Understanding social responses to artificial agents : building blocks for persuasive technology

Understanding social responses to artificial agents

Building blocks for persuasive technology

Maaike Roubroeks

UNDERSTANDING SOCIAL RESPONSES TO ARTIFICIAL AGENTS

BUILDING BLOCKS FOR PERSUASIVE TECHNOLOGY

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de Technische Universiteit Eindhoven, op gezag van de rector magnificus prof.dr.ir. C.J. van Duijn, voor een commissie aangewezen door het College voor Promoties, in het openbaar te verdedigen op dinsdag 27 mei 2014 om 16:00 uur

door

Maaike Anna Johanna Roubroeks

geboren te Vlodrop

Dit proefschrift is goedgekeurd door de promotoren en de samenstelling van de promotiecommissie is als volgt:

voorzitter:	prof.dr. J. de Jonge
1e promotor:	prof.dr. C. J. H. Midden
2e promotor:	prof.dr. E. O. Postma (Universiteit Tilburg)
Copromotor(en):	dr. J. R. C. Ham
Leden:	prof.dr. F. A. Eyssel (Universität Bielefeld)
	prof.dr. V. Evers (Universiteit Twente)
	prof.dr. P. Markopoulos
	prof.dr. W. A. IJsselsteijn

Chapter 1

General Introduction

Persuasive communication is probably one of the most frequent forms of social interaction. In daily life most humans are often subjected to attempts to change their attitudes and actions. The government may try to persuade you to eat healthy, the community to keep your environment clean, and your partner tries to persuade you to buy those expensive shoes. Sometimes these persuasive attempts are successful, other times they are not. Designing effective persuasive communication can be a challenge. For example, interventions aimed at changing energy-consumption behavior face difficulties to yield lasting effects (see e.g., McCalley & Midden, 2002). Behavior change is usually studied in human-human communication. Many features of those interactions have been studied as determinants of effective persuasion (see e.g., Cialdini, 1993; Cialdini, 2009). Among these, factors related to the message source have been studied in depth, which has led to important insights in the roles of trustworthiness, attractiveness and the social behavior of the source (see e.g., Miller & Baseheart, 1969; Petty, & Cacioppo, 1984; Smith, De Houwer, Nosek, 2013; Yoo, Gretzel & Zanker, 2013). Recently, persuasive communication has changed as it is not anymore the exclusive domain of human sources, but may also be provided by persuasive technology: smart systems designed to influence human behavior through persuasion.

Persuasive technology emerged during the last decade (e.g., Fogg, 2002; IJsselsteijn, De Kort, Midden, Eggen, & Van den Hoven, 2006), and has been defined as “any interactive computing system designed to change people’s attitudes or behaviors” (Fogg, 2002, p. 1). Similarly, IJsselsteijn and colleagues (2006) defined it as “a class of technologies that are intentionally designed to change a person’s attitude or behavior” (IJsselsteijn et al., 2006). Persuasive technology does not use coercive or deceptive means to change attitudes or behaviors. Instead, it aims at inducing voluntary change of behaviors (Fogg, 2002).

Applications of persuasive technology vary from helping users to adopt and maintain a healthy lifestyle to helping household members to reduce their energy consumption. In general, persuasive technology can be used to help humans accomplish their goals. An

important advantage of persuasive technology compared to human persuasion is that technology is not restricted by human feelings or emotions. That is, it does not sense social boundaries that human persuaders would not cross (e.g., technology does not feel uncomfortable asking users about their energy bills). Therefore, persuasive systems can be persistent (even to the point of annoyance, e.g., by asking a hundred times to save energy), allow for anonymity (in case of private information, such as one's daily energy usage), have no memory limitation, switch modalities (e.g., visually or auditory feedback), are omnipresent (e.g., on the Web, on a mobile phone, on a tablet) and are time-independent (e.g., provide feedback about water consumption during showering; Fogg, 2002; IJsselsteijn et al., 2006).

When aiming at persuading humans to change their behavior through persuasive technology, making proper design choices should optimize the probability of success. Fogg (2002) proposed a framework in which persuasive technological devices can be distinguished into one of three types: (1) tools that increase the capability of a person (e.g., tracking device); (2) media that provide users with experiences (e.g., cause-and-effect simulator); or (3) social actors (i.e., social artificial agent or robot) that employ relationships with human agents. Especially the last type is interesting, because in human-human interaction social influence strategies have been shown to be very effective in changing human behavior (e.g., Mittelmark, 1999; Schultz, Nolan, Cialdini, Goldsteijn, & Griskevicius, 2007) and it may also be effective when artificial social agents interact with humans in a persuasive setting. Two lines of evidence support the persuasive effect of social influence strategies and of artificial social agents, respectively. First, a multitude of experimental studies revealed that social influence strategies induce large behavioral changes in human participants, despite the fact that the strategies are rated by the same participants as having a low impact on their behaviors (Nolan, Schultz, Cialdini, Goldsteijn, & Griskevicius, 2008). Second, there is increasing support for persuasiveness of artificial agents (Ham, & Midden, 2010; Midden & Ham, 2009; Riether, Hegel, Wrede, & Horstmann, 2012).

Persuasiveness of Social Artificial Agents

A study that illustrates the effects of artificial social agents in persuasive technology is presented by Fogg and Nass (1997a). In their study, participants performed two tasks while interacting with a computer. In the first task the computer either helped or did not help the participant. In the second task the participant had the option to help or not to help the computer. Results showed that participants helped the computer in the second task more often when the computer helped them in the first task, than when this was not the case (Fogg, & Nass, 1997a). This finding was explained by the authors in terms of reciprocity behavior towards the computer. The participants reciprocated the helpful or unhelpful behavior of the computer (as has been found in human-human persuasion; Cialdini, 1993).

In contrast to the positive findings on the effectiveness of artificial social agents in persuasive technology, there are also failures to find positive effects. In their overview of studies of the persuasive effects of artificial agents, Dehn and Van Mulken (2000) found mixed results concerning the effectiveness of such agents. Some studies did find positive effects of artificial agents (e.g., King & Ohya, 1996; Koda & Maes, 1996, Lester, et al., 1997), whereas others found no or negative effects of artificial agents (e.g., Takeuchi & Nagao, 1993). Dehn and Van Mulken (2000) suggested that the mixed results could be caused by differences in experimental settings (e.g., different dependent variables, different manipulations, or different control conditions) and, as a consequence, it was difficult to compare the various experimental results and draw firm conclusions. In addition, they noted that many of the reviewed studies contained confounding factors. For example, a study by Lee and Nass (1999) investigated participant's conformity to agents that were either represented by text boxes or by human-like agents. The text boxes communicated via text and the human-like agents communicated via speech and by showing idle behavior (e.g., breathing movements and eye blinking). With this manipulation both the embodiment of the agent (i.e., text box vs. human-like agent) and the modus of communication changed (text vs. speech) making it difficult to discern the contribution of each manipulation separately.

Still, more recent research in human-computer interaction suggested that attempts of social influence by an artificial agent can be quite effective. For example, Riether and colleagues (2012) showed that the mere presence of robots (similar to the mere presence of other humans) led to social facilitation effects (e.g., Zajonc, 1965). Social facilitation effects cause human participants to perform better on easy tasks in the company of others, and to perform worse on complex tasks in the company of others. Riether and colleagues (2012) performed a study in which participants were asked to perform easy and complex tasks in the presence or absence of a human technical assistant or a robot. The robot had an anthropomorphic head and performed human-like behavior. They found that participants' performance on easy tasks improved in the presence of the robot or assistant, as compared to when they were absent. For complex tasks, their performance deteriorated in the presence of the robot or assistant, as compared to when they were absent. In other words, the effects of robots on human performance were similar to the effects of other humans on human performance.

Midden and Ham (2009) showed that an artificial agent can also influence participant's energy consumption choices. They suggested that feedback by a social robot was more persuasive in reducing participants' energy consumption than feedback through an energy indicator (indicated by LED lights) that gave feedback about the participant's energy consumption. In this study, participants were asked to program a virtual washing machine. In one condition they received feedback about their energy-consumption level from a robot that uttered positive (e.g., "Good") or negative (e.g., "Terrible") words, and used supporting facial expressions. In the other condition an energy bar was placed on the washing machine that showed the participant's current energy level. Participants conserved more energy when they received feedback from the social robot than when they received feedback from the energy bar (i.e., factual feedback). However, it was not clear whether the effect was due to the evaluative feedback the robot gave (i.e., "Good" vs. "Terrible"), or due to the social nature of the robot's feedback (i.e., robot vs. energy bar). Follow-up research

confirmed that the social nature of the feedback was the most effective factor in changing behavior (Vossen, Ham, & Midden, 2010).

Indications for the effectiveness of artificial social agents can also be found in field research that investigated interactions with artificial social agents in elderly care (for a review see Bemelmans, Gelderblom, Jonker, & De Witte, 2012; Broekens, Heerink, & Rosendal, 2009), in education (e.g., Lester, et al., 1997; Moreno, Mayer, Spires, & Lester, 2001; for an overview see Saerbeck, Schut, Bartneck, & Janse, 2010), and in health behavior (e.g., Kiesler, Powers, Fussell, & Torrey, 2008; Looije, Cnossen, Neerincx, 2006; Looije, Neerincx, Cnossen, 2010). For instance Wada, Shibata, Saito, Sakamoto and Tanie (2005) designed a robot that was effective in influencing elderly. That is, Wada and colleagues showed that elderly humans who interacted with a seal robot (i.e., a social robot that looks like a seal; see Figure 1.1), treated it as if it was a real social agent. They bonded with the seal robot and in response their moods improved as a result of interacting with the robot (Wada, et al., 2005).

With the development of interactive, internet-based technology, effective persuasive agents are becoming more and more ubiquitous. For example, artificial social agents influencing users can be found on Web sites (e.g., virtual assistants at help desks), or implemented in household products (e.g., the electric toothbrush of Oral B that shows a smiling face if you brush your teeth effectively; see Figure 1.2).

Figure 1.1. A seal robot that is used in elderly care.

Figure 1.2. Electric toothbrush Oral B.

The recent evidence in support of the attitudinal and/or behavioral impact of artificial social agents raises the question how humans actually view artificial social agents. Do they actually believe that artificial agents are social beings with a mind of their own, or are they

merely considered to be a piece of technology that is preprogrammed to act socially? Earlier research suggested that humans respond to artificial agents as if they were responding to other human beings (e.g., Reeves & Nass, 1996). However, when asked about their responses, participants denied behaving socially to the artificial agents (e.g., Nass & Moon, 2000). So, they are consciously aware that artificial agents do not warrant social treatment, but they behave as if agents *do* warrant social treatment.

This raises the main question of the current dissertation "Why do humans exhibit social responses to artificial agents?" In the current dissertation we investigated social responses to artificial agents in an attempt to uncover why artificial agents can be effective as social agents. Before attempting to answer the main question, we elaborate on the notion of an artificial social agent.

What is an Artificial Social Agent?

The term social agent is frequently used as a synonym for a human, or something that is (temporarily) experienced to have humanness or human characteristics. An important human characteristic is human agency. According to Bandura (2001), human agency is characterized by four core features: intentionality, forethought, self-reactiveness and self-reflectiveness. Intentionality in social agents means that they plan and act out their behavior with the intention to accomplish a goal. Forethought in social agents implies that they anticipate on actions to reach potential consequences and adjust their actions in such a way that the goal has the best chance of being accomplished. Self-reactive social agents monitor their actions and motivate and regulate their execution. Finally, self-reflective social agents reflect on their actions and make choices in case of conflicting courses of action (as described in Bandura, 2001).

Earlier research suggested that artificial agents could also function as social agents. Fong, Nourbakhsh, and Dautenhahn (2003) defined artificial social agents as being socially intelligent; meaning that they are capable of showing *human social intelligence*, like the ability to learn, react and interact with the social environment (see also Bandura, 2001), and build up social relationships (see also Levy, 2007). Artificial agents can emulate human

social intelligence by using social cues. For instance, social cues can take the form of a face or a voice (Fogg, 2002), or as human behavior implying a personality or an emotional state (Blascovich, 2002; Epley, Waytz, & Cacioppo, 2007). Epley and colleagues (2007) argued that when humans observe the actions of an artificial agent (behavioral cues), they attribute human personality traits and human mental states (e.g., emotional states) to the artificial agent (see "Theories about social responses to artificial agents" for more details). For example, when a robot hits everyone who comes near, a human observer might infer the robot has an aggressive personality, and when a robot has a smiling face, a human observer might infer that the robot likes the human observer. So, based on earlier theories and research, we argue that to make an artificial agent social, it needs to exhibit social cues that give humans the experience that the agent is a social being. In addition to the social cues formed by a face, by a voice or by human behavior (Blascovich, 2002; Epley et al., 2007; Fogg, 2002), research is needed that investigates which other factors could contribute to humans experiencing the agent as a social being.

Returning to our main question "Why do humans exhibit social responses to these artificial agents?", we will start by reviewing recent theories about social responses to artificial agents.

Theories about Social Responses to Artificial Agents

In the literature on human-computer interaction several theories have been proposed that can give guidelines for explaining human social responses to artificial agents. Various researchers have posed theories and investigated factors related to the human part and the agent part of the human-agent interaction that could explain social responses to artificial agents (e.g., Blascovich, 2002; Epley, Waytz, Akalis, & Cacioppo, 2008; Epley, et al., 2007; Guadagno, Blascovich, Bailenson, & McCall, 2007; Nass, & Moon, 2000; Von der Pütten, Krämer, Gratch, & Kang, 2010; Reeves & Nass, 1996; Waytz, Cacioppo, & Epley, 2010). In the following sections we will discuss those theories seeking an explanation of why humans exhibit social responses to artificial agents.

