

TRUSTING HUMANS AND AVATARS: BEHAVIORAL AND NEURAL EVIDENCE

Completed Research Paper

René Riedl

University of Linz
Altenberger Strasse 69,
4040 Linz, Austria
rene.riedl@jku.at

Peter N. C. Mohr

Freie Universität Berlin
Habelschwerdter Allee 45,
14195 Berlin, Germany
peter.mohr@fu-berlin.de

Peter H. Kenning

Zeppelin University Friedrichshafen
Am Seemooser Horn 20,
88045 Friedrichshafen, Germany
peter.kenning@zeppelin-university.de

Fred D. Davis

University of Arkansas
Business Building 204,
Fayetteville, AR 72701, USA
fdavis@walton.uark.edu

Hauke R. Heekeren

Freie Universität Berlin
Habelschwerdter Allee 45,
14195 Berlin, Germany
hauke.heekeren@fu-berlin.de

Abstract

Over the past decade, information technology has dramatically changed the context in which economic transactions take place. Increasingly, transactions are computer-mediated, so that, relative to human-human interactions, human-computer interactions are gaining in relevance. Computer-mediated transactions, and in particular those related to the Internet, increase perceptions of uncertainty. Therefore, trust becomes a crucial factor in the reduction of these perceptions. To investigate this important construct, we studied individual trust behavior and the underlying brain mechanisms through a multi-round trust game. Participants acted in the role of an investor, playing against both humans and avatars. The behavioral results show that participants trusted avatars to a similar degree as they trusted humans. Participants also revealed similarity in learning an interaction partner's trustworthiness, independent of whether the partner was human or avatar. However, the neuroimaging findings revealed differential responses within the brain network that is associated with theory of mind (mentalizing) depending on the interaction partner. Based on these results, the major conclusion of our study is that, in a situation of a computer with human-like characteristics (avatar), trust behavior in human-computer interaction resembles that of human-human interaction. On a deeper neurobiological level, our study reveals that thinking about an interaction partner's trustworthiness activates the mentalizing network more strongly if the trustee is a human rather than an avatar. We discuss implications of these findings for future research.

Keywords: Avatar, agent, brain, fMRI, mentalizing, NeuroIS, theory of mind (TOM), trust

Introduction

Over the past two decades, information technology (IT) has dramatically changed the context in which economic transactions take place (e.g., Kohli and Devaraj 2003). Computers, and in particular the Internet, have made possible new forms of organizational communication (e.g., video conferences), inter-organizational collaborations (e.g., e-business), and distribution of goods and services to the end customer (e.g., e-commerce). In today's society, as computers and the Internet pervade almost every corner of life, the impact of IT on economic behavior is definitive.

As a result of this shift from face-to-face conversation and bricks-and-mortar business to IT-based communication and transactions, traditional human-human interactions in economic exchange are becoming increasingly more computer-mediated, thereby augmenting the relevance of *human-computer interactions*. However, although using IT may lead to notable benefits (e.g., Keeney 1999), it often increases a perception of uncertainty in economic exchange (e.g., Pavlou et al. 2007). In order to reduce this perception, consequently, *trust* becomes a crucial issue (Gefen et al. 2008). Moreover, studying *trust in computers* is gaining importance as the economic trends continually increase in orientation toward IT (International Telecommunication Union 2010).

Trust is typically conceptualized either as a belief or as a behavior in information systems (IS) research (e.g., Gefen et al. 2003; McKnight and Chervany 2001; McKnight et al. 2002), as well as in economics and the management sciences (e.g., Coleman 1990; Fehr 2009a). Because both beliefs and behaviors have origins in neurobiological processes (e.g., Cacioppo et al. 2000), research in the fields of social cognitive neuroscience (e.g., Lieberman 2007) and neuroeconomics (e.g., Sanfey et al. 2006) has begun to investigate the brain processes underlying human trust (e.g., Delgado et al. 2005; King-Casas et al. 2005). Recently, this development has also been extended to IS research (Dimoka 2010; Riedl et al. 2010a).

Many neurobiological investigations in both social cognitive neuroscience and neuroeconomics that normally use functional magnetic resonance imaging (fMRI) have applied the *trust game* (Berg et al. 1995) to study the underlying neural processes of trust. In this game, an individual plays for monetary payoffs, using two different playing strategies, one representing trust behavior (hence, being vulnerable to the actions of the other player) and one representing distrust behavior (hence, lacking vulnerability, but losing an opportunity for greater monetary gain). As well, the game demonstrates trust or distrust behavior based on an individual's inferences about the thoughts and intentions of the other player. The ability to infer the internal states of other actors in order to predict their behavior is known as *theory of mind* (TOM), and the underlying inference process is commonly referred to as *mentalizing* (e.g., Frith and Frith 2003; Premack and Woodruff 1978; Singer 2009). Hence, TOM and mentalizing are fundamental concepts in trust situations because the decision to trust involves thinking about an interaction partner's thoughts and intentions (e.g., Dimoka 2010; Fehr 2009b; Krueger et al. 2007). Moreover, Bente et al. (2008) explain that this thinking involves both cognition-based trust (where an individual judges the competence and dependability of her or his partner) and affect-based trust (where between the partners there is a shared confidence that the other is protective of his or her interests, and that each is genuinely concerned for the welfare of the other).

A detailed look at the game literature as it pertains to trust and mentalizing reveals that most studies are designed so that participants play against human opponents, which simulates economic exchange in traditional bricks-and-mortar business rather than in computer-mediated contexts. However, a few game studies have also investigated participants' brain activation while playing against computer opponents (e.g., McCabe et al. 2001). The findings of these studies indicate that behavior *and* the underlying neural processes typically differ, depending on a participant's interaction partner. Playing against another human more strongly activates several brain regions than does playing against a computer. Most notably, participants playing against computer opponents are normally informed that "[the computer opponent] would play a fixed probabilistic strategy" (McCabe et al. 2001, p. 11832). Moreover, participants lying in the fMRI scanner are often shown a picture of a computer (Rilling et al. 2004). Such designation of the opponent as computer creates a sharp differentiation between humans and computers. In particular, this emphasis on the distinction between humans and computers as players may explain the differences in behavior and in the underlying brain mechanisms, because computers tend to be perceived as mindless, while humans are not.

As a result of recent technological advancements, many computers no longer have the appearance of a mechanistic device and, increasingly, computers today have a human appearance (Al-Natour et al. 2006; Davis et al. 2009; MacDorman et al. 2009; Qiu and Benbasat 2009, 2010). These new human-like forms of computers, *avatars* (Bailenson and Blascovich 2004), are typically designed with faces that closely resemble those of humans (Nowak and Rauh 2005). Attention to a realistic portrayal of a human face is important because the human face has been shown to serve the interpersonal function of allowing one person to predict another's personality traits and behavior (e.g., Ekman 1982; Knutson 1996), which has been demonstrated to apply, particularly, to trustworthiness predictions (Todorov 2008; Winston et al. 2002). Consequently, it is possible that in human collaborations with computers, endowing a computer with a human face alters trust behavior and underlying brain mechanisms. The primary argument supporting this reasoning is that perceiving an object's likeness to humans may result in perceptions of mind, thereby making it possible that objects (such as computers) are recognized as, and are treated as, humans.

To date, it remains unclear whether humans in a trust situation perceive and treat avatars as mechanistic computers, or as humans. To address this open question, we used fMRI and a trust game to investigate how trust behavior (including the learning of an interaction partner's trustworthiness) and the underlying neurobiological mechanisms differ between human-human and human-avatar interactions. Specifically, we conducted a multi-round trust game experiment in which participants played the role of investor, with both humans and avatars in the role of trustee. We contrasted the existing studies, which operationalized computer game partners as mechanical devices, and which informed participants that the computer would play a predefined probabilistic strategy (e.g., McCabe et al. 2001; Rilling et al. 2004). Instead, we used the human-like form of computers, avatars, and informed our participants that the avatars would have a particular character as either relatively trustworthy or relatively untrustworthy.

The remainder of this article is structured as follows: In the second section, we begin with a discussion of human-computer interaction and trust, and continue with a literature review on human-computer interaction in economic games. Based on this theoretical fundament, we describe two behavioral hypotheses (H1a and H1b), as well as one neural hypothesis (H2). The third section describes the research methodology in detail. An outline of the research results in the fourth section precedes a segment that more thoroughly discusses the results and their implications for future research. In the final section, we summarize the contribution of this article, present limitations, and provide a concluding comment.

Literature Review

Human-Computer Interaction and Trust

There is no doubt that trust may be applied to humans (Gambetta 1988; Luhmann 1979), but there is debate over whether it is valid for such technological artifacts as computers to be considered as recipients of trust (e.g., Wang and Benbasat 2005). Those who are of the opinion that computers cannot be recipients of trust argue that machines lack important properties of social actors, such as consciousness or experience of betrayal if trust is breached (Friedman and Millett 1997; Friedman et al. 2000). According to this view, computers cannot be trusted in the literal sense—they can only be relied upon.

In sharp contrast to this view, research grounded in Nass's *Computers Are Social Actors (CASA) Paradigm* has shown that people follow similar social rules and heuristics when they interact with computers as they do when interacting with humans (Nass et al. 1993, 1994; Reeves and Nass 1996). Human social behavior toward computers is interpreted as "ethopoeia," which means that when computers are endowed with personality-like characteristics (e.g., with respect to their appearance or behavior) people will respond to them *as if* they have personalities, despite the fact that these persons will claim they do *not* believe that computers actually have personalities (Nass et al. 1995).

In the following, we briefly discuss two concepts that are commonly recognized as predictors of mutual trust in interaction among humans, namely reciprocity (King-Cases et al. 2005; Ostrom 2003) and empathy (Semmes 1991; Teven and Hanson 2004). These two constructs demonstrate that in their interactions with computers, people often apply established rules of social interaction among humans, in particular those that are closely associated with trust.

Reciprocity is considered to be one of the fundamental characteristics of interaction among humans (Henrich et al. 2004), but it is not considered to be relevant in human-computer interaction. However, experimental evidence challenges this notion. The Fogg and Nass (1997) experiment involved two tasks—a task in which a computer helped a user, and a task in which the user was asked to help a computer. In the first task, participants conducted a series of web searches with a computer; the search results were either very useful or not useful at all. In the second task, participants worked with a computer that was given the task of creating a color palette to match human perceptions of color. The participants were told that by making accurate comparisons of sets of presented colors, they could help the computer to create this palette. Participants were free to choose the number of comparisons to make. Obviously, the more comparisons they made, the more they would help the computer. In one experimental condition, participants performed both tasks on the same computer. Participants in another experimental condition used one computer for the first task, and a second, though identical, computer for the second task.

The findings of the study (Fogg and Nass 1997) are consistent with reciprocity norms; that is, participants who worked with a helpful computer in the first task and then returned to the same computer in the second task performed significantly more work for the computer in the second task, compared to those participants who used two different computers for the two tasks. Moreover, when participants worked with a computer that was not very helpful in the first task, and then returned to the same computer in the second task, they made significantly fewer comparisons than did participants who used different computers.

