H. (1953). Communication and persuasion. Psychological studies of opinion change. New Haven, CT: Yale University Press.

Hudson, M. & Cairns, P. (2014). Interrogating social presence in games with experiential vignettes. *Entertainment Computing, 5*(2), 101–114.

Isbister, K. & Nass, C. (2000). Consistency of personality in interactive characters: verbal cues, non-verbal cues, and user characteristics. *International Journal of Human-Computer Studies, 53*(2), 251–267.

Janssen, M. A. (2008). Evolution of cooperation in a one-shot prisoner's dilemma based on recognition of trustworthy and untrustworthy agents. *Journal of Economic Behavior & Organization, 65*(3), 458–471.

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

Jordan, J. J., Hoffman, M., Nowak, M. A. & Rand, D. G. (2016). Uncalculating cooperation is used to signal trustworthiness. *Proceedings of the National Academy of Sciences, 113*(31), 8658–8663.

Judd, C. M., James-Hawkins, L., Yzerbyt, V. & Kashima, Y. (2005). Fundamental dimensions of social judgment: Understanding the relations between judgments of competence and warmth. *Journal of Personality and Social Psychology, 89*(6), 899–913.

Kaniarasu, P. & Steinfeld, A. M. (2014). Effects of blame on trust in human robot interaction. In: *RO-MAN: the 23rd IEEE International Symposium on Robot and Human Interactive Communication, 2014*. Piscataway, NJ: IEEE, 850–855.

Khooshabeh, P., McCall, C., Gandhe, S., Gratch, J. & Blascovich, J. (2011). Does it matter if a computer jokes. In: *CHI'11 Extended Abstracts on Human Factors in Computing Systems*. ACM, 77–86.

Kiesler, S., Sproull, L. & Waters, K. (1996). A prisoner's dilemma experiment on cooperation with people and human-like computers. *Journal of Personality and Social Psychology, 70*(1), 47.

Kim, P. H., Ferrin, D. L., Cooper, C. D. & Dirks, K. T. (2004). Removing the shadow of suspicion: the effects of apology versus denial for repairing competence-versus integrity-based trust violations. *Journal of Applied Psychology, 89*(1), 104.

Kim, P. H., Dirks, K. T., Cooper, C. D. & Ferrin, D. L. (2006). When more blame is better than less: The implications of internal vs. external attributions for the repair of trust after a competence- vs. integrity-based trust violation. *Organizational Behavior and Human Decision Processes, 99*(1), 49–65.

Kirsh, D. & Maglio, P. (1994). On distinguishing epistemic from pragmatic action. *Cognitive Science, 18*(4), 513–549.

Klein, G., Woods, D., Bradshaw, J., Hoffman, R. & Feltovich, P. (2004). Ten Challenges for Making Automation a "Team Player" in Joint Human-Agent Activity. *IEEE Intelligent Systems, 19*(06), 91–95.

Kok, I. de, Hough, J., Hülsmann, F., Botsch, M., Schlangen, D. & Kopp, S. (2015). A Multimodal System for Real-Time Action Instruction in Motor Skill Learning. In: *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 355–362.

Kopp, S. & Wachsmuth, I. (2004). Synthesizing multimodal utterances for conversational agents. *Computer Animation and Virtual Worlds, 15*(1), 39–52.

Kopp, S. et al. (2006). Towards a common framework for multimodal generation: The behavior markup language. In: *International Conference on Intelligent Virtual Agents*. Springer, 205–217.

Kosfeld, M., Heinrichs, M., Zak, P. J., Fischbacher, U. & Fehr, E. (2005). Oxytocin increases trust in humans. *Nature, 435*(7042), 673–676.

Krach, S., Hegel, F., Wrede, B., Sagerer, G., Binkofski, F. & Kircher, T. (2008). Can machines think? Interaction and perspective taking with robots investigated via fMRI. *PLOS ONE, 3*(7), e2597.

Krämer, N. C., Pütten, A. M. Rosenthal-von der & Hoffmann, L. (2015). Social effects of virtual and robot companions. In: *The Handbook of the Psychology of Communication Technology*. Ed. by S. S. Sundar. John Wiley & Sons, 137–159.