Theory of anthropomorphism

Several theories focus on characteristics of the human part of the human-agent interaction that enhance the likelihood that humans exhibit social responses towards artificial agents. One of these theories is Epley's anthropomorphism theory. Epley's theory about anthropomorphism suggests that humans have the tendency to infer human traits and properties from observed behavior (Epley, et al., 2007). When humans anthropomorphize, they attribute human characteristics to non-human agents (e.g., artificial agents). According to Epley and colleagues (2007) anthropomorphism is an inductive inference process (i.e., bottom-up reasoning) that is determined by cognitive and motivational factors. That is, when humans anthropomorphize artificial social agents, they first activate (highly accessible) knowledge structures about humans as an anchor or inductive base, which they may correct later on by using additional information about artificial agents. For instance, when humans see a robot waving its arms heavily and hitting everything that comes near it, they may infer that the robot is aggressive (inductive base). In other words, humans use their knowledge about "other humans that wave their arms heavily and hit everyone who comes near them (i.e., aggressive humans)" as an inductive base when they observe a robot doing the same thing. However, when they later observe that the robot is just broken, they may correct for their initial inferences and change which traits they attributed to the robot.

Epley and colleagues (2007) have proposed three factors that increase the likelihood that humans anthropomorphize non-human agents; elicited agent knowledge (cognitive factor), effectance motivation (motivation factor), and sociality motivation (motivation factor). The first factor (elicited agent knowledge) means that when humans have knowledge about other humans in general, and knowledge about themselves specifically, they use this knowledge as a starting point and attribute it to artificial agents. Recent work of Eyssel, Kuchenbrandt, Hegel and De Ruiter (2012) indicated that the judgments of a robot were influenced when manipulating the robot's vocal cues (i.e., elicited agent knowledge). However, Epley and colleagues (2007) proposed that humans anthropomorphize less when they are motivated and have the cognitive opportunity to correct for already gained

information about artificial agents (e.g., when they know that artificial agents work on algorithms) than when they do not have this motivation or opportunity for correction (Epley et al., 2007). Thus, Epley and colleagues (2007) suggested that humans have an automatic tendency to anthropomorphize artificial agents, but when they acquire technical knowledge about artificial agents, and are able to use this knowledge, they reduce this automatic tendency and anthropomorphize less. In other words, in that case social responses to artificial agents will be minimized.

The second factor (effectance motivation) means that when humans want to make sense of the interaction with an unknown artificial agent, they are motivated to use human traits to explain the artificial agent's behavior. In other words, they are motivated to interact effectively with the artificial agent (White, 1959). Epley and colleagues (2007) proposed that humans do this in order to reduce uncertainty or ambiguity about the situation and increase the predictability of future behavior of the artificial agent. Research showed that when humans have the need to reduce uncertainty about an unknown situation and to increase the predictability of future interactions, they anthropomorphize more (Epley, et al., 2007; Epley, et al., 2008; Waytz, et al., 2010). That is, they will respond more socially to artificial agents.

The third factor (sociality motivation) means that humans have a higher tendency to describe artificial agents in terms of human traits, when they are lonely or are in need of social connection. Epley and colleagues (2007) explain that humans do this by paying selective attention to possible social cues and seeking social connection wherever it is possible. More specifically, they seek human traits in artificial agents that reduce their feelings of loneliness. Research showed that when participants have a need for social connection, they anthropomorphize more (Epley, et al., 2007; Epley, et al, 2008; Epley, et al., 2008; Eyssel & Reich, 2013). Thus, respond more socially to artificial agents.

In short, when humans use information about human beings to explain the behavior of artificial agents, when they want to reduce uncertainty about the interaction with artificial agents, or when they feel lonely, they are more inclined to anthropomorphize artificial

agents. These characteristics on the human part of the human-agent interaction are supposed to lead humans to respond socially to these artificial agents.

Threshold model of social influence

Other theories describe factors at the agent part of the human-agent interaction, or contextual factors that increase the social responses to artificial agents. *The threshold model of social influence* by Blascovich (2002) suggested that when humans interact with artificial agents, they engage in a process of social verification. Relevant for our research question is that Blascovich suggested that humans use cues of agency (e.g., a human-like face) and behavioral realism (e.g., human behavior) to verify that they are in a meaningful interaction. If they believe being in a meaningful interaction, they respond socially to artificial agents. Blascovich (2002) broadly defined agency as the extent to which humans perceive artificial agents as (representations of) real persons, which possesses uniquely human properties, like a consciousness, emotions, and intentionality (Waytz, et al., 2010). Behavioral realism is the extent to which humans believe the agent behaves realistically; as if it was part of a real interaction (Blascovich, 2002). In Blascovich's (2002) view, agency and behavioral realism are complimentary to each other. That is, if humans perceive that an artificial agent has a low level of agency, a high level of behavioral realism is necessary to lead to social verification of the artificial agent (and vice versa).

Von der Pütten and colleagues (2010) performed a systematic test of this model by manipulating agency and behavioral realism. In their study they manipulated agency by telling participants either that they would be interacting with a computer algorithm (low level of agency) or with another participant (high level of agency). Participants were led to believe that both the computer algorithm as the other participant controlled an onscreen character. Von der Pütten and colleagues (2010) found that it did not matter whether participants believed that the onscreen character had a low or a high level of agency; the evaluations of the virtual character were largely the same whether participants believed the virtual character was controlled by a human or by a computer algorithm. In a more recent study Midden and Ham (2012) also did not find an effect of agency when manipulating whether a

robot's actions seemed to be controlled by a human (high level of agency), whether the robot seemed to control its own action (moderate level of agency), or even whether it seemed that the robot's actions were random (low level of agency). Next to investigating agency, Von der Pütten and colleagues (2010) manipulated behavioral realism by programming the onscreen character to show human listening behavior or not. To measure a wide array of evaluations of the virtual character and participant's behavioral responses they analyzed twenty dependent factors (e.g., positive and negative affect to the virtual character, person perception, social presence, and intimacy to the virtual character). Von der Pütten and colleagues (2010) furthermore found that their behavioral realism manipulation had an effect on three out of twenty dependent measures. In other words, in line with Blascovich (2002), they concluded that a higher level of behavioral realism (i.e., more social cues) led to more social behavior of the participants. However, considering the fact that seventeen of the twenty measures Von Pütten and colleagues (2010) used, did not show the expected effect, it may be that using twenty tests increased the possibility of a false positive outcome for the remaining behavioral realism measures (type I error). Taking these studies together, we conclude that there is no univocal evidence for the model of Blascovich (2002).

The media equation hypothesis

According to the media equation hypothesis (Reeves & Nass, 1996) humans automatically respond socially when interacting with artificial agents. That is, they respond as if they are responding to another human being. Reeves and Nass (1996) explained that responding socially to artificial agents is not something only children or lowly educated humans would do. They claimed that all humans automatically respond social to artificial social agents, because they most often respond mindlessly. When responding mindlessly, humans respond without thorough thinking, and are not aware of responding socially. Furthermore, Reeves and Nass claimed that these responses are (almost) inescapable (Reeves, & Nass, 1996). Supporting this theory, numerous studies have shown social responses to computers or artificial agents (e.g., Fogg & Nass, 1997a; Fogg & Nass, 1997b Moon, 2000; Nass & Moon, 2000; Nass, Moon, & Carney, 1999; Nass, Moon, & Green,

1997; for an overview see Reeves & Nass, 1996). For example, earlier research demonstrated that participants respond politely to a computer when that computer also obliges to rules of politeness (Nass, et al., 1999), participants stereotyped computers (Nass, et al., 1997), and showed reciprocity to computers (Fogg & Nass, 1997a; Moon, 2000). However, when asked directly whether participants thought artificial agents require social treatment, they knew that these responses were not appropriate and even felt offended by the question (Reeves & Nass, 1996). Hence, when participants were confronted with their own responses (Reeves & Nass, 1996), they seemed to treat the artificial agent more as an object.

Nass (2004) and Nass and Moon (2000) have tried to explain these contradictory responses to artificial agents. Nass (2004) has suggested that these human-like responses could be explained by an evolutionary account. Humans have evolved in a world in which other humans were associated with gaining opportunities and getting help when problems arose. As Nass (2004) explains "*…there would be a significant evolutionary advantage to the rule: If there's even a low probability that it's human-like, assume it's human-like"* (p. 37). In other words, if an artificial agent "behaves" like a human, then it will be assumed that it is human-like and act accordingly (Nass, 2004; Reeves & Nass, 1996). Nass and Moon (2000) proposed that the underlying process for applying social rules and expectations towards computers can be attributed to mindlessness (see also Johnson & Gardner, 2007; Johnson & Gardner, 2009; Johnson, Gardner & Wiles, 2004; Lee, 2008). In their research (Nass & Moon, 2000; see also Reeves & Nass, 1996) they refer to the theory of mindfulness/ mindlessness by Langer (1992) for explaining these apparent automatic, social responses. Langer (1992) describes the process of mindlessness as "*…being in a state of mind characterized by an overreliance on categories and distinctions drawn in the past and in which the individual is context-dependent and, as such, is oblivious to novel (or simply alternative) aspects of the situation*" (p. 289). In other words, humans who are interacting with artificial agents find themselves in a "mindless state", comparable to a form of automatic responding (see Bargh, 1984; Bargh, 1990) in which they only attend to the social cues, but

seem to (temporarily) ignore that the artificial agents are just technology (in the section "Conclusions" and in Chapter 4 we will discuss this in greater detail). As a consequence, they automatically respond socially to the artificial agents.

Nass and Moon (2000) suggested that various cues could heighten the possibility of social responses to artificial agents. For example, computers that possess human-like characteristics (like a voice, face, etcetera) are probably able to trigger social responses. Other examples of cues that heighten the possibility of social responses are: attitudes and behaviors that are controlled predominantly by automatic processes (e.g., humans can automatically determine whether a voice is female or male); attitudes and behaviors that are used often (e.g., politeness); and attitudes and behaviors that do *not* violate expectations about artificial agents (e.g., Mori, 1970; Nass & Moon, 2000). According to Nass and Moon (2000) these attitudes and behaviors lead to a higher chance of humans responding socially. *The uncanny valley hypothesis* is related to this latter suggestion (violation of expectation). This hypothesis proposes that when humans are reminded that the interaction partner is not real, expectations about the interaction partner are violated and result in feelings of uncanniness and aversion. For example, in human-human interaction the feelings of uncanniness can arise when during a handshake a prosthetic, rather than a human hand is felt (Mori, 1970). Thus if you notice non-social cues while interacting with artificial agents, you are reminded that you are actually interacting with a piece of technology and get snapped back to reality.

Reeves and Nass (1996) argue that it is not clear whether multiple social cues also would lead to more social responses. However, several theories proposed that the more social cues are available, the more social the interaction becomes (e.g., *Media richness theory* by Daft & Lengel, 1986; *Social agency theory* by Mayer, Sobko & Mautone, 2003; *Social cue hypothesis* by Louwerse, Graesser, Lu, & Mitchell, 2005; *Social presence theory* by Biocca, Harms & Burgoon, 2003; *The threshold model of social influence* by Blascovich, 2002). These theories claim that when more social cues are available humans are more

inclined to respond socially. However, more research is needed to draw any firm conclusions.

Conclusions

In the previous paragraphs we outlined several theories about social responses to artificial agents. To come back to our research question: What can the discussed theories learn us about why humans respond socially to artificial agents? First of all, Epley and colleagues (2007) proposed factors on the human part of the human-agent interaction that make humans respond socially to artificial agents. More interesting for the current dissertation are the theories on the agent part of the human-agent interaction, which may offer design characteristics for persuasive agents that are able to persuade humans to change their behavior. Comparing the theoretical model of social influence (Blascovich, 2002) with the media equation hypothesis (Reeves & Nass, 1996) we can see that they are not in agreement with each other. That is, Blascovich (2002) suggested that agency and behavioral realism trigger social responses to artificial agents. In contrast, Reeves and Nass (1996) claimed that it does not matter whether the agent shows human agency or behavioral realism. Rather, their research suggested that simple social cues are enough to make participants respond automatically social, as if they were responding to other humans.

Furthermore, these three theories suggest that social responses to artificial agents happen automatically. Although the theory of anthropomorphism (e.g., Epley, et al., 2007) and the theoretical model of social influence (Blascovich, 2002) only *implied* the automaticity of social responses to artificial agents, the media equation hypothesis (Reeves & Nass, 1996) explicitly stated that humans respond automatically social to artificial agents.

These theories put forward a question that has not been yet experimentally investigated: “Are social responses to artificial agents indeed automatic?” Furthermore, they raise the question what exactly is meant with “automaticity”. Reeves and Nass (1996) equal automaticity with mindlessness. More recent work in implicit social cognition, however, showed that automaticity is not a single concept, but that automaticity consists of multiple components. According to Bargh (1994), automaticity consists of four components; (1)

awareness, in which humans can be unaware of the stimulus (e.g., an artificial agent behaving socially), the interpretation of the stimulus, or the effect of the stimulus; (2) intentionality (e.g., unintentionally responding socially to an artificial agent); (3) controllability (e.g., control responding socially when observing an artificial agent behaving socially); and (4) efficiency (e.g., whether responding socially to the artificial agent requires cognitive resources; also see Moors, & De Houwer, 2006).