In the same way as reciprocity, *empathy* (here defined as other-oriented emotions demonstrating that one person cares about another; Singer 2009; Singer and Lamm 2009) is typically considered as a major characteristic in interaction among humans, but not in interactions between humans and computers (Batson et al. 1997). However, experimental evidence supports a view that empathy also plays a crucial role in human-computer interaction.

In one experiment (Brave et al. 2005), participants played a casino-style game of blackjack with human-like computer agents, who were manipulated based on two factors: expression of self-oriented emotion (absent versus present) and expression of empathic—other-oriented—emotion (absent versus present). The computer agent with self-oriented emotion was characterized by facial expressions that were programmed to correspond to that computer-agent's specific blackjack outcomes, while the agent without self-oriented emotion did not respond with appropriate facial expressions. In this case, empathic emotion was manipulated by the agent's reaction to the participant's performance. As an example, in the empathic emotion condition, the agent made a happy facial expression when the participant won, and the agent made a sad facial expression when the participant lost.

The results of this experiment show that empathic, but not self-oriented, emotional expressions of the computer agent do have a positive effect on the perceived trustworthiness of the agent (Brave et al. 2005). This finding is intriguing, because “[a]lthough the agent was a computer-generated, lifeless artifact which lacked ‘genuine feelings,’ the computer user still found the agent more trustworthy when it manifested caring orientations toward the user than when it did not” (Lee and Nass 2010, p. 7).

Considering the empirical evidence that when humans interact with *computers*, they do apply trust-related *social* rules, norms, and expectations (e.g., reciprocity), and taking into account the many other experimental findings of the CASA Paradigm (for reviews see, for example, Nass and Moon 2000, as well as Lee and Nass 2010) and similar research programs (e.g., Nowak and Biocca 2003), it is reasonable to assume that humans often treat computers as social actors rather than mechanistic technological artifacts. Consequently, not only humans, but also computers, can be recipients of trust. Other researchers confirm this view, as in the example of Wang and Benbasat (2005, p. 76, emphasis added), who write:

[W]hile it may at first appear debatable that technological artifacts can be objects of trust, and that people assign human properties to them, evidence from a variety of relevant literature supports this argument. People respond socially to technological artifacts and perceive that they possess human characteristics (e.g., motivation, integrity, and personality). In particular, research findings have demonstrated that components of trust in humans and in technological artifacts do not differ significantly. This indicates that people not only utilize technological artifacts as tools, but also form social and trusting relationships with them.

Human-Computer Interaction in Economic Games

Research originating from behavioral economics and neuroeconomics, which is often based on economic games (for a review, see Krueger et al. 2008), shows that brain activation *and* subsequent behavior (in particular trust behavior, McCabe et al. 2001) often differ, depending on the interaction partner. Humans as playing partners often activate specific brain regions more strongly than do computer partners, thereby elucidating distinct forms of behavior such as accepting or rejecting an offer (Sanfey et al. 2003). The following literature review summarizes important findings on human-computer interaction in economic games.

The design of one study with significant relevance (Gallagher et al. 2002) asked participants to play a computerized version of the game rock-paper-scissors. For the first experimental condition, denoted as *mentalizing*, participants believed they were playing against another human (although, in fact, they were playing against a random selection strategy). In the second condition, denoted as *rule solving*, participants were informed that they were playing against a computer with a predefined algorithm, and that the responses would be based on simple rules related to the participant's previous response (e.g., the computer would select the response that would have beaten the participant's last response). In the third condition, denoted as *random selection*, participants were informed that they were playing against a computer with a random selection strategy, and they were asked to respond randomly, as well. The brain imaging contrast between mentalizing and rule solving revealed a higher activation in the medial prefrontal cortex (MPFC, Brodmann Area, BA 9 and 32; Gallagher et al. 2002, p. 818). Moreover, the contrast between mentalizing and random selection identified higher activation not only in the anterior paracingulate cortex, but also in the right inferior frontal cortex and the cerebellum. No activation was seen in the paracingulate cortex when rule solving and random selection were compared, however. Because it was only in the mentalizing condition that participants believed they were playing against another human (who would have beliefs, desires, and intention to interpret and predict the behavior of others), the bilateral anterior paracingulate cortex is considered to be a crucial mentalizing brain region (e.g., Singer 2009), and seems to be, thereby, a fundamental brain circuit in human-human interaction rather than in human-computer interaction.

In a notable fMRI experiment from this area of research (McCabe et al. 2001), participants played the trust game with both human and computer counterparts, for cash rewards. Participants playing against the computer were told that it would play a fixed probabilistic strategy of (i) 100% trusting behavior in the role of the investor (i.e., the computer would always trust the trustee) and (ii) 75% trustworthiness as the trustee (i.e., the computer would reciprocate trust in three of four cases). Moreover, when the computer moved, it always did so immediately, in order to reduce any likelihood that participants would anthropomorphize the computer responses. As a consequence of this design, participants perceived the computer as a machine rather than a human. The neurobiological results of the fMRI experiment show that participants with the highest cooperation scores reveal significant increases in activation in the MPFC (BA 10; McCabe et al. 2001, p. 11834) during human-human interaction (as compared to human-computer interaction), though within the group of non-cooperators the results did not show significant differences in MPFC activation between the human and computer condition. Therefore, one central conclusion of the McCabe et al. (2001) study is that the MPFC constitutes an "active convergence zone" that binds joint attention to mutual gains with the inhibition of immediate reward gratification, allowing cooperative decisions in human-human interaction, but not in human-computer interaction.

A foundational experiment by Sanfey et al. (2003) applies the ultimatum game (Güth et al. 1982), and participants were made to believe that they were playing against both human partners and a computer. In fact, both the human partners and the computer played the same predetermined strategy. The behavioral results showed that unfair offers of \$1 and \$2 (stake size: \$10) made by human partners were rejected at a significantly higher rate than were those offers made by the computer, suggesting that participants had a stronger emotional reaction to unfair offers from humans than to the same offers from a computer. Among the brain regions showing greater activation for unfair compared with fair offers from human partners were bilateral anterior insula, dorsolateral prefrontal cortex (DLPFC, BA 46), anterior cingulate cortex (ACC, BA 24 and 32), and middle frontal gyrus (BA 9) (see Table S1 in the supporting online material, Sanfey et al. 2003). Moreover, the magnitude of activation was also significantly greater for unfair offers from human partners, as compared to unfair offers from the computer. These results suggest

that it is not the submitted offer itself that determines neural activation and subsequent behavior, but it is, rather, the perception of the opponent as a human or machine that creates the response.

Another fMRI study (Rilling et al. 2004) tested whether inferring the intentions of others would activate mentalizing structures, and whether activated areas would show a response in two different economic games—the ultimatum game and the prisoner’s dilemma game (Axelrod 1984). Interestingly, for both games, the study found activation in the anterior paracingulate cortex, spanning MPFC (BA 9; Rilling et al. 2004, p. 1699) and the rostral ACC (rACC, BA 32; Rilling et al. 2004, p. 1699), which are two classic mentalizing areas (Singer 2009). Both regions responded to decisions from human and computer partners, but showed stronger responses to human partners in both games. The stronger response to human partners is consistent with the behavioral data showing that participants distinguished between human and computer partners, rejecting unfair offers from human partners more often in the ultimatum game and cooperating more frequently with human partners in the prisoner’s dilemma game.

A further fMRI experiment (Rilling et al. 2002) investigated brain activation of participants playing an iterated version of the prisoner’s dilemma game. When subjects were instructed that they were playing the game with a computer rather than with another person, evidence of cooperation was less common (although participants were actually playing against the same strategy in both conditions). The neural results reveal that cooperation with a computer activated regions of the ventromedial/orbital frontal cortex that were also activated by human playing partners. In particular, this activation could be elicited by interactive computer programs, which are programmed to be responsive to the partner’s behavior. However, mutual cooperation with a computer did *not* activate the rACC (BA 32; Rilling et al. 2002, p. 403), which was observed only for human playing partners, suggesting that rACC activation may relate specifically to cooperative social interactions with human partners, rather than with computers.

A recent study focused on strategic reasoning by Coricelli and Nagel (2009) investigated how a player’s mental processing incorporates the thinking processes of others. In the competitive interactive setting of the beauty contest game (Nagel 1995), playing against human opponents (relative to playing against a computer) was shown to activate regions associated with mentalizing, including MPFC, ventromedial prefrontal cortex (VMPFC), and rACC (Coricelli and Nagel 2009, p. 9165). This result suggests that these regions encode the complexity underlying human-human interactions, but not those of human-computer interactions.

One final fMRI study relevant to our research (Krach et al. 2009) investigated participants’ brain activation during an iterated prisoner’s dilemma game. Playing against putative human and computer partners resulted in activity increases in the classic mentalizing network. However, MPFC and ACC activity were significantly more pronounced when participants believed they were playing against the human partner rather than a mindless machine.

















Hypotheses

A recent summarization of the evidence on human trust behavior found in the CASA Paradigm establishes that “[o]ne of the key take-home messages of the CASA research paradigm on trust in computers is that people tend to look for similar qualities of trustworthiness in computers as they do in other people” (Lee and Nass 2010, p. 11), and indicates as a primary conclusion that *trust behavior* seems to function independently from the interaction partner, which can be either another human or a computer. Thus, we have reason to assume that both trust decisions (H1a) and the learning of trustworthiness (H1b) do *not* differ on the basis of whether people interact with humans, or with avatars. Because the decision to trust implies thinking about a trustee’s thoughts and intentions, mentalizing is a significant concept in trust research (e.g., Dimoka 2010; Krueger et al. 2007; Winston et al. 2002). Therefore, in a trust decision task as used in the present study, we would expect activation in mentalizing brain areas, a notion widely supported in the literature (e.g., Fehr 2009b, p. 228; Dimoka 2010, p. 377). However, as outlined by our review of the economic games literature, people generally perceive humans and computers differently. In particular, brain activation patterns differ significantly between human-human and human-computer interactions. A major factor which could explain these differences is that people attribute consciousness to humans, but not to computers (e.g., Friedman and Millett 1997; Friedman et al. 2000). Because computers are typically perceived as mindless, mentalizing should not play a significant role in human-computer interaction, which gives us reason to assume that mentalizing brain regions are more strongly activated when people interact with humans, relative to interacting with avatars (H2).

Research Methodology

Stimulus Selection

In our trust game experiment, participants played in the role of the investor against both humans and avatars in the role of the trustee. The objective of the stimulus selection was, therefore, to identify a number of human and avatar faces for the fMRI study. Ultimately, both the human and avatar groups comprised four faces with a high degree of trustworthiness and four faces with a low degree of trustworthiness. To identify 16 faces appropriate for the experiment, we pre-tested a set of 40 human faces taken from the FACES database (Ebner et al. 2010; 20 faces of each sex) and 40 avatar faces (which we developed ourselves using Reallusion iClone 3.2 EX edition software, San Jose, CA, USA; 20 faces of each sex). In a pre-test, 45 subjects (33 males, 12 females, mean age = 24.07 years, SD = 5.62) rated the trustworthiness of the 80 actors on a 7-point Likert scale (1 = “not at all trustworthy” and 7 = “very trustworthy”).

| | | | |
|---|---|---|---|
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

Note: Rows 1 and 2 = high trustworthiness; rows 3 and 4 = low trustworthiness.