Krueger, F. et al. (2007). Neural correlates of trust. *Proceedings of the National Academy of Sciences, 104*(50), 20084–20089.

Krueger, J. I. & Acevedo, M. (2007). Perceptions of self and other in the prisoner's dilemma: Outcome bias and evidential reasoning. *The American Journal of Psychology, 120*(4), p 593–618.

Krumhuber, E., Manstead, A. S. R., Cosker, D., Marshall, D., Rosin, P. L. & Kappas, A. (2007). Facial dynamics as indicators of trustworthiness and cooperative behavior. *Emotion, 7*(4), 730–735.

Kulms, P. & Kopp, S. (2016). The effect of embodiment and competence on trust and cooperation in human–agent interaction. In: *Intelligent Virtual Agents*. Springer. Berlin, Heidelberg, 75–84.

— (2018). A social cognition perspective on human–computer trust: The effect of perceived warmth and competence on trust in decision-making with computers. *Frontiers in Digital Humanities, 5*, 14.

Kulms, P., Krämer, N., Gratch, J. & Kang, S.-H. (2011). It's in their eyes: A study on female and male virtual humans' gaze. In: *Intelligent Virtual Agents*. Springer. Berlin, Heidelberg, 80–92.

Kulms, P., Kopp, S. & Krämer, N. C. (2014a). Let's be serious and have a laugh: Can humor support cooperation with a virtual agent? In: *Intelligent Virtual Agents*. Springer. Berlin, Heidelberg, 250–259.

Kulms, P., Welbergen, H. van & Kopp, S. (2014b). Prototyping von intuitiven und interaktiven Benutzerschnittstellen: Schnelles und einfaches Design von Anwendungen mit virtuellen Agenten [Prototyping of intuitive and interactive interfaces: Easily designing

virtual agent applications]. In: *Technische Unterstützungssysteme, die die Menschen wirklich wollen*. Ed. by W. Robert & R. Tobias. Hamburg: Helmut-Schmidt-Universität, 30–38.

Kulms, P., Mattar, N. & Kopp, S. (2015). Modeling decision-making in cognitive systems for social cooperation games. In: *Proceedings of the 2015 Workshop on Cognitive Systems*. Bielefeld.

— (2016). Can't do or won't do? Social attributions in human–agent cooperation. In: *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 1341–1342.

Kulms, P., Welbergen, H. van & Kopp, S. (2018). MultiPro: Prototyping Multimodal UI with Anthropomorphic Agents. In: *Mensch und Computer 2018 - Tagungsband*. Ed. by R. Dachselt & G. Weber. Bonn: Gesellschaft für Informatik e.V., 23–32.

Lafferty, J., Eady, P. & Elmers, J (1974). The desert survival problem. Plymouth, MI: Experimental Learning Methods.

Lee, J. D. & 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. J., Knox, W. B., Wormwood, J. B., Breazeal, C. & DeSteno, D. (2013). Computationally modeling interpersonal trust. *Frontiers in Psychology, 4*.

Lee, J. & Moray, N. (1992). Trust, control strategies and allocation of function in human–machine systems. *Ergonomics, 35*(10), 1243–1270.

Lewicki, R. J. & Bunker, B. B. (1996). Developing and maintaining trust in work relationships. In: *Trust in Organizations: Frontiers of Theory and Research*. Ed. by R. M. Kramer & T. R. Tyler. Thousand Oaks, CA: Sage, 114–139.

Lewicki, R. J. (2006). Trust, trust development, and trust repair. In: *The Handbook of Conflict Resolution*. Ed. by M. Deutsch, P. T. Coleman & E. C. Marcus. San Francisco: Jossey-Bass Publishers, 92–119.

Lewis, J. D. & Weigert, A. (1985). Trust as a social reality. *Social Forces, 63*(4), 967–985.

Liebrand, W. B. (1984). The effect of social motives, communication and group size on behaviour in an N-person multi-stage mixed-motive game. *European Journal of Social Psychology, 14*(3), 239–264.