A better understanding of the nature of human social responses to artificial agents requires a closer look at how these social responses relate to these components of automaticity. This is one of the goals of the current research. Therefore, we investigated a social response that is established to be automatic on all four components of automaticity: spontaneous trait inferences (Uleman, Saribay, and Gonzalez, 2008; see also Uleman, Newman, & Moskowitz, 1996). Spontaneous trait inferences are drawn when humans observe the behavior of a person and infer a trait to this person that could be implied from the behavior. For example, when you observe a man kicking a puppy on the street, you could infer that the man is cruel (Carlston & Skowronksi, 1994). Research suggested that participants are not aware of drawing spontaneous trait inferences, and that they draw them efficiently (i.e., not using a lot of cognitive capacity) and spontaneously, and they (almost) cannot control drawing spontaneous trait inferences (Uleman, et al., 1996; Uleman, et al., 2008). We suggest that spontaneous trait inferences about artificial agents are also automatic on all four components of automaticity. In the context of persuasion it is important that humans are able to control their responses, because persuading humans does not mean to coerce or deceive them. For that reason we particularly focused on the controllability of social responses to artificial agents. That is, we will use a research paradigm to find out whether it is possible to control for social responses to artificial agents.

In summary, the aim of the current dissertation is to investigate the nature of human social responses to artificial agents and to test to what extent these responses occur automatically. The results of this research could give insights in why humans exhibit social responses to artificial agents and are able to form a social bond with artificial agents, which

are considered as important factors in persuasion (see also Brehm, 1966). Most important for the current dissertation is that this knowledge could help us to design artificial agents that are effective in persuading humans to change their behavior. The current research did *not* investigate direct responses of humans to persuasive agents in a social interaction. Instead, we investigated social responses to artificial agents that are paired with social cues. The artificial agents were either pictures of agents or moving agents that were paired with social cues. The focus of the social responses we measured was related to the subject of persuasion (see next paragraph). Our research results can be used as building blocks for persuasive agents. To attain these goals, we investigated social responses to artificial agents experimentally by means of research paradigms used in human-human interaction studies of social influence and person perception.

The Current Dissertation: Overview of Empirical Chapters

In this dissertation we report on our investigations of participants' automatic and controlled responses to artificial agents[1]. In Chapter 2 we aimed to demonstrate that participants respond socially to an artificial agent and explored factors that possibly contribute to the underlying mechanisms of these social responses. Furthermore, in the context of persuasion, we investigated whether, just like in human-human interaction, persuasive attempts of an artificial agent could backfire. In Chapter 3, we investigated the automaticity of participants' social responses to artificial agents. In this chapter we compared participants' automatic and controlled responses towards humans, artificial agents, and inanimate objects. In this way, we investigated Reeves and Nass' (1996) suggestion that human automatic responses to artificial agents are similar to their responses to other humans. In addition, as previously implied (Reeves & Nass, 1996), we investigated whether

[1] In all of our studies we treated our Likert scales as interval data and analyzed these data by means of ANOVA or t-tests. This was done for several reasons. We treated the Likert scales as interval because they consisted of sums across many items. That is, we used mean scores instead of a single-item score. In this way, our data is more similar to interval than to ordinal. Treating the Likert scales as merely ordinal would lose information. Furthermore, the Likert scales we used were designed to have equal distances between each value and the items in the Likert scale measured a single latent variable (Norman, 2010). Treating our data as interval would be more reliable in this case than treating it as ordinal. Also, parametric tests can be used for Likert scale data because of the Central Limit Theorem (Norman, 2010). To be sure, we randomly compared some analyses with non-parametric tests to ANOVA and found similar results (Field, 2009). Finally, most psychological research uses ANOVA (or t-tests) as its method of analysis and to keep our research results comparable to this research we followed this procedure. Therefore, we analyzed the Likert scales as interval data using ANOVA or t-tests.

human controlled responses to artificial agents are more similar to their responses to objects. Moreover, we investigated some specific aspects of this underlying process (discussed below) that could lead to social responses to artificial agents. All these aspects of automaticity can be useful when designing persuasive agents. For example, if simple social cues are sufficient to trigger social responses to artificial agents, then it is not necessary to spend a lot of money on socially-enriched agents. Finally, in Chapter 4, we investigated whether humans can control for their social responses. Our persuasive agent used empathy-provoking information as a social influence strategy. We observed whether humans showed empathy (i.e. a social response) to the persuasive agent. Empathy-provoking information was for example used by means of a persuasive agent telling a story about its dog dying. When humans feel empathy, they are more prone to show prosocial behavior in response (e.g., Twenge, Baumeister, DeWall, Ciarocco, & Bartels, 2007). We also investigated whether humans were able to control for their social reactions. In the context of persuasion it is important to know whether technology persuades humans, or 'accidentally' triggers automatic responses that are uncontrollable altogether. If humans can control for their social responses, they can protect themselves from social influence strategies, like empathy-provocation. In the following paragraphs we will give an overview of the research conducted per chapter.

Chapter 2:

In this chapter we explored social responses to a robotic agent and aimed to show that these also occur in the domain of negative responses. We sought to demonstrate that, just like in human-human interaction, humans become psychologically reactant when a robot threatens their autonomy to choose. According to Brehm (1966; Brehm & Brehm, 1981), humans will go against this threat to confirm or restore their autonomy. For example, when a person requests somebody to save energy, that other person may decide to consume even more energy. Furthermore, we investigated two factors that could influence the underlying process of social responses to a robotic agent. First, we investigated whether there is a linear relationship between social agency of an artificial agent and the degree of social

responses of the user. We did this by directly manipulating the social agency in the interaction. Second, we observed participants' social responses when we indirectly induced social agency by implying that the robot has its own goals.

Chapter 3:

In Chapter 3, we investigated participants' responses in a paradigm used in person perception that was designed to compare automatic and controlled responses. Furthermore, we compared responses to humans with responses to artificial agents and responses to inanimate objects. We added the latter category to see whether participants' controlled responses to artificial agents are more similar to their responses to inanimate objects or to their responses to humans. The automatic social responses we investigated in these studies were spontaneous trait inferences (STIs), while the controlled responses were intentional trait inferences (ITIs). Humans draw STIs when they observe someone's behavior and spontaneously draw trait inferences about the actor of the behavior (Uleman, et al., 1996). ITIs are drawn intentionally (Uleman, et al., 1996). We chose to study STIs because STIs are automatic on all four components of automaticity (Uleman, et al., 2008; see also Uleman, et al., 1996). In other words, humans draw STIs unintentionally, they are not aware of drawing STIs, they can (almost) not control themselves drawing STIs, and STIs are drawn efficiently. Also, we investigated some specific aspects of the possible underlying process for the automatic social responses to artificial agents. That is, we investigated whether participants actually inferred traits to artificial agents automatically, or whether they made automatic associations instead. Automatic associations are made when humans link two stimuli (e.g., traits and artificial agents) in their memory (Bassili, 1989). Inferences go beyond mere linkages between two stimuli and indicate that humans truly infer personality traits to artificial agents. Thus, do humans really believe the agent could possess human personality traits, or do they only make associations between human personality traits and artificial agents?

Chapter 4:

In this chapter we investigated whether it is possible to control the automatic social responses to artificial agents. In other words, we investigated whether participants are doomed to respond socially when observing an artificial social agent, or whether they are still able to control their automatic social responses. So, in Chapter 4, we reminded participants that they were interacting with technology, rather than with a person. We therefore focused the attention of participants on the technical characteristics of an artificial agent and measured whether their social responses diminished.

Chapter 5:

In Chapter 5 we summarized conclusions, and discussed limitations, and possible utilizations and implications of our results for the design of persuasive agents. We speculated about: whether simple social cues are enough to trigger social responses or whether it is necessary to let humans believe the artificial agent possesses human characteristics; about what social cues are effective for persuasive agents; whether it is better to trigger automatic or controlled social responses; and whether humans are able to control their social responses. In this way, we can provide building blocks for designing persuasive agents.

Chapter 2

Don't Tell Me What To Do, Robot!

Demonstrating Psychological Reactance To an Artificial Agent*

---

* This chapter is partly based on two publications:

Roubroeks, M., Ham, J., & Midden, C. (2011). When artificial social agents try to persuade people: The role of social agency on the occurrence of psychological reactance. International Journal of Social Robotics, 3, 155-165. doi: 10.1007/s12369-010-0088-1;

Roubroeks, M., Ham, J., & Midden, C. (2010). The dominant robot: Threatening robots cause psychological reactance, especially when they have incongruent goals. In T. Ploug, P. Hasle, & H. Oinas-Kukkonen (Eds.). *Lecture Notes in Computer Science: Vol. 6137. Persuasive Technology* (pp. 174-184). Berlin Heidelberg, Germany: Springer-Verlag. doi:10.1007/9783-642-13226-1_18.

John sees himself as a green person and wants to save energy. He decides to get a robot that helps him with his energy consumption. The robot greets John and John says “Hi” back. John is very excited about using the robot and connects the robot to his smart meter. The next morning, when John starts his daily routine, the robot starts to give John advice about his energy consumption: “You have to shut off the TV,” “Switch off that light in the hallway!,” and “You need to wash your laundry at a lower temperature.” After several of these pieces of advice, John starts to experience feelings of anger, and negative thoughts come to mind like “I don’t *have* to do anything!” This is a typical example of a negative outcome of persuasion, labeled psychological reactance (Brehm, 1966).

Recently, the research area of persuasive technology has emerged (Fogg, 2002). This research area investigates the way in which computers and artificial agents persuade humans. Persuasive technology was defined as “*any interactive computing system designed to change people’s attitudes or behaviors”* (Fogg, 2002, p. 1). Fogg (2002) argued that artificial agents could be used as persuasive social actors. He described five social cues that can be used to persuade humans; (1) Physical cues, like a face, eyes, or a body; (2) Psychological cues, like a personality or cues inducing feelings of empathy; (3) Language, written or spoken; (4) Social dynamics, like turn taking in a conversation, cooperation or reciprocity; and (5) Social roles, like being a friend or opponent. When humans observe these cues, they could make inferences of the artificial agent being a social entity and therefore act in a way they would do when observing a human. Also, they could be influenced through social influencing mechanisms used in human-human interaction (Fogg, 2002). For example, research suggested that when participants had to program a washing machine, social feedback (e.g., positive and negative facial expressions and utterances) of a robotic agent led to more energy reduction than factual feedback (i.e., kW/h information) from an energy bar (Midden, & Ham, 2008).

In the current chapter we aim to demonstrate that humans respond socially to an artificial agent and explore two factors (Study 1 and Study 2) that possibly contribute to the underlying mechanisms of these social responses. Furthermore, in the context of

persuasion, we will investigate whether persuasion could backfire. We argue that if humans can be persuaded by artificial agents, as was proposed by Fogg (2002), they can also experience psychological reactance in response to being influenced by artificial agents (as in the example described in the first paragraph). If so, this could be detrimental for persuasive technology and designers should take these results into account when designing a persuasive agent. For instance, instead of lowering their energy consumption, humans could start consuming even more energy if they experience a persuasive message about energy conservation as a threat[2] to their autonomy. This would mean that designers accomplished the opposite of the behavior they intended to stimulate. Earlier research assessed reactance in human-human interactions, in which the persuasive social actor was always human (for an overview see Burgoon, Alvaro, Grandpre, & Voulodakis, 2002). We argue that experiences of reactance will also occur when humans are persuaded by artificial agents in human-agent interactions. Support for this argument can be found in the *Computers As Social Actors paradigm* (CASA paradigm; Reeves & Nass, 1996), which stated that participants respond to artificial agents (e.g., computers) as if they were responding to other participants. The CASA paradigm (Nass & Moon, 2000; Reeves & Nass, 1996) suggested that social cues trigger social behavior rules and humans respond in accordance with these social behavior rules (e.g., Eyssel & Hegel, 2012). We suggest that this also holds for the social behavior rules of psychological reactance. That is, when humans feel threatened in their autonomy by technology that attempts to persuade them, they will experience psychological reactance, and as a result want to restore their autonomy. Therefore, we expect that a "social" artificial agent can also trigger psychological reactance.

Psychological Reactance in Human-Human Interaction

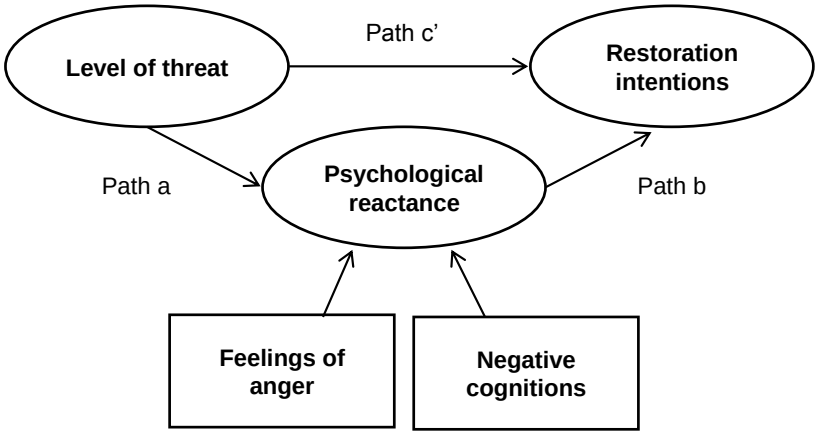
Importantly, persuasive messages inherently contain some kind of directive to change a specific attitude or behavior. Humans could experience these directives as a threat to their autonomy, and consequently might experience some sort of arousal. In response, humans can experience a strong desire to restore this feeling of autonomy (Brehm, 1966;

[2] In this article the term "threat" is always referred to as an intrusion to one's autonomy to choose.