To arrive at the final set of 16 faces, we performed a median split on both the human and the avatar groups. The average trustworthiness of each face was used as the criterion for the median split. We selected four faces out of each group, based on the following criteria: (i) each group is gender-balanced, consisting of two male and two female actors; (ii) the average trustworthiness of the faces from the upper half is significantly higher than the average trustworthiness of the faces from the lower half (humans: $t = 5.39$, $df = 44$, $p < 0.0001$; avatars: $t = 4.52$, $df = 44$, $p < 0.0001$); (iii) the average trustworthiness of the human faces from the upper half does not significantly differ from the average trustworthiness of the avatars from the upper half ($t = 0.32$, $df = 44$, $p = 0.754$); (iv) the average trustworthiness of the human faces from the lower half does not significantly differ from the average trustworthiness of the avatars from the lower half ($t = 0.35$, $df = 44$, $p = 0.725$). The resulting set of 16 faces used in our fMRI study is presented in Table 1 (rows 1 and 2 = high trustworthiness; rows 3 and 4 = low trustworthiness).

Subjects

For the main study, we selected eleven male and eight female subjects (none of whom had participated in the stimulus selection pre-test). All subjects were healthy and reported no history of neurological or

psychiatric diseases. One subject (female) had to be excluded from further analyses, as she indicated after the fMRI session that she was afraid while in the fMRI machine, and as a result was not really thinking of the assigned task. Therefore, the sample size underlying our data analysis is $N = 18$. All participants were familiar with the Internet, and had been using it for many years (mean = 10.77, SD = 3.40, min = 4, max = 18). By design, our investigation is focused on experienced computer and Internet users rather than novices. All participants were paid for their participation, and gave written informed consent. The study was approved by the Freiburg Ethics Commission International (FECI), Germany.

We were specific in the age range of participants that we sought, taking note that trust increases almost linearly from early childhood, but stays relatively constant within different age groups (Philips and Stanton 2004; Sutter and Kocher 2007). Hence, to avoid confounding effects due to age differences, we selected subjects from the relatively narrow age group of 25 to 40, rather than using undergraduates or a blend of people from different age groups (mean age = 31.83 years, SD = 4.14, min = 26, max = 40).

Another important trait that we assessed for all participants was the general level of trust (i.e., trust propensity), which we measured by a 25-item questionnaire (Rotter 1967). At a maximum, each subject could score 125 points (high trust), and at a minimum, 25 points (low trust). Answers from all participants revealed general trust scores within the normal range of healthy subjects (mean = 79.11, SD = 5.60, min = 63, max = 88). Moreover, the results are in line with the general trust levels reported in similar studies (e.g., Riedl et al. 2010a). We therefore included all participants in further analyses.

Because humans tend to assess others on the basis of racial and cultural traits, we note that all participants were white. Research (Phelps et al. 2000) indicates that both activation of the amygdala (a subcortical part of the brain which plays a prominent role in emotion processing, among other functions), as well as the behavioral responses of white participants to unfamiliar black-versus-white faces, reflect cultural evaluations of social groups. Thus, it was essential for the validity of our results that we matched the race of the participants and trustees (see Table 1).

Another trait that creates a differential is handedness, so we considered that as well. Research (e.g., Cuzzocreo et al. 2009; Schachter 2000) shows that handedness is related to brain anatomy and functionality, as well as cognitive abilities (e.g., memory). Hence, it is important in a brain imaging study such as the present one that the handedness of participants be held constant. Because one of the remarkable features of motor control in humans is that more than 90% of the population is more skilful with the right hand (Sun and Walsh 2006), we have chosen only right-handed people for the experiment. Our results, therefore, directly pertain to the vast majority of the human population.

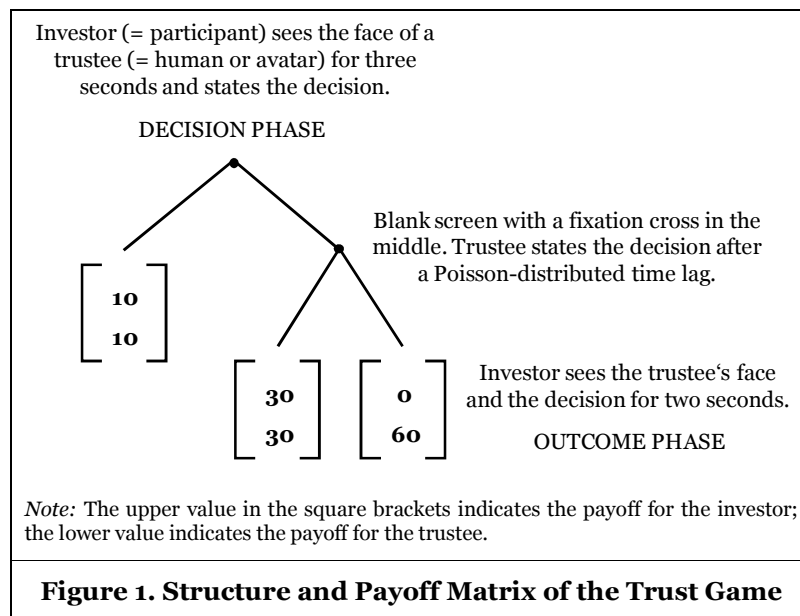
Experimental Procedure and Stimulus Presentation

We used a multi-round trust game, an adjusted version of the original trust game (Berg et al. 1995), in which the investor has an initial endowment. First, the investor decides whether to keep her or his endowment, or to send it to the trustee. Then the trustee observes the investor's action and decides whether to keep the amount received, or to share it with the investor. The experimenter multiplies the investor's transfer by some factor, so that both players are advantaged, collectively, if the investor transfers money and the trustee sends back a part of it. As a behavioral measure for trust, we used the decision of the investor to send money, and as a behavioral measure for trustworthiness, we used the trustee's decision of whether or not to return money (Fehr 2009a).

In our experiment, participants played in the role of the investor against (i) humans and (ii) avatars (both playing in the role of the trustee, see Figure 1). Our game, therefore, mimics a typical interaction in a buyer-seller-relationship, both in bricks-and-mortar business (human condition), as well as in computer and online environments (avatar condition). Participants were told in advance that their playing partners (i.e., the trustees) would not be responsive to their playing strategies. We did stress, however, that each trustee has a specific character, which determines his or her trustworthiness. Half of the trustees, both humans and avatars, were predetermined by the experimenters to be relatively trustworthy, whereas the other half were predetermined as relatively untrustworthy. Trustworthy actors returned €30 in seven out of ten rounds, whereas untrustworthy actors returned €30 in only three rounds. Apart from their facial appearance, participants had no information regarding the trustees.

In each round of the game, participants were asked whether they wanted to keep their initial endowment of €10 (\approx \$13), or whether they wanted to give it to the trustee, whose face was presented to them. In the

case of giving the €10 to the trustee, the amount was multiplied by six (resulting in €60), which the trustee could then either keep, or split (i.e., return €30 to the participant). Participants played ten rounds of the game with each trustee, with three seconds in each round to make the investment decision (= DECISION PHASE, see Figure 1); note that decision times of one to three seconds are sufficient to make economic decisions in an fMRI environment (e.g., O’Doherty et al. 2004; Winston et al. 2002). After a variable time (varied based on a Poisson distribution) in which a blank screen with a fixation cross in the middle was presented, the trustee’s face and the decision was visually presented for two seconds (= OUTCOME PHASE, see Figure 1). Before the first round, after the fifth round, and after the tenth round, participants were asked to rate the trustworthiness of the trustee, which was operationalized as the probability that the trustee, in case of being given money, would behave in a trustworthy manner (i.e., returning €30). It is important to note that the fMRI (hemodynamic) response to an event (e.g., perception of the face of a trustee) begins after a delay of one to two seconds, peaks approximately five to six seconds following event onset, and returns to baseline after about fourteen to sixteen seconds (Payne and Venkatraman 2011). Because hemodynamic responses to multiple subsequent events typically sum in a linear manner (Huettel and McCarthy 2000), it is possible to investigate neural responses to multiple events that are presented in close temporal proximity.



We used the program Presentation (Neurobehavioral Systems, Albany, CA, USA) to present the stimuli in the MRI-scanner and to record the responses on a Notebook with Windows XP (Microsoft, Seattle, USA) as the operating system. Visual stimuli were presented using video goggles.

Data Collection and Analysis

MRI data were acquired using a 3T Siemens Tim Trio (Erlangen, Germany) MRI scanner. fMRI is a noninvasive neuroscientific technique with relatively good spatial and temporal resolution (Huettel et al. 2009). It takes advantage of the blood oxygen level dependent (BOLD) effect for estimating the neural activity that corresponds with experimental conditions. We acquired 2 runs of 690 functional T2*-weighted echoplanar images (EPI) [TR, 2 s; echo time (TE), 40 ms; flip angle, 90°; field of view, 256 mm; matrix, 64 x 64 mm; 26 axial slices approximately parallel to the bicommissural plane; slice thickness, 4 mm]. In addition, for registration purposes, a high-resolution T1-weighted structural image (MPRAGE) was acquired from each participant [TR, 20 ms; TE, 5 ms; flip angle, 30°; 179 sagittal slices; voxel size, 1 x 1 x 1 mm]. Initial analysis was performed using the FSL toolbox from the Oxford Centre for fMRI of the Brain (www.fmrib.ox.ac.uk/fsl). The image time-course was first realigned to compensate for small head movements (Jenkinson et al. 2002). Data were spatially smoothed using an 8 mm full-width-half-

maximum Gaussian kernel and were temporally filtered using a high-pass temporal filter (with $\sigma=100s$). Registration was conducted through a 2-step procedure, whereby EPI images were first registered to the MPRAGE structural image and then to standard MNI (Montreal Neurological Institute) space (MNI152_T1_2mm_brain), using 7 parameters for the first registration step and 12 parameters for the second (Jenkinson and Smith 2001). Statistical analyses were performed in native space, with the statistical maps normalized to standard space prior to higher-level analysis.

Statistical analysis of functional data was performed using a multi-level approach implementing a mixed-effects model treating participants as a random effect. This was initially performed separately for each participant's concatenated runs. Regressors-of-interest were created by convolving a rectangular function representing stimulus duration times with a canonical (double-gamma) hemodynamic response function. Time-series statistical analysis was carried out using FILM (FMRIB's Improved Linear Model) with local autocorrelation correction (Woolrich et al. 2001).

The functional analysis was based on two regressors-of-interest and four regressors-of-no-interest. Two binary regressors modeled the decision phase. The first regressor represents the decision in the human condition, whereas the second regressor models the decision in the avatar condition. The durations of the two regressors correspond to the decision time, which was three seconds in all cases. The final four regressors-of-no-interest modeled the trustworthiness ratings for humans and avatars, as well as the amount of stated trustworthiness for both conditions. On the group level (second level of analysis), we integrated the results from the single subject level (first level), again applying a general linear model. One binary regressor modeled a constant effect of all first-level parameter estimates on the group level.