Lin, R. & Kraus, S. (2010). Can automated agents proficiently negotiate with humans? *Communications of the ACM, 53*(1), 78–88.

Lindstedt, J. K. & Gray, W. D. (2015). Meta-T: Tetris® as an experimental paradigm for cognitive skills research. *Behavior Research Methods*, 1–21.

Lucas, G., Stratou, G., Lieblich, S. & Gratch, J. (2016). Trust me: Multimodal signals of trustworthiness. In: *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, 5–12.

Macrae, C. N. & Bodenhausen, G. V. (2000). Social cognition: thinking categorically about others. *Annual review of psychology, 51*, 93–120.

Madhavan, P. & Wiegmann, D. A. (2007). Similarities and differences between human–human and human–automation trust: An integrative review. *Theoretical Issues in Ergonomics Science, 8*(4), 277–301.

Malle, B. F. & Knobe, J. (1997). The folk concept of intentionality. *Journal of Experimental Social Psychology, 33*(2), 101–121.

Malle, B. F. & Scheutz, M. (2014). Moral competence in social robots. In: *2014 IEEE International Symposium on Ethics in Science, Technology and Engineering*. IEEE, 1–6.

Malle, B. F., Guglielmo, S. & Monroe, A. E. (2014). A theory of blame. *Psychological Inquiry, 25*(2), 147–186.

Mattar, N., Welbergen, H. van, Kulms, P. & Kopp, S. (2015). Prototyping user interfaces for investigating the role of virtual agents in human–machine interaction. In: *Intelligent Virtual Agents*. Springer. Berlin, Heidelberg, 356–360.

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

McAllister, D. J. (1995). Affect- and cognition-based trust as foundations for interpersonal cooperation in organizations. *The Academy of Management Journal, 38*(1), 24–59.

McCabe, K., Houser, D., Ryan, L., Smith, V. & Trouard, T. (2001). A functional imaging study of cooperation in two-person reciprocal exchange. *Proceedings of the National Academy of Sciences, 98*(20), 11832–11835.

Mealey, L., Daood, C. & Krage, M. (1996). Enhanced memory for faces of cheaters. *Ethology and Sociobiology, 17*(2), 119–128.

Melo, C. M. de & Gratch, J. (2015). Beyond believability: Quantifying the differences between real and virtual humans. In: *Intelligent Virtual Agents*. Berlin, Heidelberg: Springer, 109–118.

Melo, C. M. de, Carnevale, P. J., Read, S. J. & Gratch, J. (2014). Reading people's minds from emotion expressions in interdependent decision making. *Journal of Personality and Social Psychology, 106*(1), 73–88.

Melo, C. M. de, Gratch, J. & Carnevale, P. J. (2015). Humans versus computers: Impact of emotion expressions on people's decision making. *IEEE Transactions on Affective Computing, 6*(2), 127–136.

Miwa, K., Terai, H. & Hirose, S. (2008). Social responses to collaborator: Dilemma game with human and computer agent. In: *Proceedings of the 30th Annual Conference of the Cognitive Science Society*. Ed. by B. C. Love, K. McRae & V. M. Sloutsky. Austin, TX: Cognitive Science Society, 2455–2460.

Moon, Y. (2003). Don't blame the computer: When self-disclosure moderates the self-serving bias. *Journal of Consumer Psychology, 13*(1), 125–137.

Muir, B. M. & Moray, N. (1996). Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics, 39*(3), 429–460.

Muir, B. M. (1987). Trust between humans and machines, and the design of decision aids. *International Journal of Man-Machine Studies, 27*(5-6), 527–539.

Muir, B. M. (1994). Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics, 37*(11), 1905–1922.

Mumm, J. & Mutlu, B. (2011). Designing motivational agents: The role of praise, social comparison, and embodiment in computer feedback. *Computers in Human Behavior, 27*(5), 1643–1650.

Nass, C. & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues, 56*(1), 81–103.

Nass, C., Steuer, J. & Tauber, E. R. (1994). Computers are social actors. In: *CHI '94 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York: ACM, 72–78.

Nass, C., Moon, Y., Fogg, B. J., Reeves, B. & Dryer, D. (1995). Can computer personalities be human personalities? *International Journal of Human-Computer Studies, 43*(2), 223–239.