Brehm & Brehm, 1981). The theory of psychological reactance argued that the following elements (see Figure 2.1) are relevant to reactance; (1) a perceived autonomy (e.g., autonomy to conserve energy or not), (2) a threat to that perceived autonomy (e.g., a robot persuading you to conserve energy), (3) the experience of psychological reactance (i.e., as indicated by feelings of anger and negative cognitions), and (4) a desire to restore the threatened autonomy (e.g., the desire to consume even more energy).

The restoration of autonomy can occur directly, by doing the forbidden act; or indirectly, by an increase of attractiveness or liking of the eliminated option (Brehm, Stires, Sensenig, & Shaban, 1966), by denying the existence of the threat to autonomy, by exercising a similar autonomy to gain a feeling of control, or by derogating the source (described in Dillard & Shen, 2005). Quick and Stephenson (2007b) recently developed a scale to measure restoration behavior. They distinguished three different forms of restoration. For example, when you persuade humans to conserve energy with their washing machine they can do several things to restore their freedom; (1) boomerang restoration, in which humans consume even more energy when using the washing machine; (2) related boomerang restoration, in which humans consume more energy in a related context (e.g., setting the thermostat at a higher temperature); or (3) vicarious boomerang restoration, in which humans watch others consuming more energy with the washing machine (Quick & Stephenson, 2007b).

Figure 2.1. Model of psychological reactance described in Dillard & Shen (2005).



Psychological reactance seems to be a very fundamental human response. That is, the theory of psychological reactance (Brehm, 1966; Brehm & Brehm, 1981) is in line with self-determination theory (Ryan, & Deci, 2000) that stated that one of the basic humans needs is the need for autonomy. According to self-determination theory humans have an intrinsic motivation to continuously strive for autonomy (Ryan, & Deci, 2000). The need for autonomy is even so important that humans automatically become psychologically reactant when they feel an intrusion of their autonomy. A recent study (Chartrand, Dalton, & Fitzsimons, 2007) showed that participants, even non-consciously, acted reactant when confronted with someone who wanted to control them. In this study they were either subliminally primed with a highly-controlling significant other and the goal to either work hard or have fun on a subsequent task. It showed that participants, without awareness, pursued the opposite goal that was intended when they were primed with the high-controlling significant other (see also Fitzsimons & Lehmann, 2004). In a second study it was shown that this effect was especially true for participants who scored high on trait reactance (Chartrand, et al., 2007).

Recent work has investigated one of the most important cues that caused psychological reactance: the way in which a message was formulated (e.g., Buller, Borland, & Burgoon, 1998; Dillard & Shen, 2005; Grandpre, Alvaro, Burgoon, Miller, & Hall, 2003; Miller, Lane, Deatrick, Young, & Potts, 2007; Quick & Consodine, 2008; Quick & Stephenson, 2007a; Quick & Stephenson, 2008; Rains & Turner, 2007; Reinhart, Marshall, Feeley, & Tutzauer, 2007). This research indicated that when language became more controlling, participants were more prone to respond in a reactant way. For example, research by Grandpre and colleagues (2003) suggested that adolescents who were provided with explicit (i.e., containing clear intentions) messages against smoking (e.g., “Do not smoke”) showed more signs of psychological reactance and initiated more smoking behavior than adolescents who were provided with implicit (i.e., containing no clear intentions) messages against smoking (e.g., “Smoking is not cool”). Intriguingly, this effect was reversed for pro-smoking messages. That is, whether adolescents did smoke or not smoke after the

experiment did not depend on the content of the messages (pro-smoking vs. anti-smoking), but on the intrusion of their autonomy.

A study by Miller and colleagues (2007) provided insights in the underlying processes leading to psychological reactance. This research suggested that the use of concrete (descriptive) language that avoids any doubts about the meaning of the message, resulted in more compliant behavior. However, when a message contained controlling language (i.e., containing explicit directives), persuasion was diminished and psychological reactance was enhanced. In addition, Miller and colleagues' (2007) results indicated that high-controlling language made use of imperatives (i.e., commands and orders) and controlling terms such as "have to", "must", "should", "ought to", or "need to". On the other hand, low-controlling language made use of terms (i.e., suggestions) like "could", "can", "may", "might", or "could try to" (Miller, et al., 2007). High-controlling messages were experienced as directive and these messages clearly showed the persuaders intentions, whereas low-controlling messages were experienced more as suggestions and as autonomy-supportive (Miller, et al., 2007). In short, Miller and colleagues (2007) suggested that when using a (low-controlling) concrete message, persuasion will be enhanced. But when the same message contains high-controlling language, compliance will decrease and might even lead to psychological reactance.

Psychological Reactance as a Measurable Concept

Recent research proposed that psychological reactance can be measured by assessing feelings of anger and negative cognitions (Dillard & Shen, 2005). That is, in the original theory of psychological reactance, Brehm and Brehm (1981) explained that psychological reactance cannot be measured directly. This is because it is a motivational state that is the consequence of a threat to autonomy. They concluded that only the consequences of psychological reactance can be measured (like restoration behavior; Brehm & Brehm, 1981). However, a more recent study by Dillard and Shen (2005) proposed a different conceptual perspective on psychological reactance. In their study, they presented two concepts that could be used as indicators of psychological reactance; feelings of anger

and negative cognitions. Other research showed that the influence of these two factors on psychological reactance could be best explained by an intertwined model—the two indicators contributed to psychological reactance in an intermingling way (Dillard & Shen, 2005; Quick & Stephenson, 2007a; Rains & Turner, 2007). That is, anger and negative cognitions were best explained as an inseparable construct. The implication of the intertwined model is that one should assess both anger and negative cognitions when measuring psychological reactance. Designers of persuasive agents could use these measures of anger and negative cognition to explore whether their persuasive agents trigger psychological reactance in humans. In this way, the probability of reactance could be minimized before the persuasive agent is introduced to real users.

Psychological Reactance and Social Agency

In earlier research, psychological reactance is defined as a social phenomenon that is the outcome of an interaction between human social actors (Brehm & Brehm, 1981). Brehm (1966) proposed that "…reactance will frequently occur in response to restrictions or threats thereof imposed by social entities" (p. 387). As stated, we argue that psychological reactance can also occur in interactions between humans and artificial agents. However, since Brehm and Brehm (1981) defined reactance as a social phenomenon we also argue that for reactance to occur, the human who is influenced might need to perceive the artificial agents as a social entity. In other words, experiencing the artificial agent as a social actor with human-like characteristics (like intentionality). That is, we argue that humans have to verify that the interaction with artificial agents is a meaningful, social interaction. According to the theoretical model of social influence (Blascovich, 2002), this kind of social verification depends on two variables: perceived agency and perceived behavioral realism. Blascovich (2002) defined perceived agency as the extent to which humans perceive artificial agents as representations of real humans. Humans can infer agency by interpreting the actions of the artificial agents (e.g., seems like controlled by a human) or by prior knowledge (e.g., prior to the interaction observe that a human controls the agent as suggested by Guadagno, et al., 2007). Behavioral realism is the extent to which humans perceive artificial agents as

behaving similar to humans in a real social interaction (human-human interaction). Blascovich (2002) describes that realism is only of value to the extent that it facilitates the social interaction. Thus it is important that the social cues in the interaction that induce realism facilitate the social interaction. He also proposes that agency and behavioral realism are complementary to each other. That is, when a human perceives a computer interaction partner as having a low level of agency, the level of behavioral realism has to be high for social verification to occur (and vice versa). Importantly, Blascovich (2002) argued that only if social verification occurs, persuasion by that computer becomes possible. In other words, only when humans perceive enough agency or behavioral realism and subsequently social verification occurs, then humans will interact with artificial agents as if they were interacting with real humans. In line with this proposition, the CASA paradigm (Reeves & Nass, 1996) also stated that humans respond in the same way to artificial agents as to humans, but (in contrast to Blascovich, 2002) added that only simple social cues were needed to trigger this response.

Other theories, like social agency theory (Mayer, et al., 2003) and the social-cue hypothesis (Louwerse, et al., 2005) stated that when social cues (e.g., voice, presence of a face, facial expressions) are available in an interaction, the interaction becomes more social. That is, the human-agent interaction is experienced to a higher degree as a human-human interaction. Additionally, when more social cues are available, humans try to understand the relationship with the other actor better (Louwerse, et al., 2005; Mayer, et al., 2003). Related to that, research concluded that a more realistic interaction led to a greater influence of the agent (Blascovich, 2002; Guadagno et al., 2007). More support for the hypothesis that more social cues lead to more realistic interactions comes from media richness theory (Daft & Lengel, 1986). In this theory it is proposed that when humans interact with “richer” media (with face-to-face interaction as the richest possible cue), humans will behave as if they were in a human-human interaction. For example, research from Kidd and Breazeal (2005) showed that artificial agents that had a higher social presence were seen as more convincing, more entertaining and were more engaged with these agents than artificial

agents with lower social presence. In other words, a higher social presence led humans to behave more as if they were interacting with other humans. In our research we will explore whether simple social cues are enough to trigger social responses to artificial agents.

Based on the theories and research findings presented in the previous paragraph, we propose that when a human observes an artificial agent, persuasion will become more effective if the agent has a high level of social agency (compared to a low level of social agency). We further reason that when such an artificial agent (with a high level of social agency) is more persuasive, the chance that humans experience this agent as a threat will be greater. A higher threat will lead to more reactance. Therefore, we assumed that a higher social agency will be seen as more threatening (than a lower social agency), and therefore psychological reactance will be greater. In addition, it could be that humans perceive an agent with a high level of social agency as having a higher amount of control than an agent with a low level of social agency, and consequently they experience a greater threat to their autonomy.

Although there are many definitions and types of agency (e.g., Bandura, 2001; Blascovich, 2002; Gray, Gray, & Wegner, 2007; Hernandez & Iyengar, 2001; Kashima, et al., 2005; Mayer, et al., 2003; Wegner, Sparrow, & Winerman, 2004; Wojciszke, Abele, & Baryla, 2009), we will use the term social agency to indicate the degree to which an artificial agent is perceived as a social entity, similar to Bandura's (2001) human agency concept. In other words, social agency is the degree to which an artificial agent is perceived as being capable of social behavior that resembles human behavior in a human-human interaction. Humans do not have to believe that an artificial social agent is an autonomous social entity that has its own intentionality (Bandura, 2001). Instead, they treat social agents as if they were humans (even if it is only temporary) if they experience some kind of humanness (see also Reeves, & Nass, 1996).

The Current Research

Previous research already suggested that participants respond socially to artificial agents (Reeves & Nass, 1996). In the current two studies we investigated whether participants experienced psychological reactance (i.e., responding socially) towards an artificial agent that tried to persuade them. Furthermore, we investigated whether manipulating social agency (Study 1) or goal congruency (Study 2) would lead to an increase of psychological reactance (i.e., social responses). Study 1 is an initial test to investigate whether participants showed psychological reactance to an artificial agent. In addition, we investigated whether participants who were persuaded by an artificial agent that used more social cues led to an increase in psychological reactance. In Study 1, participants read a short instruction on how to conserve energy when using the washing machine. This instruction either used low-threatening (e.g., "You could…") language or used high-threatening (e.g., "You have to…") language. In line with previous research in human-human interaction (e.g., Miller, et al., 2007), we hypothesized that participants would report more psychological reactance after reading an instruction using high-threatening language than after reading an instruction using low-threatening language. We also manipulated social agency of the actor as the source of the instruction text by; presenting some participants only with text describing the instruction (low level of social agency); some participants with the same text paired with a still picture of an artificial agent (medium level of social agency); and some participants with the same text paired with a short film clip of the same artificial agent (high level of social agency). We hypothesized that after participants had read instructions that used high-threatening language and that were paired with a short film clip of an artificial agent, they would report the highest levels of psychological reactance.

Study 1

Method

Participants and Design

We recruited 138 participants[3] (70 males, 68 females; age *M* = 35.17, *SD* = 15.64) on the Internet via a Web link that was posted on Facebook and Hyves (Dutch social network site), and sent to the participant database[4] of the Eindhoven University of Technology. All participants were native Dutch speakers. Participants were randomly assigned to a 3 (threat level: no threat vs. low threat vs. high threat) x 3 (social agency level: low social agency vs. medium social agency vs. high social agency) between-subjects design. Additionally, we measured whether the relationship between threat and psychological reactance was mediated by restoration behavior. The experiment lasted about 15 minutes, for which participants were paid €3 (approximately $3.75 U.S. at the time this study was conducted), and participants were provided with the opportunity to win an additional price of €100 (approximately $125 U.S. at the time the experiment was conducted).

Materials

Instruction text. The text of the instruction described how to conserve energy when using a washing machine. To manipulate threat level, this text either mainly used non-controlling language (e.g., “People could save energy when using the washing machine.”), low-controlling language (e.g., “You could save energy when using the washing machine.”), or high-controlling language (e.g., “You have to save energy when using the washing machine.”). The language use in the texts was based on previous studies about psychological reactance (Dillard & Shen, 2005; Miller, et al., 2007; Quick & Considine, 2008).