Results

Behavioral Results

Analyzing whether or not participants differ in their decisions to trust (i.e., to invest their initial endowment of €10) when playing against humans or avatars, we found that participants showed considerable trust behavior (see Table 2), independent of playing against humans or avatars (average trust in the human condition: in 52 of 80 games; average trust in the avatar condition: in 51.39 of 80 games). There was no significant difference between the number of decisions to trust in the human and avatar conditions ($t = 0.4455$, $df = 17$, $p = 0.66$). This result supports H1a.

To better understand participants' trust behavior, we modeled the perceived trustworthiness of the trustees on a round-by-round basis. The underlying assumption of this approach is that perceived trustworthiness of a trustee is based on the facial appearance before the first round begins, and is adjusted based on its behavioral trustworthiness during the subsequent rounds of the game.

We assume that perceived trustworthiness is reflected in the subjective probability that a trustee will be trustworthy (i.e., will return €30) when a participant invests the initial endowment of €10. Moreover, a participant typically selects the gaming strategy that offers a higher expected value. The expected value of keeping the initial endowment of €10 is €10, because the probability for this gain is 100% in our game (see Figure 1, left path). The expected value of investing the endowment of €10 is equal to the product of a trustee's perceived trustworthiness and the possible payoff of €30. Hence, the trust decision only depends on the perceived trustworthiness (see Figure 1, right path).

During the ten rounds of interaction with a specific trustee, trustworthiness is expected to be updated on the basis of that trustee's behavior. In this article, we model the updating process with a *reinforcement learning model* (Behrens et al. 2008, 2009; Rescorla and Wagner 1972; Sutton and Barto 1998). Such a model generally assumes that after the decision has been made for one alternative, a received reward $R(t)$ at time t is compared to an expected value $EV(t)$, with the deviation d formalized as prediction error PE : $d(t) = R(t) - EV(t)$. A reinforcement learning model assumes that learning is driven by these deviations; hence, a PE is used to update $EV(t)$, allowing the optimization of reward predictions. The influence of a specific PE on $EV(t)$ regarding the next trust decision is determined by the learning rate.

Formally, a *reinforcement learning model* is defined as:

$$EV(t) = EV(t - 1) + \alpha \cdot d(t - 1) \quad (1).$$

From $EV(t)$, it is possible to calculate the subjective probability of trustworthy behavior at time t by dividing $EV(t)$ by €30. In our trust game, a model with a constant learning rate (see *formula 1*) would assume that the perceived trustworthiness is updated equally for trustworthy and untrustworthy behavior.

We fitted the reinforcement learning model for the interaction with humans and avatars, and instructed each participant to rate the trustworthiness of each trustee before the first round of the trust game began (based solely on facial appearance). We used these initial ratings as the starting points of the reinforcement learning process. The free parameter α was fitted by minimizing the sum of squared differences between model predictions and a participant's trustworthiness ratings after the fifth and tenth rounds.

| Table 2. Behavioral Results | |
|--|-----------------|
| | Mean (SD) |
| Decisions to Trust | |
| Decisions to Trust (Human) ^a | 52.00 (11.55) |
| Decisions to Trust (Avatar) ^a | 51.39 (13.57) |
| Trustworthiness Learning Rate | |
| Sum of Squared Differences (Constant Learning Rate, Human) | 0.2343 (0.1322) |
| Sum of Squared Differences (Constant Learning Rate, Avatar) | 0.2773 (0.1950) |
| Alpha (Human) | 0.2139 (0.0825) |
| Alpha (Avatar) | 0.2122 (0.0916) |
| Correct Model Predictions (Constant Learning Rate, Human) | 73.61% (9.40%) |
| Correct Model Predictions (Constant Learning Rate, Avatar) | 71.51% (13.92%) |
| <i>Note:</i> ^a Total number of rounds played = 160 (humans = 80 rounds, avatars = 80 rounds). | |

We further investigated whether learning rates were significantly different between the human and avatar conditions (see Table 2 for descriptive statistics). As expected, we found no significant differences ($t = 0.1510$, $df = 17$, $p = 0.8820$). This result supports H1b.

Assuming a deterministic decision strategy stating that when two alternatives offer a higher-than-expected value, people always choose the one that offers a higher expected value, the reinforcement learning model, on average, correctly predicts 73.61% (71.51 %) of the trust decisions in the human (avatar) condition. Importantly, the predictive power of this model is significantly higher than chance level (human: $t = 10.6540$, $df = 17$, $p < 0.001$; avatar: $t = 6.870$, $df = 17$, $p < 0.001$).

In the two hypotheses that pertain to the behavioral level of analysis, we predicted that both trust decisions (H1a) and the learning of trustworthiness (H1b) do *not* differ based on whether people interact with humans, or with avatars. Thus, in order to find support for these two hypotheses, we had to test null hypotheses (i.e., $H_0: \text{Condition}_{\text{Human}} = \text{Condition}_{\text{Avatar}}$). Statistical power is defined as the probability of rejecting H_0 when it is false; major factors that influence this power are (i) effect size, (ii) level of significance, and (iii) sample size (Desmond and Glover 2002).

When compared to traditional behavioral research (both in the behavioral sciences and the IS discipline), sample size is relatively small in neuroscience studies. A recent review of papers, including studies in highly prestigious journals such as *Neuron*, *Science*, and *Nature*, found that, for example, the average sample size is $N = 18$ in neuroscience studies (Lieberman et al. 2009, p. 301). Moreover, we investigated the sample sizes of the fMRI studies published in IS outlets: $N = 6$ (Dimoka and Davis 2008), $N = 15$

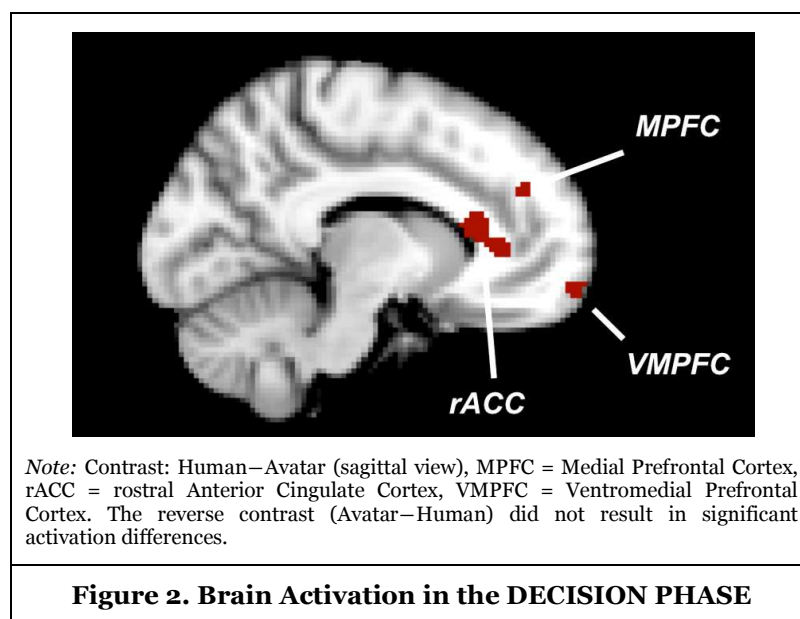
(Dimoka 2010), $N = 20$ (Riedl et al. 2010a), and $N = 24$ (Benbasat et al. 2010). Based on these inquiries, we conclude that our sample size is similar to those reported in other fMRI studies, both in other neuroscience disciplines (e.g., neuropsychology, social neuroscience, neuroeconomics, neuromarketing) and IS research.

Despite this fact, however, it is impossible to rule out the eventuality that a larger sample size in our study could affect the results. In particular, a larger sample size could lead to the identification of statistically significant differences in both trust decisions (H1a) and the learning of trustworthiness (H1b), thereby disproving our hypothesized similarity in trust behavior toward humans and avatars.

On the basis of additional statistical tests, we demonstrate the robustness of our results related to H1a and H1b. For H1a, the observed effect size is very small (*Cohen's d* = 0.1085), indicating that there is indeed negligible difference between the human and avatar conditions. Further support for H1a can be provided when we assume that all decisions of all subjects are independent. Because the amount of payment every subject received was based on the outcome of one randomly selected round (out of all rounds played during the experiment), this assumption is not inappropriate. From this assumption, a binomial test further validates our findings, as it reveals no significant difference between the number of decisions to trust in the human and avatar conditions (human = 936 [out of 1,440 decisions throughout the entire experiment: 8 human opponents \times 10 rounds \times 18 subjects], avatar = 925 [1,440 decisions], $p = 0.39$). Importantly, this binomial test has the power to detect even small effects (*Hedges g* < 0.05). The findings of these additional statistical tests allow us to confirm that our data strongly support H1a. Finally, for H1b, the observed effect is even smaller than the one for H1a (*Cohen's d* = 0.0355), indicating that there also is negligible difference between the learning of human and avatar trustworthiness. Thus, we also argue that our data strongly support H1b.

Neuroimaging Results

In the present study, we focus on the investigation of brain activation during the DECISION PHASE and *not* during the OUTCOME PHASE (see Figure 1). In the decision phase, participants were able to see the trustees' faces, and had to decide whether or not to send their initial endowment of €10. Among activity changes in other brain areas, we found that activity in the MPFC, rACC, and VMPFC was significantly higher in the human condition compared to the avatar condition ($z > 3.09$; cluster size > 25; see Figure 2). Previous studies show these three brain regions to be associated with mentalizing (for reviews, see Amodio and Frith 2006, as well as Singer 2009). This result supports H2.



Discussion and Future Research

In this study, we investigated human trust behavior and the underlying neurobiological mechanisms, measuring by means of fMRI in a multi-round trust game. Participants taking on the role of an investor played against humans and avatars (i.e., human-like virtual agents) playing in the role of the trustee. Participants exhibited similar trust behavior when they played against humans and against avatars. Moreover, participants' learning rates regarding the trustworthiness of humans and avatars were also similar. At the neurobiological level, however, we observed significant differences in brain activation during the DECISION PHASE (in which the task was to invest the initial endowment of €10 or not) when contrasting the human versus avatar condition. Specifically, we found higher activation in a brain network associated with mentalizing, namely MPFC, rACC, and VMPFC. These three brain regions correspond approximately to BA 9, 10, 24, and 32, and they have been summarized under the label arMFC (anterior region of the rostral medial frontal cortex; Amodio and Frith 2006). In this regard, our brain imaging results are in line with two recent fMRI studies from the field of robotics.

The first study (Krach et al. 2008) investigated whether an increase of human-likeness of the interaction partners (computer partner < functional robot < anthropomorphic robot < human partner) modulates people's mentalizing activity. Using the prisoner's dilemma game, the study was designed so that during the experiment participants were unaware that they always played against a random sequence. The brain imaging data revealed a linear increase in cortical activity in the MFC, in correspondence with the increase in human-likeness of the interaction partner. These results, together with the findings of the present study, strongly suggest that humans seem to have a propensity to build a model of another's 'mind' in accordance with its perceived human-likeness.