Nass, C., Fogg, B. & Moon, Y. (1996). Can computers be teammates? *International Journal of Human-Computer Studies, 45*(6), 669–678.

Nguyen, T.-H. D. et al. (2015). Modeling warmth and competence in virtual characters. In: *Intelligent Virtual Agents*. Berlin, Heidelberg: Springer, 167–180.

Nicholson, C. Y., Compeau, L. D. & Sethi, R. (2001). The role of interpersonal liking in building trust in long-term channel relationships. *Journal of the Academy of Marketing Science, 29*(1), 3.

Niewiadomski, R., Demeure, V. & Pelachaud, C. (2010). Warmth, competence, believability and virtual agents. In: *Intelligent Virtual Agents*. Berlin, Heidelberg: Springer, 272–285.

Nishio, S., Ogawa, K., Kanakogi, Y., Itakura, S. & Ishiguro, H. (2012). Do robot appearance and speech affect people's attitude? Evaluation through the ultimatum game. In: *2012 RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication*, 809–814.

Nowak, K. L. & Biocca, F. (2003). The effect of the agency and anthropomorphism on users' sense of telepresence, copresence, and social presence in virtual environments. *Presence: Teleoperators and Virtual Environments, 12*(5), 481–494.

Olson, J. M. & Zanna, M. P. (1993). Attitudes and attitude change. *Annual Review of Psychology, 44*(1), 117–154.

Oosterhof, N. N. & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences, 105*(32), 11087–11092.

Parasuraman, R. & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors, 52*(3), 381–410.

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

Parise, S., Kiesler, S., Sproull, L. & Waters, K. (1996). My partner is a real dog: Cooperation with social agents. In: *Proceedings of the 1996 ACM Conference on Computer Supported Cooperative Work*. ACM, 399–408.

Parise, S., Kiesler, S., Sproull, L. & Waters, K (1999). Cooperating with life-like interface agents. *Computers in Human Behavior, 15*(2), 123–142.

Parsons, S. & Wooldridge, M. (2002). Game theory and decision theory in multi-agent systems. *Autonomous Agents and Multi-Agent Systems, 5*(3), 243–254.

Peeters, G. (2002). From good and bad to can and must: Subjective necessity of acts associated with positively and negatively valued stimuli. *European Journal of Social Psychology, 32*(1), 125–136.

Picard, R. W. (1997). Affective computing. Cambridge, MA: MIT Press.

Preacher, K. J. & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods, 40*(3), 879–891.

Pütten, A. M. von der, Krämer, N. C., Gratch, J. & Kang, S.-H. (2010). "It doesn't matter what you are!" Explaining social effects of agents and avatars. *Computers in Human Behavior, 26*(6), 1641–1650.

Qiu, L. & Benbasat, I. (2009). Evaluating anthropomorphic product recommendation agents: A social relationship perspective to designing information systems. *Journal of Management Information Systems, 25*(4), 145–182.

— (2010). A study of demographic embodiments of product recommendation agents in electronic commerce. *International Journal of Human-Computer Studies, 68*(10), 669–688.

Reeder, G. D., Kumar, S., Hesson-McInnis, M. S. & Trafimow, D. (2002). Inferences about the morality of an aggressor: the role of perceived motive. *Journal of Personality and Social Psychology, 83*(4), 789–803.

Reeves, B. & Nass, C. (1996). The Media Equation: How people treat computers, television, and new media like real people and places. Cambridge, UK: Cambridge University Press.

Rempel, J. K., Holmes, J. G. & Zanna, M. P. (1985). Trust in close relationships. *Journal of Personality and Social Psychology, 49*(1), 95–112.

Rilling, J. K., Sanfey, A. G., Aronson, J. A., Nystrom, L. E. & Cohen, J. D. (2004). The neural correlates of theory of mind within interpersonal interactions. *NeuroImage, 22*(4), 1694–1703.

Rilling, J. K. et al. (2008). The neural correlates of the affective response to unreciprocated cooperation. *Neuropsychologia, 46*(5), 1256–1266.