[3] In total we recruited 176 participants. However, because of technical failures, data of 38 participants got lost. Analyses are done on the data of the remaining 138 participants.
[4] The database of the Eindhoven University of Technology consisted of people whose ages ranged from 18 to 40, and who lived in or near Eindhoven.

To manipulate social agency[5] of the source of this instruction text, we presented this text without any accompanying display of a source (low social agency), a still picture of an artificial agent (medium social agency), or a short film clip of the same artificial agent (high social agency).

Picture. The still picture showed a robotic agent (i.e., the virtual iCat[6]) that had a neutral facial expression. The instruction text was placed within a speech-bulb that was connected to the robotic agent, clearly implying that the robotic agent was uttering the words (medium social agency condition; see Figure 2.2).

Figure 2.2. The medium social agency condition.

Short film clip. The short film clip showed the same robotic agent (i.e., the virtual iCat) as the picture, but now the mouth of the robotic agent made speech movements indicating that it was saying the words, although no actual sound was audible (high social agency condition).

Perceived threat to autonomy. To check the effect of our manipulation of threat level, we used a Dutch translation of a measure of perceived threat to freedom developed by Dillard and Shen (2005). For each question, participants could indicate their agreement to a statement on a five-point Likert scale, ranging from (1) *Completely disagree to* (5) *Completely agree*. These four statements were: "*The advice restricted my autonomy to choose how I wanted to do the laundry,*" "*The advice tried to manipulate me,*" "*The advice tried to make a decision for me,*" and "*The advice tried to pressure me.*" The mean score on

[5] We tried to do a manipulation check of social agency. However, it was hard to assess the degree of social agency participants inferred to the low-agency source, because no source was available to judge. However, we did find that the medium social agency agent and the high social agency agent did not differ in anthropomorphism, $F < 1$, $p > .05$. (Godspeed questionnaire of Bartneck, 2000). This is also supported by our research results.

[6] The iCat is a yellow robotic agent that is designed by the Philips Research corporation. It resembles a torso of a cat. It can show facial expressions and can be programmed to talk. The virtual iCat is the virtual image of the real robotic agent.

these four statements formed a reliable measure (α = .90) for the perceived threat to autonomy score.

Feelings of anger. To measure feelings of anger we used a Dutch translation of the measure of anger developed by Dillard and Shen (2005). For each question, participants could indicate their agreement to a statement on a four-point Likert scale ranging from (1) *Not at all to* (4) *Completely*. These four statements were: "*I was irritated,*" "*I was angry,*" "*I was annoyed,*" and "*I was aggravated*". The mean score on these four statements formed a reliable measure (α = .83) for feelings of anger.

Negative cognitions. To measure negative cognitions we used a thought-listing task. This task was based on a measure that was used in Dillard and Shen (2005) and translated to Dutch. In this task, participants were asked to report all thoughts they had while reading the instruction text, even when those thoughts had nothing to do with the instruction text. There was no time limit, and participants could report their thoughts by typing as much text as they wanted in a text box (the amount of thoughts the participants typed in ranged in amount from 1 to 10, with M = 3.2, SD = 1.9). After listing all thoughts, participants had to indicate for all thoughts whether a thought was either negative (by typing an "*N*" behind that thought description), positive (by typing a "*P* " behind that thought description), or neutral (by typing "Neu" behind that thought description). Examples of negative thoughts that participants reported were "Stupid cat!" and "I don't HAVE TO do anything" (translated to English). We followed the procedure of Dillard and Shen (2005) for removing emotions to minimize overlap between affect and cognition in the reported thoughts. For each participant, we calculated a score we labeled the negative cognitions score by taking the total amount of reported thoughts and calculating the percentage of the reported negative thoughts.

Restoration intentions. To measure restoration intentions, we used a Dutch translation of the reactance restoration scale (RRS; Quick & Stephenson, 2007b). This measure consists of three questions and uses a seven-point continuum, with the following anchor points "*motivated–unmotivated*", "*determined–not determined*", "*encouraged–not encouraged*", and "*inspired–not inspired*". The following three items were assessed: "*Right*

now, I am [. . .] to save energy when doing the laundry" (reversed scored; boomerang restoration), "*Right now, I am [. . .] to be around others that save energy when doing the laundry*" (reversed scored; vicarious boomerang restoration), and "*Right now, I am [. . .] to do something totally energy consuming*" (related boomerang restoration). The mean score of these three restoration questions formed a reliable measure (α = .92) for the boomerang restoration score; the mean score of the vicarious boomerang restoration items formed a reliable measure (α = .97) for the vicarious boomerang restoration score; and the mean score of the related boomerang restoration items formed a reliable measure (α = .97) for the related boomerang restoration score.

Procedure

Participants were invited to participate in an online experiment on energy conservation in the household. They were asked to read an instruction text about energy conservation. Participants in the no threat condition read a non-threatening instruction text, participants in the low threat condition read a low-threatening instruction text, and participants in the high threat condition read a high-threatening instruction text. For some participants, the instruction text was either provided as text-only (low social agency); for some participants the instruction text was accompanied by a still picture of a robotic agent (medium social agency; see Figure 2.2); and for some participants the instruction text was accompanied by a short film clip, in which the same robotic agent moved its mouth as if it was talking (high social agency). After reading the instruction text, participants were asked to perform the thought-listing task, and were asked to fill in the perceived threat to autonomy scale, the feelings of anger scale, and fill in the restoration intentions scale. Finally, participants were thanked for their participation, debriefed, paid, and reminded of the opportunity to win €100.

Results

Manipulation Check of Threat Level

An analysis of the effect of threat level on threat score suggested that our threat level manipulation was successful. That is, a One-Way Analysis of Variance (ANOVA), with threat level as the independent variable and threat score as the dependent variable showed a significant effect of our threat manipulation[7], $F(2, 135) = 17.31$, $p < .001$, $n_p^2 = .20$. Planned comparison analyses indicated that participants in the high threat condition reported more perceived threat ($M = 3.44$, $SD = 1.21$) than participants in the low threat condition ($M = 2.20$, $SD = 1.04$), $t(135) = 5.40$, $p < .001$, $r = .42$, and reported more perceived threat for the high threat condition than the no threat condition ($M = 2.28$, $SD = 1.00$), $t(135) = 5.00$, $p < .001$, $r = .39$. However, these analyses did not indicate that participants in the no threat condition reported a different perceived threat to autonomy from participants in the low threat condition, $t(135) = .38$, $p = .700$. A possible explanation for this might be that participants in the no threat condition still experienced the instruction text as directed at them personally, and therefore experienced it as a low threat.

Threat and Social Agency

In line with previous research in human-human interaction (e.g., Dillard & Shen, 2005) we expected that participants who were exposed to a high-threatening artificial agent would be more psychologically reactant (indicated by higher scores for reported negative cognitions and feelings of anger) than participants who were exposed to a low-threatening artificial agent. To analyze this, the negative cognitions score and the anger score were submitted to a 2 (psychological reactance measure: negative cognitions vs. feelings of anger) x 3 (threat level: no threat vs. low threat vs. high threat) x 3 (social agency level: low social agency vs. medium social agency vs. high social agency) Repeated Measures

[7] We also tested whether the manipulation of social agency by itself led to more experienced threat. This was done in order to exclude the possibility that people experienced more psychological reactance because a higher social agent is seen as more threatening. Results showed that social agency did not lead to differences in the amount of experienced threat, $F(2,135) = 1.88$, $p = .16$. That is, the agent did not cause any threat by itself. We also measured liking towards the agent, but we did not find any effects between conditions, $F < 1$, $p > .05$.These findings suggest that liking towards the agent was equally high for participants in all conditions.

ANOVA[8], with the first factor serving as a within-subjects factor[9]. Confirming our expectations, results showed that there was a significant main effect of threat level, $F(2, 129) = 13.81$, $p < .001$, $n_p^2 = .18$[10] (see Table 2.1 for mean scores and standard deviations).

Table 2.1. Mean scores psychological reactance score (and standard deviations between brackets) in all conditions for the complete MANOVA design.

	Threat level		
	No threat	Low threat	High threat
Psychological reactance	11.46_a (13.69)	16.85_a (16.31)	28.04_b (14.75)

Note: Mean scores in rows that do not share the same subscript differ significantly, $p < .001$.

Planned comparisons between the three levels of threat level indicated that participants in the high threat condition reported more feelings of psychological reactance than either participants in the low-threat condition, or participants in the no threat condition, both *p*s < .001. Again, there was no significant difference between the no threat condition and the low-threat condition, $p = .191$. Furthermore, the effect of threat level was qualified by an interaction of Threat Level X Psychological Reactance Measure, $F(2, 129) = 13.12$, $p < .001$, $n_p^2 = .17$, indicating that the effect of threat level is different on the measure of feelings of anger than on the measure of negative cognitions (see Table 2.2). Closer examination of the separate effects on feelings of anger and negative cognitions showed that the effect of threat level was significant for both measures, both *p*s < .001, but that there was a stronger effect size of threat level on the measure of feelings of anger, $F(2, 129) = 19.23$, $n_p^2 = .23$, than of threat level on the measure of negative cognitions, $F(2, 129) = 13.47$, $n_p^2 = .17$ (see Table 2.2). Although the correlation between negative cognitions and feelings of anger was quite high (Pearson's $r = .61$, $p < .001$) we found a different effect on the two psychological reactance measures. That is, it seemed that the effects are not completely inseparable. This result was a bit surprising, because previous research by Dillard and Shen (2005) suggested

[8] We also found a main effect of social agency, $F(2, 129) = 3.66$, $p = .028$, $n_p^2 = .05$, and an interaction of Social Agency Level X Type of Psychological Reactance $F(2, 129) = 3.66$, $p = .029$, $n_p^2 = .05$.

[9] Field (2009) described to use a Repeated Measures ANOVA instead of separate ANOVAs when the dependent variables are correlated. We did find that anger and negative cognitions were highly correlated and therefore used a Repeated Measures ANOVA.

[10] The Levene's test of the anger score was significant, so we note that caution should be taken for the interpretation of our results.

that the effects of their distinct effects cannot be disentangled. However, these authors do note in their discussion that this does not always have to be the case, because "*. . . cognition and affect are phenomena of rapid change*" (Dillard & Shen, 2005, p. 160).

Table 2.2. Means scores on the negative cognitions score and feelings of anger score (and standard deviations between brackets) for the threat manipulation.

	Threat		
Psychological Reactance	No Threat	Low Threat	High Threat
Negative cognitions	21.51_a (26.86)	32.25_a (32.16)	54.00_b (28.76)
Anger	1.40_a (.53)	1.45_a (.46)	2.08_b (.74)

Note: Mean scores in rows that do not share the same subscript differ significantly, $p < .05$.

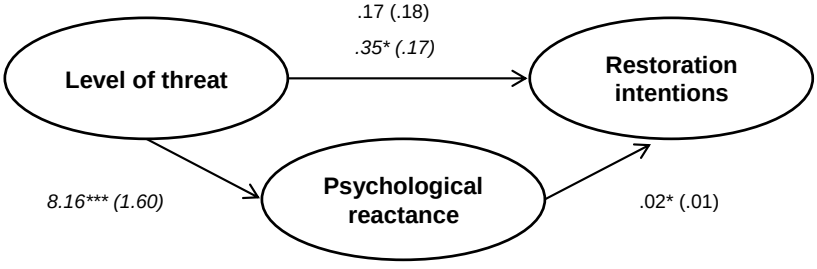
We also expected that an artificial agent with a higher level of social agency would be better able to persuade participants. Therefore, we suggested that when humans feel threatened by this artificial agent with a high level of social agency, they would even become more psychologically reactant. However, the expected interaction effect of Threat Level X Social Agency Level was not found, $F(4, 129) = .33$, $p = .858$. That is, we did not find that when participants who were exposed to a high-threatening agent with a high level of social agency led to a significant increase of psychological reactance. A possible explanation for this null effect could be that our manipulation of social agency was not powerful enough to elicit an additional effect on psychological reactance.

Restoration

In line with previous research about psychological reactance investigating reactance in human-human interaction (e.g., Quick & Stephenson, 2007b), we expected that threat level would predict restoration intentions, and that this relationship would be mediated by psychological reactance (see Figure 2.3). In other words, when participants felt threatened they became psychologically reactant and in response experienced restoration intentions. To analyze whether psychological reactance mediated the relationship between threat level and restoration intentions we used the method proposed by Baron and Kenny (1986). Step 1 of the analysis is to check for a positive relationship of threat level with restoration. We analyzed this for the three types of restoration intentions (boomerang restoration, related

boomerang restoration, and vicarious boomerang restoration) separately. A regression analysis provided no evidence that threat level predicted the boomerang restoration score, $t(137) = 1.59$, $p = .113$, nor the related boomerang score, $t(137) = 1.04$, $p = .303$. However, a regression analysis provided evidence that threat level predicted the vicarious boomerang restoration score ($B = .35$, $SE = .17$), $t(137) = 2.08$, $p < .039$, $R^2 = .03$ (path *c*). We therefore followed the procedure for this effect. Step 2 of the analysis is to check for a positive relationship between threat and psychological reactance (i.e., the combined score of feelings of anger and negative cognitions). Results showed that this was indeed the case ($B = 8.16$, $SE = 1.60$), $t(137) = 5.09$, $p < .001$, $R^2 = .16$ (path *a*). Subsequently, we checked whether the mediator affected the outcome. Results suggested that when entering vicarious boomerang restoration as the dependent variable, and entering psychological reactance and threat as the two independent variables, the outcome of psychological reactance on vicarious boomerang restoration was affected ($B = .02$, $SE = .01$), $t(137) = 2.47$, $p = .015$, $R^2 = .07$ (path *b*). This analysis further showed that the effect of threat on vicarious boomerang restoration became non-significant ($B = .17$, $SE = .18$), $t(137) = .96$, $p = .340$, $R^2 = .07$ (path *c*'). Because the effect of threat on vicarious boomerang restoration became almost zero (i.e., $B = .17$), this suggested a full mediation of psychological reactance on the relationship between threat and vicarious boomerang restoration. A Sobel test confirmed this suggestion, $z = 2.20$, $SE = .08$, $p = .028$. In short, it seems that psychological reactance served as a mediator in the relationship between threat and vicarious boomerang restoration.