In the second study (Kircher et al. 2009), the objective was to assess an interactive task involving social partner feedback that can be used to infer intent, in order to determine the degree to which the task would engage the MPFC. Participants' brains were scanned as they played a special version of the prisoner's dilemma game with partners outside the scanner, alleged to be human and computer partners. The MPFC was activated when participants played both human and computer partners. However, the direct contrast revealed significantly stronger activation during the human-human interaction. These findings suggest a link between activation in the MPFC and the interaction partner (human or computer) played in a mentalizing task. Hence, these findings provide further support for the brain activation patterns found in our study.

Altogether, our behavioral results suggest that trust in human-like computers (avatars) resembles trust in humans (H1a and H1b). On a deeper neurobiological level, however, the human brain distinguishes between humans and avatars (H2). People seem to attribute the concept of a mind to humans, but not to the same extent for avatars. We conclude, therefore, that our brain imaging study sheds light on the differences in the mentalizing process in a trust situation. Because many cognitive processes underlying human behavior are not accessible to consciousness (Lieberman 2007), and thus not open to introspection and self-reporting (e.g., Dimoka et al. 2010), our investigation adds an expanded view to the existing IS trust literature.

Research in the IS discipline is deeply rooted in the behavioral sciences (e.g., Frank et al. 2008). Consequently, the differences identified in the present study—differences in brain mechanisms, but *not* in a specific form of subsequent behavior—might be perceived as irrelevant, particularly due to the putative lack of behavioral implication. Specifically, in the present experiment we could not identify statistically significant differences in trust behavior toward humans and avatars. However, research indicates that a difference in brain activation can be expected to have some form of behavioral consequence (e.g., Kolb and Whishaw 2009, chapter 2; in evolutionary psychology, a similar concept has been proven to advance IS theorizing, Kock 2004, 2005, 2009). A lack of difference in one specific form of behavior (here, trust) implies that behavioral variance is likely influenced by multiple mental processes. Yoon et al. (2006, p. 38, emphasis added) present findings from a consumer neuroscience study that confirm this view:

A fair criticism of the sort of differences documented ... is that each could well be, to invoke a line beloved of philosophy professors, "a distinction without a difference." For example, can one brain region process a particular sort of stimulus (e.g., human traits) while a second, physically separate region processes a very different one (brands), yet with little discernible difference in substantive outcome measures? It is impossible to rule this out

entirely. Nonetheless, specialization of the cortex has long been recognized at the gross level of the cortical lobes as well as the cellular level, and recent findings of functional specialization in focal regions ... further push the envelope on cortical specialization. As such, the likelihood that brand judgments lie in one area and human judgments in an entirely separate one, but that they are otherwise fundamentally alike, can only be viewed as remote. Rather, it would appear that the behavioral measures employed in prior studies have not, as of yet, been able to ferret out these distinctions, which seem to be separate processes subserved by different brain regions.

Drawing upon this statement, as well as on similar views held by neuroeconomists (e.g., Clithero et al. 2008), we argue that future research should seek to identify possible behavioral consequences that may result from the significant higher activation in the mentalizing brain areas in the human condition compared to the avatar condition. One promising strategy to identify relevant behavioral consequences is to proceed along the causal chain under investigation. We exemplify this based on an example (Figure 3).

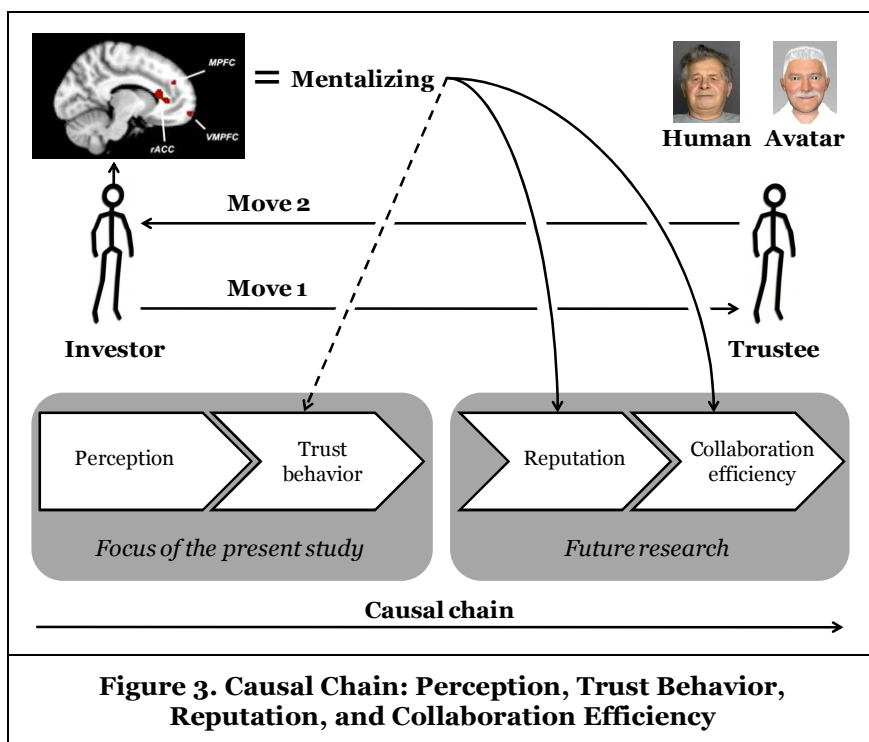


Figure 3 illustrates the major focus of the present study (left), as well as possible directions for future research (right). In essence, once an investor perceives stimuli related to the trustee (e.g., facial appearance), he or she has information on which a trust decision can be based (move 1). Next, the trustee can either keep or split the amount (move 2), and this behavioral information, in turn, can be used by an investor to learn a trustee's trustworthiness.

In the present study, we found that differences in brain activation (notably, the significantly higher activation in mentalizing areas during interaction with humans rather than avatars) do *not* lead to differences in trust behavior (see the dashed arrow in Figure 3). Because differences in brain activation generally have behavioral consequences (e.g., Kolb and Whishaw 2009), future research could look for corresponding outcome variables by proceeding along the causal chain under investigation. Specifically, the causal chain illustrated in Figure 3 shows that trust is an antecedent of reputation (e.g., King-Cases et al. 2005), which in turn is an antecedent of collaboration efficiency (e.g., Nowak and Sigmund 2005; Raub and Weesie 1990).

Reputation among humans is a crucial phenomenon that plays a significant role in social interactions (Fehr 2004), and in particular, in trust situations (King-Cases et al. 2005). Research suggests that one force driving humans to trust and cooperate is the desire to build a good reputation (Frith and Frith 2010). Most people strive to be perceived positively by other individuals, and humans typically care what others think of them. Hence, reputation building implies mentalizing (Frith and Frith 2010, see the arrow in Figure 3).

Considering this link between reputation building and mentalizing, we hypothesize that the activation difference in the arMFC between the human and avatar condition could be caused by the participants' stronger desire build a reputation with humans than with avatars. The participants in our experiment could have reflected more intently on their gaming partners' thoughts about the trust decisions if the partner was a human rather than an avatar. Questions our participants might have considered could be, for example, "What does the gaming partner think when I do not trust him/her, and thus do not send the €10?" and "What does the gaming partner think when I trust, and thus send the €10?" Importantly, positive correlations between reputation and efficiency in social and economic interactions are reported in the literature (e.g., Nowak and Sigmund 2005; Raub and Weesie 1990). Consequently, considering the causal chain shown in Figure 3 (trust behavior → reputation → collaboration efficiency), together with our brain imaging findings, we hypothesize that interactions with avatars, in contrast to interactions with humans, could result in decreased levels of collaboration effectiveness (due to decreased levels of activation in mentalizing brain areas; see the arrow in Figure 3). Future research could test this hypothesis, with the possibility of identifying an important economic consequence of the brain activation findings from the present study.

In addition to the avenue for extended research illustrated in Figure 3, another important question could be addressed by adding a further experimental condition. In the present study participants had no information about whether the avatars actually represent real humans or not. The subjects were told that the trustees would not be responsive to their playing strategy, but they were also told that each trustee has a specific character, which determines his or her trustworthiness. Thus, one further useful manipulation could be adding an experimental condition in which participants are told that the avatars represent real humans. Such a manipulation could affect both activation in mentalizing brain areas and trust behavior, which would have far-reaching practical implications. If this condition produced the same behavioral and neural results as our study, understanding the implications of such a finding would present intriguing issues that deserve consideration.

The value of knowing that a human-avatar interaction would produce responses similar to those when the human interacts with another human relies, in one instance, on an application of media richness theory (Daft and Lengel 1986; Daft et al. 1987), which is a framework that describes a communication medium's ability to reproduce the information sent over it. A video conference, for example, allows for a real-time visual and auditory experience in a way that e-mail does not. Hence, video conferencing is considered to be "richer" than e-mail, particularly because its degree of *social presence* is higher (Rice 1993). In essence, media richness theory states that a sender should use the richest possible medium to communicate a desired message, thereby increasing information richness—the ability of information to change understanding within a time interval.

Real-world practicalities often require communication through media that are less rich than what might be available. For example, although it might be theoretically beneficial to collaborate with a colleague or a business partner in a face-to-face meeting, economic considerations (e.g., travel costs) might preclude this type of communication. Hence, less expensive but still acceptable media have to be selected. If the human-avatar communication produced results as we have described, use of avatars (which represent real humans rather than computer-controlled virtual assistants, just as in virtual worlds such as *Second Life*), as an alternative to video conferencing, could be effectively used in organizational contexts.

Avatars have, in fact, been recently suggested as a new type of communication in business organizations, thereby complementing existing media such as face-to-face meetings, text chat, e-mail, telephone, or videoconferencing (Nowak and Rauh 2005). In this context, a recent study (Bente et al. 2008) investigated the influence of avatars on social presence, interpersonal trust, perceived communication quality, nonverbal behavior, and visual attention in network-based collaborations. A real-time communication window including an avatar-based interface was integrated into a shared collaborative workspace, and this medium was compared to text chat, audio, and audio-video. The results show that

differences exist between text chat and the other three modalities in perceived intimacy, co-presence, and emotionally-based trust. Moreover, a detailed analysis of nonverbal activity and visual attention revealed similarities between the video and avatar modalities, both demonstrating higher levels of exposure to the virtual other and visual attention. Altogether, these *behavioral* results suggest that avatar-based communication can be as rich as video communication.

However, bearing our research results in mind, it is likely that the behavioral outcome similarities between avatar-based and video communication are accompanied by differences on the neurobiological level. In particular, building a reputation in the communication process is more likely to happen during the interaction via video (because real humans are visible) when compared to interaction via avatars. This difference, in turn, may influence important outcome variables on the behavioral level, particularly efficiency in social and economic interactions (see Figure 3). Thus, our neuroscientific focus contributes to a more complete understanding of possible limitations of avatar-based communication in business organizations.