Rom, S. C., Weiss, A. & Conway, P. (2017). Judging those who judge: Perceivers infer the roles of affect and cognition underpinning others' moral dilemma responses. *Journal of Experimental Social Psychology, 69*, 44–58.

Rosenberg, S., Nelson, C. & Vivekananthan, P. S. (1968). A multidimensional approach to the structure of personality impressions. *Journal of Personality and Social Psychology, 9*(4), 283–294.

Rousseau, D. M., Sitkin, S. B., Burt, R. S. & Camerer, C. (1998). Not so different after all: A cross-discipline view of trust. *The Academy of Management Review, 23*(3), 393–404.

Sagar, M., Seymour, M. & Henderson, A. (2016). Creating connection with autonomous facial animation. *Communications of the ACM, 59*(12), 82–91.

Salem, M., Lakatos, G., Amirabdollahian, F. & Dautenhahn, K. (2015). Would you trust a (faulty) robot? In: *Proceedings of the Tenth International Conference on Human-Robot Interaction*. New York: ACM, 141–148.

Sally, D. (2000). A general theory of sympathy, mind-reading, and social interaction, with an application to the prisoners' dilemma. *Social Science Information, 39*(4), 567–634.

Sandoval, E. B., Brandstetter, J., Obaid, M. & Bartneck, C. (2015). Reciprocity in human-robot interaction: A quantitative approach through the prisoner's dilemma and the ultimatum game. *International Journal of Social Robotics*.

Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E. & Cohen, J. D. (2003). The neural basis of economic decision-making in the ultimatum game. *Science, 300*(5626), 1755–1758.

Sheridan, T. B. & Parasuraman, R. (2005). Human-automation interaction. *Reviews of Human Factors and Ergonomics, 1*(1), 89–129.

Shore, L. M., Tetrick, L. E., Lynch, P. & Barksdale, K. (2006). Social and economic exchange: Construct development and validation. *Journal of Applied Social Psychology, 36*(4), 837–867.

Sidner, C. L., Lee, C., Kidd, C. D., Lesh, N. & Rich, C. (2005). Explorations in engagement for humans and robots. *Artificial Intelligence, 166*(1), 140 –164.

Stirrat, M. & Perrett, D. (2010). Valid facial cues to cooperation and trust. *Psychological Science, 21*(3), 349–354.

Todorov, A. (2008). Evaluating faces on trustworthiness. *Annals of the New York Academy of Sciences, 1124*(1), 208–224.

Tomlinson, E. C., Dineen, B. R. & Lewicki, R. J. (2004). The road to reconciliation: Antecedents of victim willingness to reconcile following a broken promise. *Journal of Management, 30*(2), 165–187.

Trivers, R. L. (1971). The evolution of reciprocal altruism. *The Quarterly Review of Biology, 46*(1), 35–57.

Tsai, J., Bowring, E., Marsella, S., Wood, W. & Tambe, M. (2012). A study of emotional contagion with virtual characters. In: *Intelligent Virtual Agents*. Springer, 81–88.

Tzeng, J.-Y. (2004). Toward a more civilized design: studying the effects of computers that apologize. *International Journal of Human-Computer Studies, 61*(3), 319–345.

Utz, S., Ouwerkerk, J. W. & van Lange, P. A. M. (2004). What is smart in a social dilemma? differential effects of priming competence on cooperation. *European Journal of Social Psychology, 34*(3), 317–332.

Van Lange, P. A. & Kuhlman, D. M. (1994). Social value orientations and impressions of partner's honesty and intelligence: A test of the might versus morality effect. *Journal of Personality and Social Psychology, 67*(1), 126.

Van Lange, P. A., Joireman, J., Parks, C. D. & Van Dijk, E. (2013). The psychology of social dilemmas: A review. *Organizational Behavior and Human Decision Processes, 120*(2), 125–141.

Van't Wout, M., Kahn, R. S., Sanfey, A. G. & Aleman, A. (2006). Affective state and decision-making in the ultimatum game. *Experimental Brain Research, 169*(4), 564–568.

Visser, E. J. de et al. (2016). Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology: Applied, 22*(3), 331–349.