Figure 2.3. Mediation analyses psychological reactance on the relationship between threat level and restoration intentions.



Note: The numbers in the figure are the standardized regression coefficients and standard errors of the paths. The italicized text is the relationship between threat level and restoration intentions, without controlling for psychological reactance. *$p < .05$, ***$p < .001$.

Discussion

The current research was, to our knowledge, the first to investigate whether humans would experience psychological reactance towards an artificial agent. According to Reeves and Nass (1996) humans respond to artificial agents as if they were responding to other humans. Therefore, we suggested that humans could also experience psychological reactance when an artificial agent threatened their autonomy. In line with previous research on reactance in human-human interaction (e.g., Dillard & Shen, 2005), we found that participants experienced more psychological reactance when reading a high-threatening instruction text than when reading a low- or non-threatening instruction text. Another explanation for our results could be that participants did not experience threat, but instead reacted negatively because they did not like the artificial agent. We did measure participants' liking towards the artificial agent. However, we did not find an effect of liking on our results. That is, liking did not differ for participants in the low-threat condition, no-threat condition or high threat condition. Moreover, results indicated that psychological reactance was especially pronounced on the reactance measure of feelings of anger, but also present on the reactance measure of negative cognitions. This pattern was also found in a recent study (Quick & Consodine, 2008) that showed stronger effects related to psychological reactance on measures of anger than on measures of negative cognitions. We furthermore found that participants who experienced reactance had the intention to restore their autonomy by being around others who used extra energy while doing the laundry (vicarious boomerang intention). A Sobel test showed a full mediation of psychological reactance on the relationship between threat and vicarious boomerang restoration. In other words, when participants feel threatened in their autonomy to consume energy, they experience reactance, which results in intentions to be around others who consume extra energy while doing their laundry.

Finally, we had expected that participants would experience more reactance when they read a high-threatening instruction text stemming from an artificial agent with a high level of social agency. Results did not provide evidence for this interaction. This is in contrast

with previous theories about multiple social cues (e.g., Biocca, et al., 2003; Daft & Lengel, 1986; Mayer, et al., 2003; Short, Williams, & Christie, 1976), that suggested that if more social cues were present, humans would respond more like they would in a real social interaction. A possible explanation for this could be that the social cues of the social agent were not strong enough. Perhaps a social agent that interacts with a person provides more social cues (Nass & Moon, 2000) that leads to more social behavior. Future research could examine whether the occurrence of psychological reactance perseveres or even accumulates when humans interact with a social agent in a task that is more interactive in nature. Another explanation could be that participants in the text-only condition imagined a human actor or the experimenter when reading the text. If this would be the case then there would exist a possibility that participants would show higher reactance towards the text-only condition, because a human actor is more social than an artificial agent with social cues. However, we found the opposite. Also, research by Reeves and Nass (1996) showed that participants did not imagine an experimenter when reading a text.

Although we did find that participants became psychologically reactant when confronted with a high-threatening agent, we did not measure participants' prior intentions for energy conservation. It could be that our sample mainly consisted of participants who did not care or did not want to save energy and therefore became reactant. That is, we argue that if the goals of the artificial agent were incongruent with the goals of a participant, this participant may have felt forced in a direction opposite to his or her own personal goals. It could even be that participants got the impression that the artificial agent was working against them, and felt more threatened leading to more psychological reactance.

However, according to psychological reactance theory (Brehm, 1966), a threat-to-autonomy message will always cause psychological reactance, regardless of prior intentions. A good example of this is children in their puberty: They will do anything that is the opposite of what you propose, just to demonstrate their autonomy, even if you propose something they really would like (Burgoon, et al., 2002; Grandpre, et al., 2003).

Still, other research suggested that motivational states and goals may make humans more prone to respond more reactant. For example, a more recent study by Silvia (2005) suggested that when a person had the same goal intentions as the persuader, psychological reactance did not occur, even when high-threatening language was used. Whereas when a person had different goal intentions, psychological reactance occurred most strongly for high-threatening language (Silvia, 2005).

Therefore, we investigated whether an overlap between the intentions of the persuader and of the person being persuaded is relevant for the occurrence of reactance caused by threatening language. That is, would an artificial agent using threatening language lead to reactance independent of goal overlap like Brehm (1966) proposed? Or, would humans experience even more psychological reactance when they feel that the goals of the artificial agent are incongruent with their own goals like Silvia (2005) found more recently? In Study 2, we investigated these contradictory expectations.

Study 2

In Study 2, participants had to program a virtual washing machine. Next to the computer an iCat (i.e., robotic agent, see Picture 2.3) was placed that advised participants about the programming choices. The advice consisted of low-threatening ("You could…") language or high-threatening ("You have to…") language. In line with Study 1, we expected that participants would report more psychological reactance for high-threatening language than low-threatening language. We also explored the role of goal congruency. Therefore, we manipulated goal overlap by asking the participants to rank their own washing goals (e.g., a participant may have ranked clean laundry above energy conservation), and then indicating that the iCat either preferred the same washing goals (e.g., likewise ranked clean laundry above energy conservation; congruent goals condition) or that the iCat preferred other washing goals (e.g., ranked energy conservation above clean laundry; incongruent goal condition). In line with Silvia (2005) that studied similarity in human-human interaction, we expected that the highest level of psychological reactance would be observed when

participants interacted with an iCat that they knew had incongruent goals and that used high-threatening language.

Method

Participants and Design

We recruited 79 participants[11] (46 males, 33 females; age *M* = 20.06*, SD* = 2.00). Participants were mainly first-year undergraduates at Eindhoven University of Technology. Participants were randomly assigned to a 2 (threat level: low threat vs. high threat) x 2 (goal congruency level: congruent goal vs. incongruent goal) between-subjects design. The dependent variables were the negative cognitions score and the feelings of anger score (i.e., the predictors of psychological reactance; Dillard & Shen, 2005). The experiment lasted about 30 minutes, for which participants were paid €7.50 (approximately $11.25 at the time this experiment was conducted).

Materials

The materials were the same as in Study 1, with the following exceptions.

Feelings of anger. This measure was based on the anger questionnaire of (Dillard & Shen, 2005), and on the Dutch State-Trait Anger Expression Inventory (STAXI) scale (Van der Ploeg, Van Buuren, & Van Brummelen, 1988). Statements were presented such as "I felt irritated during the task," "I had the feeling I had to hit something during the task," and "I felt angry during the task". The mean score of the feelings of anger questionnaire formed a reliable scale (α = .83) and was labeled as "feelings of anger score".

Negative cognitions. The same thought-listing task was used as in Study 1. Additionally, participants were asked to report the frequency of every thought. As in Study 1, the percentage of the negative thoughts in the thought-listing task was labeled as "negative cognitions score".

[11] In total we recruited 80 participants. However, because of technical failures, data of 1 participant got lost. Analyses are done on the data of the remaining 79 participants.

Perceived threat to autonomy. This measure was based on the perceived threat to autonomy measure used in Dillard & Shen (2005). Statements were, for example: “I felt free to choose the way I wanted to choose,” and “I had the feeling that somebody tried to make a decision for me”. The mean score of the perceived threat-to-autonomy questionnaire formed a reliable scale (α = .85) and was labeled as “threat score”.

Perceived goal congruency. The goal congruence questionnaire assessed whether participants perceived the displayed goals of the artificial agent (i.e., the iCat called Femke) as either congruent (the same goals as the participant) or incongruent (different goals as the participant). Statements were for example: “Femke wanted to reach the same goals as I did”, and “Femke tried to hinder me in reaching my goals” [reversed coding]. The mean score of the perceived goal congruency questionnaire formed a reliable scale (α = .83) and was labeled as “goal congruency score”.

Negative evaluations of the agent. The agent (i.e., Femke) was evaluated by two items. The statements were: “Femke was an expert at doing the laundry” [reversed coding], and “Femke was friendly” [reversed coding]. The two questions correlated fairly well (R = .33, p < .005) and were combined to form the “negative evaluations score”.

Restoration thoughts. To measure restoration thoughts (i.e., cognitions about restoring the feeling of autonomy) two questions were asked: “Often, I had the tendency to do just the opposite of what Femke recommended,” and “Often, I deliberately tried to ignore the advice given by Femke”. The two questions correlated fairly high (Pearson’s r = .58, p < .001) and were combined to form the “restoration thoughts score”.

Restoration behavior. We considered participant’s restoration behavior (i.e., actual behavior to restore the feeling of autonomy) as refraining from the proposed behavior after receiving an advice. In the washing trials, participants were asked to make programming choices. After each programming choice, participants were provided with advice from the iCat. We checked whether participants adjusted their score after receiving the advice. We calculated a difference score by taking the initial choice the participant made which was

subtracted of the final choice the participant made after he/she received the advice of the iCat.

Procedure

Participants were invited to participate in a study about technology and interaction. When arriving at the laboratory, participants were seated behind a desk top computer. All instructions were presented on the computer screen. The participants were shown an introduction about programming a virtual washing machine (see Figure 2.4).

Figure 2.4. Virtual washing machine panel.

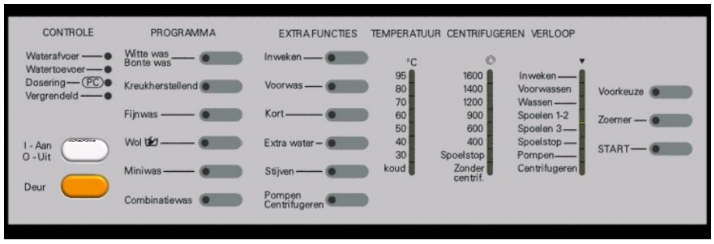


Figure 2.5. The iCat (Femke).

Note: At the washing machine panel the labels say “Washing program”, “Extra functions”, “Temperature”, “Spinning speed”.

Next, participants were introduced to the iCat, called Femke (see Figure 2.5) and were told that Femke would give advice on their programming choices. After the introduction of the task, participants were presented with two opposing goals that humans can have during washing their laundry; “A clean laundry,” and “Energy conservation.” They were asked to rank these two goals on how important they thought the goals were to them. After stating their preference, participants were told that Femke also had made a preference ranking. In the congruent goal condition, participants were told that Femke had made the same ranking (e.g., when participants preferred a clean laundry, Femke also preferred a clean laundry). Then, participants could push a button and Femke introduced itself. Femke stated its preference ranking again and explicated that this ranking was the same as the participant’s ranking. In the incongruent goal condition, participants were told that Femke had made the opposite ranking (e.g., when participants preferred a clean laundry, Femke preferred energy conservation). After pushing the button, Femke introduced itself and stated again its preferences and explicated that this ranking was the opposite as the participant’s ranking.

Participants were reminded that they were not obliged to follow Femke's advice, and that the final programming choices were their *own choices*. Next, participants had to program the virtual washing machine (1 practice trial and 10 experimental trials). In the low threat condition, participants received advice from Femke consisting of low-threatening language (e.g., "You could set the temperature to 40 C°). In the high threat condition, participants received advice from Femke consisting of high threatening language (e.g., "You have to set the temperature to 40 C°). Before the experimental trials started, Femke again explained its preference ranking. Important to note is that the advice from Femke was (apart from the low-threatening language use or high-threatening language use) exactly the same for the congruent goal as for the incongruent goal conditions. After completing the 10 trials, participants were asked to answer some questionnaires. Finally, participants were thanked, paid € 7.50 (approximately $ 11.25 at the time this experiment was conducted), debriefed, and dismissed.

Results

Manipulation Check of Threat Level

An analysis of the effect of threat level on threat score suggested that our threat level manipulation was successful. A One-Way Analysis of Variance (ANOVA) with threat level as the independent variable and threat score as the dependent variable showed a significant effect of our threat manipulation, $F(1, 77) = 16.54$, $p < .001$. That is, participants in the high threat condition reported more perceived threat ($M = 3.52$, $SD = .56$) than participants in the low threat condition ($M = 2.98$, $SD = .62$).

Manipulation Check of Goal Congruency Level

An analysis of the effect of goal congruency level on goal congruency score suggested that our goal congruency level manipulation was successful. A One-Way ANOVA with goal congruency level as the independent variable and goal congruency score as the dependent variable suggested that participants in the congruent goal condition perceived

Femke's (i.e., the iCat) goals as more congruent (M = 3.79, SD = .79) than participants in the incongruent goal condition (M = 2.83, SD = .79), $F(1, 77) = 33.10$, $p < .001$.