Another variation in future research might be to vary the degree of human-likeness of the avatars. The avatars that we used in the present study (see Table 1) have a medium degree of human-likeness (i.e., they are not completely simplistic cartoon-like characters, but also are not photorealistic portrayals). It would be reasonable to assume that avatars with increasing human-likeness should trigger brain activation patterns monotonically more similar to that of real humans. Research suggests, however, that such an assumption would be too simplistic, because as avatars approach photorealistic perfection but do not fully accomplish it, they cause humans to feel negative emotions (e.g., Davidson 2003) that neurologically resemble distrust reactions (e.g., Dimoka 2010). This effect is referred to as the *uncanny valley effect* (e.g., Geller 2008; Mori 1970). Considering this, future research could test the behavioral and neural effects of avatars with varying degrees of human-likeness in trust game settings. One prediction based on the uncanny valley effect is that there is no linear relationship between human-likeness and (i) trust behavior and (ii) brain trust (MacDorman 2009). Specifically, we hypothesize that “uncanny valley avatars,” in contrast to the other avatars, lead to the strongest distrust perceptions, both on behavioral and neural levels.

Contribution, Limitations, and a Concluding Note

In the following, we summarize the contribution of the present article. Our review of the social neuroscience and neuroeconomics literature on game playing (e.g., the trust game) revealed that in most studies participants play against human opponents, thereby simulating economic interaction in traditional bricks-and-mortar business and non-computer-mediated environments. Due to the increasing importance of computer-mediated communication and transactions, human-computer interactions are becoming increasingly relevant. A few game studies, therefore, have also begun to investigate participants’ brain activation while playing against computer opponents (e.g., McCabe et al. 2001). Altogether, the results of these studies show that trust behavior *and* the underlying neural mentalizing processes differ depending on the interaction partner.

However, a detailed look at the experimental tasks of these studies reveals that participants playing against computer opponents were typically informed that the computers would play a fixed probabilistic strategy, and participants in the fMRI scanner usually were shown a simple picture of a computer. When presented in this objectified format, computers were likely to be perceived as mindless. Obviously, such a clear differentiation between computers and humans may explain the observed differences in behavior and the underlying brain mechanisms.

Recently, technological advancements have made possible the development of new human-like forms of computers, referred to as *avatars*. No published study, to the best of our knowledge, has investigated whether avatars are perceived and treated as mechanistic computers *or* as humans in a trust situation. To close this research gap, we addressed this open question on the basis of fMRI. We conducted a multi-round trust game experiment in which participants played in the role of the investor against both humans and avatars, both playing in the role of the trustee.

Our behavioral results show no differences between the human and avatar condition. Specifically, participants trusted avatars to a similar degree as they trusted humans; the learning of an interaction partner’s trustworthiness was similar, as well (see Table 2). However, our neurobiological findings show

significant differences in brain activation in a mentalizing network within the arMFC, spanning MPFC, rACC, and VMPFC (see Figure 2). Consequently, our study demonstrates that trust behavior in human-computer interaction resembles that in human-human interaction, if a computer is presented as an avatar and not as a mechanistic artifact. However, on a more intensive neurobiological level, the human brain still distinguishes between humans and avatars.

In the IS discipline, research is typically focused on the behavioral level. However, as the present study investigating trust in humans and avatars indicates, similar patterns of behavior may be associated with different underlying neurobiological processes. One general implication developed from our research is that IS research can substantially benefit from neuroscience (e.g., Benbasat et al. 2010; Dimoka et al. 2007, 2010, 2011; Riedl et al. 2010b), because the use of brain imaging makes possible the identification of mental processes that may be difficult to capture based on behavioral data alone (e.g., mentalizing in a trust situation). Thus, neuroscientific approaches help understand human cognition and affect, as well as subsequent IT-related behavior. However, because differences in brain activation typically have behavioral consequences (e.g., Kolb and Whishaw 2009), we argued that future research could look for such important outcome variables by proceeding along the causal chain under investigation; we discussed possible candidates such as reputation and collaboration efficiency (see Figure 3), with both having trust as an important antecedent (e.g., King-Cases et al. 2005; Nowak and Sigmund 2005).

As is common in scientific research, the present study has limitations that should be taken into account. First, the interpretation of our empirical findings is based on a simple game-playing task in a controlled laboratory environment. Second, during the experiment, participants were required to lie still and were restrained with pads to prevent motion during measurement sessions. Third, the present study investigates the potential of static pictures of humans and avatars in order to activate mentalizing brain mechanisms. Indeed, we found activation in a network spanning the arMFC. However, inferring another actor's thoughts and intentions is not based solely on the processing of static information. Motion (e.g., gestures), as well, has been demonstrated to enable inferences regarding another actor's mind (e.g., Frith and Frith 2010). Future investigations could replicate our study using non-static stimulus material. Considering recent findings regarding the positive effects of both visual and behavioral realism of avatars on outcome variables such as affect-based trustworthiness (e.g., Bente et al. 2008; Groom et al. 2009; Qiu and Benbasat 2005), we hypothesize that an increasing degree of an avatar's human-likeness induced by non-static information could reduce the neurobiological differences between humans and avatars in mentalizing circuits (Morris et al. 2005). In this context, one study (Gazzola et al. 2007) already found that the mirror neuron system—which transforms observed actions into the neural representations of these actions—responds to both human and robotic actions, with no significant differences between these two agents. Despite these limitations, however, we believe that the present study contributes to a better understanding of trust and mentalizing in IS research.

A recent paper by Lieberman and Eisenberger (2009) presents neuroscience evidence showing that “responses to [complex social] events rely on much of the same neural circuitry that underlies the simplest physical pains and pleasures” (p. 890), and “the brain may treat abstract social experiences and concrete physical experiences as more similar than is generally assumed” (p. 891). Trust is a remarkably complex social phenomenon (Gambetta 1988; Luhmann 1979). Given that physical needs (e.g., food and drink) and perilous threats (e.g., a dangerous animal) are more critical to survival than social phenomena (e.g., trust), the question arises as to why the brain would have evolved to treat them as motivationally similar. One popular explanation based on evolutionary theory (Lieberman and Eisenberger 2009) states that for a great part of human history, being part of a social group promoted survival. The division of basic human activities (e.g., one group member takes care of food acquisition, another provides protection from adversarial groups, and another cares for offspring) has resulted in a greater certainty of survival, while social groups without this structure have not achieved the same result. Most notably, the division of activities among group members implies cooperative behavior, which in turn is strongly based on trust and mentalizing. Bearing these explanations in mind, it becomes clear that trust and mentalizing, in particular, are intriguing social phenomena that have evolved in order to secure the survival of humankind during the past thousands of years. Though computer-mediated forms of communication and cooperation, particularly via avatars, have only recently begun to emerge, they promise to bring revolutionary changes in the organization and interaction of societies. It will be rewarding to see what insight future research in the IS discipline will reveal regarding the behavioral and neurobiological mechanisms underlying social and economic interaction in a computerized world.

Acknowledgements

We would like to thank the track chairs Hock Hai Teo and Hock Chuan Chan, as well as the anonymous associate editor and the reviewers for their excellent work in providing guidance on ways to improve the article. Moreover, we are grateful to Cornelia Huber for her support in developing the avatars. We also appreciate the generous and visionary support of Novomind AG, Germany (www.novomind.com). Also, we appreciate the generous support of Schindler Parent Meersburg who, through the Schindler Parent Distinguished Guest Lecturer for Marketing, supported René Riedl's work on NeuroIS and consumer neuroscience projects at Zeppelin University Friedrichshafen. Peter N. C. Mohr's work on this project was supported by the "Collaborative Research Center 649: Economic Risk" funded by the German Research Foundation (DFG). Fred D. Davis's work on this project was supported by Sogang Business School's World Class University Program (R31-20002) funded by the Korea Research Foundation. Finally, we thank Deborah C. Nester for proof-reading.

References

- Al-Natour, S., Benbasat, I., and Cenfetelli, R. T. 2006. "The Role of Design Characteristics in Shaping Perceptions of Similarity: The Case of Online Shopping Assistants," *Journal of the Association for Information Systems* (7), pp. 821-861.
- Amodio, D. M., and Frith, C. D. 2006. "Meeting of Minds: The Medial Frontal Cortex and Social Cognition," *Nature Reviews Neuroscience* (7), pp. 268-277.
- Axelrod, R. 1984. *The Evolution of Cooperation*, Basic Books, New York.
- Bailenson, J. N., and Blascovich, J. 2004. "Avatars," in *Encyclopedia of Human-Computer Interaction*, W. S. Berkshire (ed.), Great Barrington, MA, Berkshire Publishing Group, pp. 64-68.
- Batson, C. D., Early, S., and Salvarani, G. 1997. "Perspective Taking: Imagining How Another Feels Versus Imagining How You Would Feel," *Personality and Social Psychology Bulletin* (23), pp. 751-758.
- Behrens, T. E. J., Hunt, L. T., and Rushworth, M. F. S. 2009. "The Computation of Social Behavior," *Science* (324), pp. 1160-1164.
- Behrens, T. E. J., Hunt, L. T., Woolrich, M. W., and Rushworth, M. F. S. 2008. "Associative Learning of Social Value," *Nature* (456), pp. 245-249.
- Benbasat, I., Dimoka, A., Pavlou, P. A., and Qiu, L. 2010. "Incorporating Social Presence in the Design of the Anthropomorphic Interface of Recommendation Agents: Insights from an fMRI Study," in *Proceedings of the 31st International Conference on Information Systems*, St. Louis, Missouri, USA, December 12 - 15, pp. 1-22.
- Bente, G., Rüggenberg, S., Krämer, N. C., and Eschenburg, F. 2008. "Avatar-Mediated Networking: Increasing Social Presence and Interpersonal Trust in Net-Based Collaborations," *Human Communication Research* (34), pp. 287-318.
- Berg, J., Dickhaut, J., and McCabe, K. 1995. "Trust, Reciprocity, and Social History," *Games and Economic Behavior* (10), pp. 122-142.
- Brave, S., Nass, C., and Hutchinson, K. 2005. "Computers that Care: Investigating the Effects or Orientation of Emotion Exhibited by an Embodied Computer Agent," *International Journal of Human-Computer Studies* (62), pp. 161-178.
- Cacioppo, J. T., Berntson, G. G., Sheridan, J. F., and McClintock, M. K. 2000. "Multilevel Integrative Analyses of Human Behavior: Social Neuroscience and the Complementing Nature of Social and Biological Approaches," *Psychological Bulletin* (126), pp. 829-843.
- Clithero, J. A., Tankersley, D., and Huettel, S. A. 2008. "Foundations of Neuroeconomics: From Philosophy to Practice," *PLoS Biology* (6), pp. 2348-2353.
- Coleman, J. S. 1990. *Foundations of Social Theory*, Harvard University Press, Cambridge, MA.
- Coricelli, G., and Nagel, R. 2009 "Neural Correlates of Depth of Strategic Reasoning in Medial Prefrontal Cortex," *Proceedings of the National Academy of Sciences* (106), pp. 9163-9168.
- Cuzzocreo, J. L., Yassa, M. A., Verduzco, G., Honeycutt, N. A., Scott, D. J., and Bassett, S. S. 2009. "Effect of Handedness on fMRI Activation in the Medial Temporal Lobe during an Auditory Verbal Memory Task," *Human Brain Mapping* (30), pp. 1271-1278.
- Daft, R. L., and Lengel, R. H. 1986. "Organizational Information Requirements, Media Richness and Structural Design," *Management Science* (32), pp. 554-571.