Visser, E. J. de et al. (2017). A little anthropomorphism goes a long way: Effects of oxytocin on trust, compliance, and team performance with automated agents. *Human Factors, 59*(1), 116–133.

Voiklis, J., Cusimano, C. & Malle, B. (2014). A Social-Conceptual Map of Moral Criticism. In: *Proceedings of the 36th Annual Conference of the Cognitive Science Society*. Ed. by P. Bello, M. Guarini, M. McShance & B. Scassellati. Austin, TX: Cognitive Science Society, 1700–1705.

Vugt, H. C. V., Bailenson, J. N., Hoorn, J. F. & Konijn, E. A. (2010). Effects of facial similarity on user responses to embodied agents. *ACM Transactions on Computer-Human Interaction (TOCHI), 17*(2), 7.

Walter, S. et al. (2014). Similarities and differences of emotions in human–machine and human–human interactions: What kind of emotions are relevant for future companion systems? *Ergonomics, 57*(3), 374–386.

Waytz, A., Gray, K., Epley, N. & Wegner, D. M. (2010). Causes and consequences of mind perception. *Trends in Cognitive Sciences, 14*(8), 383–388.

Waytz, A., Heafner, J. & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology, 52*, 113–117.

Weber, J. M., Kopelman, S. & Messick, D. M. (2004). A conceptual review of decision making in social dilemmas: Applying a logic of appropriateness. *Personality and Social Psychology Review, 8*(3), 281–307.

Welbergen, H. van, Ding, Y., Sattler, K., Pelachaud, C. & Kopp, S. (2015). Real-time visual prosody for interactive virtual agents. In: *International Conference on Intelligent Virtual Agents*. Springer, 139–151.

Williams, L. E. & Bargh, J. A. (2008). Experiencing physical warmth promotes interpersonal warmth. *Science, 322*(5901), 606–607.

Winston, J. S., Strange, B. A., O'Doherty, J. & Dolan, R. J. (2002). Automatic and intentional brain responses during evaluation of trustworthiness of faces. *Nature Neuroscience, 5*(3), 277–283.

Wissen, A. van, Gal, Y., Kamphorst, B. A. & Dignum, M. V. (2012). Human–agent teamwork in dynamic environments. *Computers in Human Behavior, 28*(1), 23–33.

Wojciszke, B. (1994). Multiple meanings of behavior: Construing actions in terms of competence or morality. *Journal of Personality and Social Psychology, 67*(2), 222–232.

— (2005). Morality and competence in person-and self-perception. *European Review of Social Psychology, 16*(1), 155–188.

Wojciszke, B., Bazinska, R. & Jaworski, M. (1998). On the dominance of moral categories in impression formation. *Personality and Social Psychology Bulletin, 24*(12), 1251–1263.

Yaghoubzadeh, R., Kramer, M., Pitsch, K. & Kopp, S. (2013). Virtual agents as daily assistants for elderly or cognitively impaired people. In: *Intelligent Virtual Agents*. Springer International Publishing, 79–91.

Yoshida, W., Seymour, B., Friston, K. J. & Dolan, R. J. (2010). Neural mechanisms of belief inference during cooperative games. *The Journal of Neuroscience, 30*(32), 10744–10751.

Zhou, S., Bickmore, T., Paasche-Orlow, M. & Jack, B. (2014). Agent-user concordance and satisfaction with a virtual hospital discharge nurse. In: *International Conference on Intelligent Virtual Agents*. Springer, 528–541.

Zimbardo, P. G. & Leippe, M. R. (1991). The psychology of attitude change and social influence. New York, NY: McGraw-Hill.

Zlotkin, G. & Rosenschein, J. S. (1989). Negotiation and task sharing among autonomous agents in cooperative domains. In: *International Joint Conference on Artificial Intelligence*. Vol. 11, 912–917.

DECLARATION OF AUTHORSHIP

I, Philipp Kulms, declare that this thesis titled 'Trust in interdependent and task-oriented human–computer cooperation' and the work presented in it are my own original work. Furthermore, I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Bielefeld, January 11, 2018

Philipp Kulms