Threat and Goal Congruency

As in Study 1, we tested whether participants reported more psychological reactance when they received high-threatening advice compared to low-threatening advice. The negative cognitions score and the feelings of anger score[12] were submitted to a 2 (psychological reactance measure: negative cognitions vs. feelings of anger) x 2 (threat level: low threat vs. high threat) x 2 (goal congruency level: congruent goal vs. incongruent goal) Repeated Measures ANOVA[13], with the first factor serving as a within-subjects factor. Replicating Study 1, we found that participants in the high threat condition experienced more psychological reactance (M = 25.76, SD = 13.34) than participants in the low threat condition (M = 15.40, SD = 13.38), $F(1, 75) = 12.06$, $p = .001$. Furthermore, as in Study 1, the effect of threat level was qualified by an interaction of Threat Level X Psychological Reactance Measure, $F(1, 75) = 11.69$, $p = .001$, $n_p^2 = .14$, indicating that the effect of threat level was different on the measure of feelings of anger than on the measure of negative cognitions. Closer examination of the simple effects of threat level on feelings of anger and on negative cognitions showed that the effect of threat level was significant for both measures, but, there was a stronger effect size of threat level on the measure of negative cognitions, $F(1, 75) = 1.88$, $p = .001$, $n_p^2 = .14$, than of threat level on the measure of feelings of anger, $F(1, 75) = 5.22$, $p = .025$, $n_p^2 = .07$. Although the correlation between negative cognitions and feelings of anger was moderately high (Pearson's $r = .36$, $p = .001$) we found a different effect on the two types of psychological reactance measures. Notably, in Study 1 we found the reversed effect; stronger effect size for feelings of anger than for negative cognitions.

[12] When analyzing the data we discovered two extreme outliers (i.e., on Mahalanobis distance), which had an effect on the feelings of anger score. However, when excluding the outliers the effects remained the same in all analyses and therefore we decided to present the data analyses of all participants.
[13] We did not find a main effect of goal congruency, $F(1, 75) = .66$, $p = .42$.

We also expected that an artificial agent that had an incongruent goal would lead to the highest level of psychological reactance. However, we did not find the expected interaction effect of Threat level x Goal congruency level, $F(1, 75) = .31$, $p = .58$. In other words, goal congruency did not lead to an increasing effect on psychological reactance. This could suggest that participants that felt threatened experienced psychological reactance, even if they and the iCat wanted to accomplish the same goals, which is in line with Brehm (1966).

Restoration

In accordance with previous research on psychological reactance, we expected a positive relationship between threat level and the three types of restoration measures (restoration behavior, restoration thoughts, and negative evaluations towards the agent), and that this relationship would be mediated by psychological reactance (e.g., Quick & Stephenson, 2007a). To analyze whether psychological reactance mediated the relationship between threat level and restoration measures we used the method proposed by Baron and Kenny (1986). We decided to do three separate mediation analyses. Step 1 of the analysis by Baron and Kenny (1986) is to check for a positive relationship of threat level with restoration. First of all, results showed that there was no positive relationship between threat and restoration behavior[14], $t(78) = 1.24$, $p = .218$[15]. Therefore, no mediation analysis was possible.

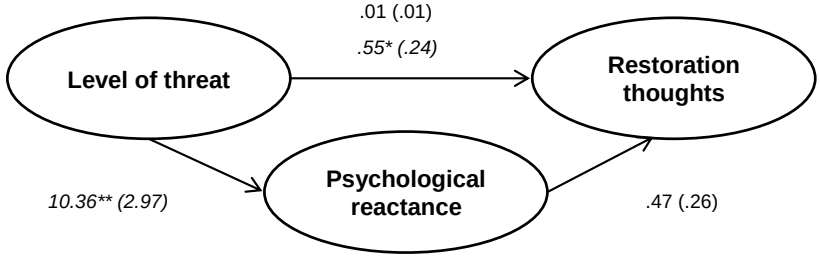
Second, the effect of threat on restoration thoughts was significant ($B = .55$, $SE = .24$), $t(78) = 2.33$, $p = .022$, $R^2 = .07$ (path *c*; see Figure 2.6). Step 2 of the analysis is to check for a positive relationship between threat and psychological reactance (i.e., the combined score of feelings of anger and negative cognitions). Results showed that this was indeed the case ($B = 10.36$, $SE = 2.97$), $t(78) = 3.49$, $p = .001$, $R^2 = .37$ (path *a*). Subsequently, we checked whether the mediator affected the outcome. Results suggest that

[14] We also explore compliance towards the agent. Unfortunately, we did not find an effect of compliance on participant's responses, $F < 1$, $p > .05$. This result is supported by the null results on the effect of threat on restoration behavior.

[15] Although there was no positive relationship between threat and restoration behavior, restoration behavior was correlated to negative cognitions (Pearson's $r = .26$, $p = .019$) and feelings of anger (Pearson's $R = .25$, $p = .028$).

when entering restoration thoughts as the dependent variable, and entering psychological reactance and threat as the two independent variables, the outcome of psychological reactance on restoration thoughts became marginally significant (*B* = .47, *SE* = .26), *t*(78*)* = 1.84, *p* = .070, R^2 = .08 (path *b*). This analysis further showed that the effect of threat on restoration thoughts also became non-significant (*B* = .01, *SE* = .01), *t*(78*)* = .88, *p* = .385, R^2 = .07 (path *c*'). These results suggest that psychological reactance did not mediate the relationship between threat and restoration thoughts, but that threat caused both psychological reactance and restoration thoughts. This was confirmed by a Sobel test, z = .86, *SE* = .10, *p* = .389.

Figure 2.6. Mediation analyses psychological reactance on the relationship between threat and restoration thoughts.

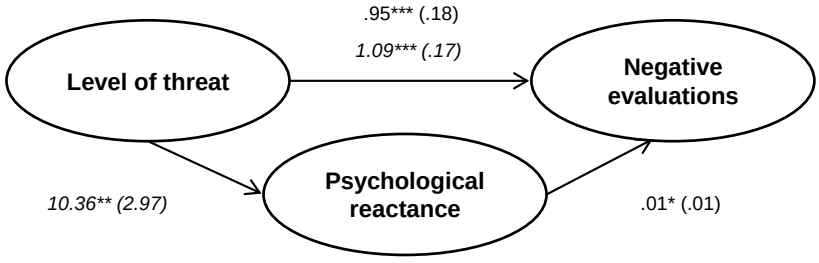


Note: The standardized regression coefficients and standard errors of the paths. The italicized text is the relationship between threat level and restoration intentions, without controlling for psychological reactance. *p < .05, **p < .01.

Finally, we checked whether the relationship between threat and negative evaluations of the agent was mediated by psychological reactance (see Figure 2.7). Results showed that the effect of threat on negative evaluations was significant (*B* = 1.09, *SE* = .17), *t*(76) = 6.50, *p* < .001, R^2 = .36 (path *c*). Step 2 of the analysis is to check for a positive relationship between threat and psychological reactance (i.e., the combined score of feelings of anger and negative cognitions). Results showed that this was indeed the case (*B* = 10.36, *SE* = 2.97), *t*(78) = 3.49, *p* = .001, R^2 = .37 (path *a*). Subsequently, we checked whether the mediator affected the outcome. Results suggest that when entering negative evaluations as the dependent variable, and entering psychological reactance and threat as the two independent variables, the outcome of psychological reactance on negative evaluations was

affected (B = .01, SE = .01), t(78) = 2.23, p = .028, R^2 = .39 (path b). This analysis further showed that the effect of threat on negative evaluations was slightly diminished but remained highly significant, B = .95, SE = .18, t(78) = 5.36, p < .000, R^2 = .39 (path c'). These results suggest that psychological reactance partially mediated the relationship between threat and negative evaluations. This was confirmed by a Sobel test, z = 1.94, SE = .07 p = .052.

Figure 2.7. Mediation analyses psychological reactance on the relationship between threat level and negative evaluations



Note: The standardized regression coefficients and standard errors of the paths. The italicized text is the relationship between threat level and restoration intentions, without controlling for psychological reactance. *p < .05, **p < .01, ***p < .001.

Discussion

Replicating Study 1, we found that participants reported more psychological reactance when interacting with a robot that gave high-threatening advice than when interacting with a robot that gave low-threatening advice. In contrast to what we found in Study 1, the effect was more pronounced on negative cognitions than on feelings of anger. An explanation for this result could be that we used slightly different measures for negative cognitions as well as for feelings of anger. In contrast to expectations, we did not find an increase in feelings of psychological reactance when participants interacted with a high-threatening robot with incongruent goals (compared to the participants' goals). A potential explanation for this is that participants may not have remembered the goals of the robot. However, on our manipulation check (located at the end of the experiment) participants reported correctly that either the goals of the robot were the same as or opposite to their own

goals. Another reason could be that the behavior of the robot was not in line with its goals. To exclude possible confounding of behavior, we kept the behavior exactly the same for both goal congruency conditions. This could have led to inconsistent behavior of the robot with its goals. Because we did not check this possibility, we cannot exclude that this is what happened.

Still, our findings seem to provide evidence in line with psychological reactance theory (Brehm, 1966) which proposed that reactance will occur independent of goal overlap. The current study was the first to test this proposal. The results seemed to suggest that it may simply not matter whether someone has the same or opposite goals. That is, when someone feels threatened in his/her autonomy, psychological reactance is experienced anyway (see also e.g., Grandpre, et al., 2003). Thus designers should take into account that even users that are willing to change their behavior in a desired way, could become reactant by a persuasive agent.

We also investigated whether psychological reactance mediated the relationship between threat and restoration behavior, restoration thoughts and negative evaluations against the robot. Although, we found that restoration behavior was significantly correlated with both feelings of anger and negative thoughts, we did not find a relationship between threat and restoration behavior. Next, we investigated whether psychological reactance mediated the relationship between threat and restoration thoughts during the task. We found that participants reported more restoration thoughts, as a result of a higher threat-to-autonomy message. That is, participants showed a higher tendency to do the opposite of what the advice-giving robot said and a higher tendency to ignore the advice completely. Results indicated that the relationship between threat and restoration thoughts was not mediated by psychological reactance. This suggests that a threatening message can trigger psychological reactance as well as restoration thoughts directly. Finally, we investigated whether psychological reactance mediated the relationship between threat and negative evaluations. We found that participants gave more negative evaluations of the robot, as a result of a higher threat-to-autonomy message. That is, they attributed a lower expertise to

the high-threatening advice-giving agent, and evaluated it as less friendly than the low-threatening advice-giving agent. For negative evaluations, it seemed that there was a partial mediation of psychological reactance. That is, the effect of threat diminished when including psychological reactance as a predictor of negative evaluations, but remained significant. This suggests that a threatening message led participants to experience psychological reactance and consequently evaluate the advice-giving robot more negatively. In summary, although participants did not show restoration behavior, they did show restoration intentions and negative feelings towards the persuasive agent. It could be only a matter of time before they also show restoration behavior. Designers could try to prevent this by designing the persuasive agent in such a way that it does not trigger psychological reactance and possible restoration behavior.

General Discussion

As explained in the introduction, every act of persuasion—and therefore also persuasion through persuasive technology—has the possibility of triggering psychological reactance (Brehm, 1966; Brehm & Brehm, 1981). Trying to change human behavior (or attitudes) may feel as a threat to autonomy. In other words, the *feeling* of threat to autonomy might be enough to cause humans to experience psychological reactance and consequently evoke behavior that is in contrast to the desired behavior. This means that every persuasive technology could suffer from the effects of psychological reactance, which could lead to opposite behavior or even rejection of the persuasive technology. In line with previous work (e.g., Buller, et al., 1998; Dillard & Shen, 2005; Grandpre, et al., 2003; Miller, et al., 2007; Quick & Consodine, 2008; Quick & Stephenson, 2007a; Quick & Stephenson, 2008; Rains & Turner, 2007; Reinhart, et al., 2007) that investigated psychological reactance in human-human interaction, we found in two experiments that participants also experienced psychological reactance when they were confronted with a persuasive agent. In Study 2, we even found that participants experienced reactance in a human-agent interaction. This was seen in higher levels of negative cognitions and feelings of anger (i.e., the two indicators of

psychological reactance). In addition, we found that threatening language did not only cause psychological reactance, but also restoration thoughts. Moreover, we found that psychological reactance mediated the relationship between perceived threat and restoration thoughts, and partially mediated the relationship between perceived threat and negative evaluations of the artificial agent. That is, when humans experienced a threat to their autonomy, they showed intentions to restore their autonomy and evaluated the artificial agent as more negative, which was mediated by their feelings of psychological reactance. In other words, humans first experienced threat, they then experienced psychological reactance, and in turn they showed restoration thoughts and evaluated the artificial agent more negatively.

Earlier research suggested that more social cues would lead to stronger social responses (e.g., Blascovich, 2002; Daft & Lengel, 1986; Guadagno et al., 2007; Louwerse, et al., 2005; Mayer, et al., 2003). To test this, we manipulated the degree of social agency by using a degree of social cues. In Study 1, we investigated whether the social agency that participants ascribe to an artificial social agent might increase the social reactance that participants experience towards that agent when it attempts to persuade too strongly (by using high-threatening language). We did not find that a higher level of social agency of the artificial agent led to even more psychological reactance. Furthermore, in Study 2 we manipulated the intentionality of the robotic agent by manipulating its goal congruency with the participant. In other words, we manipulated whether the robotic agent had a congruent or incongruent goal with the participant. Brehm (1966) proposed that an artificial agent using threatening language would lead to reactance independent of goal overlap. However, more recent work of Silvia (2005) suggested that participants experience more psychological reactance when they feel that the goals of the artificial agent or incongruent with their own goals like. Our results suggested that it did not matter whether a robot tried to work against the participant by giving advice to accomplish its own goals, or tried to help by giving the participant advice to accomplish the participant's goals. Thus in line with Brehm (1966), we

found that threat appeared to be the dominant factor of causing psychological reactance, irrespective of goal congruency between the robotic agent and the participant.