- Daft, R. L., Lengel, R. H., and Trevino, L. K. 1987. "Message Equivocality, Media Selection, and Manager Performance: Implications for Information Systems," *MIS Quarterly* (11), pp. 355-366.
- Davidson, R. J. 2003. "Affective Neuroscience and Psychophysiology: Toward a Synthesis," *Psychophysiology* (40), pp. 655-665.
- Davis, A., Murphy, J., Owens, D., Khazanchi, D., and Zigurs, I. 2009. "Avatars, People, and Virtual Worlds: Foundations for Research in Metaverses," *Journal of the Association for Information Systems* (10), pp. 90-117.
- Delgado, M. R., Frank, R. H., and Phelps, E. A. 2005. "Perceptions of Moral Character Modulate the Neural Systems of Reward during the Trust Game," *Nature Neuroscience* (8), pp. 1611-1618.
- Desmond, J. E., and Glover, G. H. 2002. "Estimating Sample Size in Functional MRI (fMRI) Neuroimaging Studies: Statistical Power Analyses," *Journal of Neuroscience Methods* (118), pp. 115-128.
- Dimoka, A. 2010. "What Does the Brain Tell Us about Trust and Distrust? Evidence from a Functional Neuroimaging Study," *MIS Quarterly* (34), pp. 373-396.
- Dimoka, A., and Davis, F. D. 2008. "Where Does TAM Reside in the Brain? The Neural Mechanisms Underlying Technology Adoption," in *Proceedings of the 29th International Conference on Information Systems*, Paris, France, December 14 – 17, pp. 1-18.
- Dimoka, A., Banker, R. D., Benbasat, I., Davis, F. D., Dennis, A. R., Gefen, D., Gupta, A., Ischebeck, A., Kenning, P., Müller-Putz, G., Pavlou, P. A., Riedl, R., vom Brocke, J., and Weber, B. 2011. "On the Use of Neurophysiological Tools in IS Research: Developing a Research Agenda for NeuroIS," *MIS Quarterly* (forthcoming).
- Dimoka, A., Pavlou, P. A., Davis, F. D. 2007. "NEURO-IS: The Potential of Cognitive Neuroscience for Information Systems Research," in *Proceedings of the 28th International Conference on Information Systems*, Montreal, Quebec, Canada, December 9 – 12, pp. 1-20.
- Dimoka, A., Pavlou, P. A., Davis, F. D. 2010. "NeuroIS: The Potential of Cognitive Neuroscience for Information Systems Research," *Information Systems Research* (forthcoming).
- Ebner, N. C., Riediger, M., and Lindenberger, U. 2010. "FACES—A Database of Facial Expressions in Young, Middle-Aged, and Older Women and Men: Development and Validation," *Behavior Research Methods* (42), pp. 351-362.
- Ekman, P. 1982. *Emotion in the Human Face*, 2nd ed., Cambridge University Press, New York.
- Fehr, E. 2004. "Don't Lose Your Reputation," *Nature* (432), pp. 449-450.
- Fehr, E. 2009a. "On the Economics and Biology of Trust," *Journal of the European Economic Association* (7), pp. 235-266.
- Fehr, E. 2009b. "Social Preferences and the Brain," in *Neuroeconomics: Decision Making and the Brain*, P. W. Glimcher, C. F. Camerer, E. Fehr, and R. A. Poldrack (eds.), Academic Press, Amsterdam, pp. 215-232.
- Fogg, B. J., and Nass, C. 1997. "How Users Reciprocate to Computers: An Experiment that Demonstrates Behavior Change," in *Proceedings of the Conference on Human Factors in Computing Systems*, Atlanta, Georgia, USA, March 22 – 27, pp. 331-332.
- Frank, U., Schauer, C., and Wigand, R. T. 2008. "Different Path of Development of Two Information Systems Communities: A Comparative Study Based on Peer Interviews," *Communications of the Association for Information Systems* (22), pp. 389-412.
- Friedman, B., and Millett, L. I. 1997. "Reasoning about Computers as Moral Agents: A Research Note," in *Human Values and the Design of Computer Technology*, B. Friedman (ed.), CSLI Publications, Stanford, CA, pp. 201-205.
- Friedman, B., Kahn, P. H. Jr., and Howe, D. C. 2000. "Trust Online," *Communications of the ACM* (43), pp. 34-40.
- Frith, U., and Frith, C. D. 2003. "Development and Neurophysiology of Mentalizing," *Philosophical Transactions of the Royal Society B: Biological Sciences* (358), pp. 459-473.
- Frith, U., and Frith, C. D. 2010. "The Social Brain: Allowing Humans to Boldly Go Where No Other Species Has Been," *Philosophical Transactions of the Royal Society B: Biological Sciences* (365), pp. 165-176.
- Gallagher, H. L., Jack, A. I., Roepstorff, A., and Frith, C. D. 2002. "Imaging the Intentional Stance in a Competitive Game," *NeuroImage* (16), pp. 814-821.
- Gambetta, D. 1988. *Trust: Making and Breaking Cooperative Relations*, Basil Blackwell, Oxford.
- Gazzola, V., Rizzolatti, G., Wicker, B., and Keysers, C. 2007. "The Anthropomorphic Brain: The Mirror Neuron System Responds to Human and Robotic Actions," *NeuroImage* (35), pp. 1674-1684.

- Gefen, D., Benbasat, I., and Pavlou, P. A. 2008. "A Research Agenda for Trust in Online Environments," *Journal of Management Information Systems* (24), pp. 275-286.
- Gefen, D., Karahanna, E., and Straub, D. W. 2003. "Trust and TAM in Online Shopping: An Integrated Model," *MIS Quarterly* (27), pp. 51-90.
- Geller, T. 2008. "Overcoming the Uncanny Valley," *IEEE Computer Graphics and Applications* (28:4), pp. 11-17.
- Groom, V., Nass, C., Chen, T., Nielsen, A., Scarborough, J. K., and Robles, E. 2009. "Evaluating the Effects of Behavioral Realism in Embodied Agents," *International Journal of Human-Computer Studies* (67), pp. 842-849.
- Güth, W., Schmittberger, R., and Schwarze, B. 1982. "An Experimental Analysis of Ultimatum Bargaining," *Journal of Economic Behavior and Organization* (3), pp. 367-388.
- Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., and Gintis, H. 2004. *Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies*. Oxford University Press, Oxford.
- Huettel, S. A., and McCarthy, G. 2000. "Evidence for a Refractory Period in the Hemodynamic Response to Visual Stimuli as Measured by MRI," *NeuroImage* (11), pp. 547-553.
- Huettel, S. A., Song, A. W., and McCarthy, G. 2009. *Functional Magnetic Resonance Imaging*, 2nd ed., Sinauer Associates, Sunderland, MA.
- International Telecommunication Union. 2010. *World Telecommunication / ICT Development Report 2010 – Monitoring the WSIS Targets: A Mid-Term Review*, Geneva, Switzerland, http://www.itu.int/ITU-D/ict/publications/wtdr_10/material/WTDR2010_e.pdf.
- Jenkinson, M., and Smith, S. 2001. "A Global Optimisation Method for Robust Affine Registration of Brain Images," *Medical Image Analysis* (5), pp. 143-156.
- Jenkinson, M., Bannister, P., Brady, M., and Smith, S. 2002. "Improved Optimization for the Robust and Accurate Linear Registration and Motion Correction of Brain Images," *NeuroImage* (17), pp. 825-841.
- Keeney, R. L. 1999. "The Value of Internet Commerce to the Customer," *Management Science* (45), pp. 533-542.
- King-Casas, B., Tomlin, D., Anen, C., Camerer, C. F., Quartz, S. R., and Montague, P. R. 2005. "Getting to Know You: Reputation and Trust in a Two-Person Economic Exchange," *Science* (308), pp. 78-83.
- Kircher, T., Blümel, I., Marjoram, D., Lataster, T., Krabbendam, L., Weber, J., van Os, J., and Krach, S. 2009. "Online Mentalising Investigated with Functional MRI," *Neuroscience Letters* (454), pp. 176-181.
- Knutson, B. 1996. "Facial Expression of Emotion Influence Interpersonal Trait Inferences," *Journal of Nonverbal Behavior* (20), pp. 165-182.
- Kock, N. 2004. "The Psychological Model: Towards a New Theory of Computer-Mediated Communication Based on Darwinian Evolution," *Organization Science* (15), pp. 327-348.
- Kock, N. 2005. "Media Richness or Media Naturalness? The Evolution of Our Biological Communication Apparatus and Its Influence on Our Behavior toward e-Communication Tools," *IEEE Transactions on Professional Communication* (48), pp. 117-130.
- Kock, N. 2009. "Information Systems Theorizing Based on Evolutionary Psychology: An Interdisciplinary Review and Theory Integration Framework," *MIS Quarterly* (33), pp. 395-418.
- Kohli, R., Devaraj, S. 2003. "Measuring Information Technology Payoff: A Meta-Analysis of Structural Variables in Firm-Level Empirical Research," *Information Systems Research* (14), pp. 127-145.
- Kolb, B., and Wishaw, I. 2009. *Fundamentals of Human Neuropsychology*. Basingstoke, Hampshire, UK: Palgrave Macmillan.
- Krach, S., Blümel, I., Marjoram, D., Lataster, T., Krabbendam, L., Weber, J., van Os, J., and Kircher, R. 2009. "Are Women Better Mindreaders? Sex Differences in Neural Correlates of Mentalizing Detected with Functional MRI," *BMC Neuroscience* (10), pp. 1-11.
- Krach, S., Hegel, F., Wrede, B., Sagerer, G., Binkofski, F., and Kircher, T. 2008. "Can Machines Think? Interaction and Perspective Taking with Robots Investigated via fMRI," *PLOS ONE* (3), pp. 1-11.
- Krueger, F., Grafman, J., and McCabe K. 2008. "Neural Correlates of Economic Game Playing," *Philosophical Transactions of the Royal Society B: Biological Sciences* (363), pp. 3859-3875.
- Krueger, F., McCabe, K., Moll, J., Kriegeskorte, N., Zahn, R., Strenziok, M., Heinecke, A., and Grafman, J. 2007. "Neural Correlates of Trust," *Proceedings of the National Academy of Sciences* (104), pp. 20084-20089.
- Lee, J.-E. R., and Nass, C. I. 2010. "Trust in Computers: The Computers-Are-Social-actors (CASA) Paradigm and Trustworthiness Perception in Human-Computer Communication," in *Trust and*

- Technology in a Ubiquitous Modern Environment: Theoretical and Methodological Perspectives*, D. Latusek and A. Gerbasi (eds.), IGI Global, pp. 1-15.
- Lieberman, M. D. 2007. "Social Cognitive Neuroscience: A Review of Core Processes," *Annual Review of Psychology* (58), pp. 259-289.
- Lieberman, M. D., and Eisenberger, N. I. 2009. "Pains and Pleasures of Social Life," *Science* (323), pp. 890-891.
- Lieberman, M. D., Berkman, E. T., and Wager, T. D. 2009. "Correlations in Social Neuroscience Aren't Voodoo," *Perspectives on Psychological Science* (4), pp. 299-307.
- Luhmann, N. 1979. *Trust and Power*, Wiley, New York.
- MacDorman, K. F., Green, R. D., Ho, C.-C., and Koch, C. T. 2009. "Too Real for Comfort? Uncanny Responses to Computer Generated Faces," *Computers in Human Behavior* (25), pp. 695-710.
- McCabe, K., Houser, D., Ryan, L., Smith, V., and Trouard, T. 2001. "A Functional Imaging Study of Cooperation in Two-Person Reciprocal Exchange," *Proceedings of the National Academy of Sciences* (98), pp. 11832-11835.
- McKnight, D. H., and Chervany, N. L. 2001. "Trust and Distrust Definitions: One Bite at a Time," in *Trust in Cyber-Societies*, R. Falcone, M. Singh, and Y.-H. Tan (eds.), Springer, Berlin/Heidelberg, pp. 27-54.
- McKnight, D. H., Choudhury, V., and Kacmar, C. 2002. "Developing and Validating Trust Measures for e-Commerce: An Integrative Typology," *Information Systems Research* (13), pp. 334-359.
- Mori, M. 1970. "Bukimi No Tani [The Uncanny Valley]," *Energy* (7), pp. 33-35.
- Morris, J. P., Pelphrey, K. A., and McCarthy, G. 2005. "Regional Brain Activation Evoked when Approaching a Virtual Human on a Virtual Walk," *Journal of Cognitive Neuroscience* (17), pp. 1744-1752.
- Nagel, R. 1995. "Unraveling in Guessing Games: An Experimental Study," *The American Economic Review* (85), pp. 1313-1326.
- Nass, C., and Moon, Y. 2000. "Machines and Mindlessness: Social Responses to Computers," *Journal of Social Issues* (56), pp. 81-103.
- Nass, C., Moon, Y., Fogg, B. J., Reeves, B., and Dryer, D. C. 1995. "Can Computer Personalities Be Human Personalities?" *International Journal of Human-Computer Studies* (43), pp. 223-239.
- Nass, C., Steuer, J. S., Henriksen, L., and Dryer, D. C. 1994. "Machines and Social Attributions: Performance Assessments of Computers Subsequent to 'Self-' or 'Other-' Evaluations," *International Journal of Human-Computer Studies* (40), pp. 543-559.
- Nass, C., Steuer, J. S., Tauber, E., and Reeder, H. 1993. "Anthropomorphism, Agency, and Ethopoeia: Computers as Social Actors," in *Proceedings of the INTERCHI '93 Conference on Human Factors in Computing Systems*, Amsterdam, Netherlands, April 24 - 29, pp. 111-112.
- Nowak, K. L., and 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), pp. 481-494.
- Nowak, K. L., and Rauh, C. 2005. "The Influence of the Avatar on Online Perceptions of Anthropomorphism, Androgyny, Credibility, Homophily, and Attraction," *Journal of Computer-Mediated Communication* (11), pp. 153-178.
- Nowak, M. A., and Sigmund, K. 2005. "Evolution of Indirect Reciprocity," *Nature* (437), pp. 1291-1298.
- O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., and Dolan, R. J. 2004. "Dissociable Roles of Ventral and Dorsal Striatum in Instrumental Conditioning," *Science* (304), pp. 452-454.
- Ostrom, E. 2003. "Toward a Behavioral Theory Linking Trust, Reciprocity, and Reputation," in *Trust and Reciprocity: Interdisciplinary Lessons from Experimental Research*, E. Ostrom and J. Walker (eds.), Russell Sage Foundation, New York, NY, pp. 19-79.
- Pavlou, P. A., Liang, H., and Xue, Y. 2007. "Understanding and Mitigating Uncertainty in Online Exchange Relationships: A Principal-Agent Perspective," *MIS Quarterly* (31), pp. 105-136.
- Payne, J. W., and Venkatraman, V. 2011. "Opening the Black Box: Conclusions to a Handbook of Process Tracing Methods for Decision Research," *A Handbook of Process Tracing Methods for Decision Research*, M. Schulte-Mecklenbeck, A. Kühberger, and R. Ranyard (eds.), Psychology Press, New York, pp. 223-249.
- Phelps, E. A., O'Connor, K. J., Cunningham, W. A., Funayama, E. S., Gatenby, J. C., Gore, J. C., and Banaji, M. R. 2000. "Performance on Indirect Measures of Race Evaluation Predicts Amygdala Activation," *Journal of Cognitive Neuroscience* (12:5), pp. 729-738.
- Phillips, D. M., and Stanton, J. L. 2004. "Age-Related Differences in Advertising: Recall and Persuasion," *Journal of Targeting, Measurement & Analysis for Marketing* (13:1), pp. 7-20.

- Premack, D., and Woodruff, G. 1978. "Does the Chimpanzee Have a Theory of Mind?" *Behavioral and Brain Sciences* (1), pp. 515-526.
- Qiu, L., and Benbasat, I. 2005. "Online Consumer Trust and Live Help Interfaces: The Effects of Text-to-Speech Voice and Three-Dimensional Avatars," *International Journal of Human-Computer Interaction* (19), pp. 75-94.
- Qiu, L., and Benbasat, I. 2009. "Evaluating Anthropomorphic Product Recommendation Agents: A Social Relationship Perspective to Designing Information Systems," *Journal of Management Information Systems* (25), pp. 145-181.
- Qiu, L., and Benbasat, I. 2010. "A Study of Demographic Embodiments of Product Recommendation Agents in Electronic Commerce," *International Journal of Human-Computer Studies* (68), pp. 669-688.
- Raub, W., and Weesie, J. 1990. "Reputation and Efficiency in Social Interactions: An Example of Network Effects," *American Journal of Sociology* (96), pp. 626-654.
- Reeves, B., and Nass, C. 1996. *The Media Equation: How People Treat Computers, Television, and New Media like Real People and Places*, Cambridge University Press/CSLI, New York.
- Rescorla, R. A., and Wagner, A. R. 1972. "A Theory of Pavlovian Conditioning: Variations in the Effectiveness of Reinforcement and Nonreinforcement," in *Classical Conditioning II: Current Research and Theory*, A. H. Black, W. F. Prokasy (eds.), Appleton-Century-Crofts, New York, pp. 64-99.
- Rice, R. E. 1993. "Media Appropriateness: Using Social Presence Theory to Compare Traditional and New Organizational Media," *Human Communication Research* (19), pp. 451-484.
- Riedl, R., Banker, R. D., Benbasat, I., Davis, F. D., Dennis, A. R., Dimoka, A., Gefen, D., Gupta, A., Ischebeck, A., Kenning, P., Müller-Putz, G., Pavlou, P. A., Straub, D. W., vom Brocke, J., Weber, B. 2010b. "On the Foundations of NeuroIS: Reflections on the Gmunden Retreat 2009," *Communications of the Association for Information Systems* (27), pp. 243-264.
- Riedl, R., Hubert, M., and Kenning, P. 2010a. "Are There Neural Gender Differences in Online Trust? An fMRI Study on the Perceived Trustworthiness of eBay Offers," *MIS Quarterly* (34), pp. 397-428.
- Rilling, J. K., Gutman, D. A., Zeh, T. R., Pagnoni, G., Berns, G. S., and Kilts, C. D. 2002. "A Neural Basis for Social Cooperation," *Neuron* (35), pp. 395-405.
- Rilling, J. K., Sanfey, A. G., Aronson, J. A., Nystrom, L. E., and Cohen, J. D. 2004. "The Neural Correlates of Theory of Mind within Interpersonal Interactions," *NeuroImage* (22), pp. 1694-1703.
- Rotter, J. B. 1967. "A New Scale for the Measurement of Interpersonal Trust," *Journal of Personality* (35), pp. 651-655.
- Sanfey, A. G., Loewenstein, G., McClure, S. M., and Cohen, J. D. 2006. "Neuroeconomics: Cross-Currents in Research on Decision-Making," *Trends in Cognitive Sciences* (10), pp. 108-116.
- Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E., and Cohen, J. D. 2003. "The Neural Basis of Economic Decision-Making in the Ultimatum Game," *Science* (300), pp. 1755-1758.
- Schachter, S. C. 2000. "The Quantification and Definition of Handedness: Implications for Handedness Research," *Side Bias: A Neuropsychological Perspective*, M. K. Mandal, B. M. Bulman-Fleming, and G. Tiwari (eds.), Kluwer Academic Publishers, Dordrecht, pp. 155-174.
- Semmes, C. E. 1991. "Developing Trust: Patient-Practitioner Encounters in Natural Health-Care," *Journal of Contemporary Ethnography* (19), pp. 450-470.
- Singer, T. 2009. "Understanding Others: Brain Mechanisms of Theory of Mind and Empathy," in *Neuroeconomics: Decision Making and the Brain*, P.W. Glimcher, C. F. Camerer, E. Fehr and R. A. Poldrack (eds.), Academic Press, Amsterdam, pp. 251-268.
- Singer, T., and Lamm, C. 2009. "The Social Neuroscience of Empathy," *Annals of the New York Academy of Sciences* (1156), pp. 81-96.
- Sun, T., and Walsh, C. A. 2006. "Molecular Approaches to Brain Asymmetry and Handedness," *Nature Reviews Neuroscience* (7), pp. 655-662.
- Sutter, M., and Kocher, M. G. 2007. "Trust and Trustworthiness across Different Age Groups," *Games and Economic Behavior* (59:2), pp. 364-382.
- Sutton R. S., Barto A.G. 1998. *Reinforcement Learning: An Introduction*, MIT Press, Cambridge, MA.
- Teven, J. J., and Hanson, T. L. 2004. "The Impact of Teacher Immediacy and Perceived Caring on Teacher Competence and Trustworthiness," *Communication Quarterly* (52), pp. 39-53.
- Todorov, A. 2008. "Evaluating Faces on Trustworthiness: An Extension of Systems for Recognition of Emotions Signaling Approach / Avoidance Behaviors," *Annals of the New York Academy of Sciences* (1124), pp. 208-224.

- Wang, W., and Benbasat, I. 2005. "Trust in and Adoption of Online Recommendation Agents," *Journal of the Association for Information Systems* (6), pp. 72-101.
- Winston, J. S., Strange, B. A., O'Doherty, J., and Dolan, R. J. 2002. "Automatic and Intentional Brain Responses during Evaluation of Trustworthiness of Faces," *Nature Neuroscience* (5), pp. 277-283.
- Woolrich, M. W., Ripley, B. D., Brady, M., and Smith, S. M. 2001. "Temporal Autocorrelation in Univariate Linear Modelling of fMRI Data," *NeuroImage* (14), pp. 1370-1386.
- Yoon, C., Gutchess, A. H., Feinberg, F., and Polk, T. A. 2006. "A Functional Magnetic Resonance Imaging Study of Neural Dissociations between Brand and Person Judgments," *Journal of Consumer Research* (33), pp. 31-40.