Coming back to John and his robot (example in introduction), we would suggest that the designers of the robot adjust their messages in such a way that John does not experience them as a threat to his autonomy (by using low-threatening concrete messages). Otherwise the effect could backfire, leading John to consume even more energy, regardless of his own intention to conserve energy. Furthermore, designers could pretest their robot on psychological reactance by using the measures of anger and negative cognitions (and possibly restoration) before exposing real users to the robot.

In this chapter we observed participants responding to an artificial agent as if it was a social entity in the domain of negative behavior (i.e., psychological reactance). Our results implied that these social responses were of an automatic nature. Although Reeves and Nass (1996) suggested that humans show automatic social behavior, more research is needed to empirically test this. If humans indeed responded automatically social to artificial agents, this would mean that it could be quite easy for designers to trigger social responses in users. However, it is not yet clear why humans respond socially to something they know does not require social treatment. In other words, what are the underlying processes of this apparent automatic social behavior to artificial agents? In the next chapter we will empirically disentangle the automatic and controlled responses of humans when responding to artificial agents and investigate the underlying processes of these different responses.

Chapter 3

Humans, Robots and Inanimate Objects:

The Influence of Actor Type on Spontaneous and Intentional Trait Inferences

---

* This chapter is partly based on:

Roubroeks, M., Ham, J., & Midden, C. (2012). Brutale Mensen, Robots, en Objecten: De Invloed van het Type Actor op Spontane en Intentionele Gevolgtrekkingen. In N. van de Ven, M. Baas, L. van Dillen, D. Lakens, A.M. Lokhorst, & M. Strick (Eds.), *Jaarboek Sociale Psychologie 2011* (pp. 189-193). Groningen: ASPO pers.

Imagine a day in the future on which John calls a robot company and asks whether their programmers could program his household robot to be a little bit friendlier. After his robot has been reprogrammed, John connects it again to his smart meter. He notices that the robot is not only friendlier when communicating about John's energy consumption, but is also friendly to John's friends when they visit; "You have nice new shoes!", "Cool haircut, it suits you", and "Funny joke! You are so much fun". For a split-second John thinks the robot has a friendly personality. However, when thinking about it some more, he figures it is just a piece of technology that has been programmed to show friendly behavior.

In line with this example, in the current chapter, we argue that in their quick and uncontrolled (automatic) judgments humans may infer human traits, based on the behavior of many things, such as this robot. In contrast, we argue that in their more slow and controlled judgments humans will take the actor type (like humans, artificial agents, or inanimate objects) into account when making social trait inferences. If we know more about people's automatic and controlled responses, then designers could decide in what way they would like their robot to be exposed to humans. For example, in a more automatic way when they would like to trigger helping behavior towards a friendly robot. Or, in a more controlled way to get a clearer answer on how people view the robot at first sight, without interacting with it.

Social Responses to Technology

Indeed, earlier research has shown that humans exhibit social responses when they interact with a computer (e.g., Fogg & Nass, 1997a; Fogg & Nass, 1997b; Nass & Moon, 2000; Nass, et al., 1999; Reeves & Nass, 1996). According to the Media Equation hypothesis (Reeves & Nass, 1996), humans exhibit social responses to computers when triggered by certain social cues. These responses are described in the Computer As Social Agents (CASA) paradigm (Reeves & Nass, 1996). This paradigm proposed that participants respond to computers as if they are social actors, while at the same time participants know that these responses are inappropriate (i.e., participants know they do not have to respond socially to a computer) and treat the computers as the objects they are. Numerous studies

have supported this notion (for an overview see Reeves & Nass, 1996). For example, it has been shown that participants respond more politely when a computer asked participants to directly judge the computer's own performance (at the same computer), than when the computer's performance was assessed by a paper-and-pencil questionnaire (Nass, et al., 1999). Such polite responses in direct interactions were also found in human-human interaction (e.g., Finkel, Guterbock, & Borg, 1991). However, when asked about their social responses to the computer, participants denied behaving socially and even felt offended by the question (Reeves & Nass, 1996). In other words, automatic responses (behavior when interacting with computers) and controlled responses (behavior when directly asked about computers) seem to differ from each other. But why would humans respond in a human-like way to something they *know* is not human?

Nass (2004) suggested that these human-like responses can be explained by an evolutionary account. He argued that humans have evolved in a world in which other humans were associated with gaining opportunities and getting help when problems arose, and that "*…there would be a significant evolutionary advantage to the rule: If there's even a low probability that it's human-like, assume it's human-like"* (Nass, 2004, p. 37). In other words, if an artificial agent "behaves" like a human, then it will be assumed that it is human-like and act accordingly (Nass, 2004; Reeves & Nass, 1996).

Automatic versus Controlled Social Responses

In their research (Nass & Moon, 2000; see also Reeves & Nass, 1996) Nass and colleagues referred to the theory of mindfulness/mindlessness by Langer (1992; see also Johnson & Gardner, 2007; Johnson & Gardner, 2009; Johnson, et al., 2004; Lee, 2008) for explaining these apparently automatic, social responses. Langer (1992) described the process of mindlessness as "*…being in a state of mind characterized by an overreliance on categories and distinctions drawn in the past and in which the individual is context-dependent and, as such, is oblivious to novel (or simply alternative) aspects of the situation*" (p. 289). In other words, humans who are interacting with artificial agents find themselves in

a “mindless state” in which they only attend to the social cues, but seem to (temporarily) ignore that the artificial agents are just a piece of technology.

However, to our knowledge, earlier research (e.g., Reeves & Nass, 1996) never directly tested the hypothesis that human initial responses to technology are indeed *automatically social*. We suggest that a mindless state can be compared to a form of automatic behavior (see Bargh; 1984; Bargh, 1990) that is used in the research area of implicit social cognition. This research area makes use of well-established theoretical and methodological tools to investigate participants’ automatic and controlled behavior. With these tools we can extend earlier research (see e.g., Reeves & Nass, 1996) because we are now able to directly compare participants’ automatic and controlled social responses to artificial agents. Furthermore, although Reeves and Nass (1996) stated that humans respond the same to humans as to artificial agents. However, to our knowledge, they never directly compared participants’ responses to humans with their responses to artificial agents. In our research, we extend the claims of Reeves and Nass (1996) by directly comparing human responses to humans, artificial agents, but also to inanimate objects (see next paragraph).

We predicted that in their more spontaneous, automatic responses, participants respond in social ways to an actor’s behavior, independent of whether that actor is a human, an artificial agent or even a completely inanimate object (e.g., a curtain or a rock). We argue that participants automatically respond in social ways when certain social cues are available, either in characteristics of the actor performing the behavior, or in the behavior itself. Humans differ from artificial agents and inanimate objects in that they are real social actors that possess personality traits and mental states (like emotions) that can be inferred from their behavior (see also Chapter 1). In other words, humans are social entities with their own intentionality, but artificial agents can only give the experience of being a social actor without really having intentionality. However, like humans, artificial agents are able to display social cues (e.g., Blascovich, 2002; Fogg, 2003) that could lead to the (temporal) experience of being social actors. This is unlike inanimate objects; a human or an artificial agent may

display a friendly face, but a rock cannot display such cues. Still, also the behavior an actor performs might contain social cues. Thereby, not only a human actor or an artificial agent might trigger social cues (e.g., by behaving aggressively), but also a completely inanimate object (e.g., a curtain) may trigger a social cue through its behavior (e.g., by hitting a person in the face). This is in line with the theory of anthropomorphism that stated that humans anthropomorphize non-human things. That is, humans attribute human-like properties (like intentionality) or mental states (like emotions) to artificial agents and inanimate objects (e.g., Epley, et al., 2008; Epley, et al., 2007; Waytz, et al., 2010). In summary, humans, as well as artificial agents and inanimate objects can exhibit social cues and therefore we expected that this would lead to (temporarily) automatic social responses to all these different types of actors.

Furthermore, also based on the presented insights into automatic versus controlled human responses, we predicted that in their more intentional, controlled responses, participants only respond in social ways to humans, less to artificial agents and even less to completely inanimate objects[16]. Because when humans think more extensively about artificial agents or inanimate objects, they become aware that they are interacting with a piece of technology or object that does not has human personality traits, nor mental states (like emotions). Still, we expected that artificial agents would not completely be categorized as inanimate objects. Rather, we expected that artificial agents would be seen as a category in between humans and inanimate objects. This is in line with robotic psychology that "studies the psychological significance of robots' behavior and its intertwining with elements of physical and social environments" (p. 295, Libin & Libin, 2004). Researchers found that participants' perceptions were different for robots compared to animals, humans, or inanimate objects (Libin & Libin, 2004). Also, there is no agreement about the definition for the term 'robot'. Even the father of robotics, Joseph Engelberger, once mentioned: "*I can't*

[16] We believe that traits might be inferred to objects because of the natural language we use. For example, when participants read the sentence "The pointer helps the teacher to clarify his presentation" they could infer the trait "helpful". This does not mean that they think the pointer has a helpful personality or has the intentionality to help the teacher, but that the pointer is a "helpful tool" that is used by the teacher.

define a robot, but I know one when I see one" (CBC, 2007). Furthermore, this is in line with monster theory. Monster theory uses the metaphor of a monster, which stems from Frankenstein's monster that existed partly of human and partly of machine, and therefore is hard to categorize (Smits, 2006). In monster theory it is stated that like a monster, new technologies (here, artificial agents) simultaneously fit into two categories (here, humans and objects) that are considered as mutually exclusive (Smits, 2006). That is, artificial agents fit in the category "humans", but also in the category "objects". However, these categories are considered to be mutually exclusive (i.e., one cannot be a human as well as an object). In other words, humans find it hard to explicitly categorize artificial agents, because humans know they are only inanimate objects, but at the same time they can experience them to possess human-like properties. For example, when interacting with a smiling robot, participants observe technical cues (e.g., mechanical sounds), but also human cues (e.g., smiling when the participant approaches the robot). This could result in feelings of both fear and fascination. We argue that a way to deal with this is to create a whole new category (similar to monster assimilation, discussed in Smits 2006). In summary, we expected that at a controlled level, humans would categorize artificial agents as a category in between humans and inanimate objects. Therefore, we expected they would show the most social response to humans, less social responses to artificial agents and the least social responses to inanimate objects.

Spontaneous and Intentional Trait Inferences to Artificial Agents

So, we predicted that, in their automatic responses, humans might respond in social ways to other humans, artificial agents, and inanimate objects, whereas in their controlled responses their responses will be influenced by actor type. To test this claim we investigated how humans infer traits from overt behaviors. Therefore, we used a well-established paradigm from the person perception literature in human-human interaction. In this paradigm we will investigate whether an automatic social behavior (labeled spontaneous trait inferences, STIs) and its more controlled counterpart (labeled intentional trait inferences, ITIs) will also occur in human-agent interaction (Uleman, et al., 1996). Furthermore, we

investigate how these two types of trait inferences depend on the type of actor of the behavior (human vs. artificial agent vs. inanimate object). Finally, we investigated a possible underlying mechanism of the automatic social responses to the different types of actors.

A variety of studies in human-human interaction indicated that STIs are drawn when a person is attending to another person's behavior and unintentionally infers a trait that could be implied by the observed behavior (e.g., Uleman, et al., 1996; Uleman, et al., 2008). For example, when you see a man quarreling with a woman, and the man constantly raises his voice and keeps hitting the table with his fist, you probably spontaneously infer that the man is aggressive. In other words, humans extract traits out of behaviors without the intention to do so. While STIs are automatic, their controlled counterparts are intentional trait inferences (ITIs).

Also, a variety of studies in human-human interaction indicated that ITIs are drawn when a person is forming an impression of another person and intentionally infers a trait that could be implied by the observed behavior of that person. When measuring ITIs humans are usually asked to intentionally attribute traits to a person (Uleman, et al., 1996; Uleman, et al., 2008). Many earlier studies suggested that humans can infer traits intentionally and spontaneously when confronted with human behavior (for an overview see Uleman, et al., 1996; Uleman, et al., 2008).

In the current research, we aim to demonstrate that participants respond automatically socially to humans, artificial agents and inanimate objects. Furthermore we aim to demonstrate that, at a controlled level, participants distinguish between humans, artificial agents and objects. Finally, we will investigate a possible underlying mechanism for the automatic social responses to artificial agents. In other words, participants will attribute human personality traits to artificial agents (and inanimate objects) when they observe social behavior cues at a spontaneous (more automatic) level. But, when participants observe social behavior cues on an intentional (more controlled) level, they will attribute traits to a lesser extent to artificial agents (and even less to inanimate objects). With this behavior (STIs) we can directly test whether participants' initial responses are indeed spontaneous,

automatic social responses. In addition, we can compare these responses with their intentional, controlled responses (ITIs) towards artificial agents.

To assess STIs, we used a well-established research paradigm often used in earlier research: the relearning paradigm (Carlston, & Skowronski, 1994; Carlston, Skowronski, & Sparks, 1995). The relearning paradigm is an indirect memory test in which participants can draw STIs (implicitly) without being asked to form impressions. The relearning paradigm is based on the notion that humans have a better memory for things that have been repeated (relearned) than for things that have not been repeated (learned once; e.g., Ham & Vonk, 2011). In the relearning paradigm (Carlston & Skowronski, 1994; Carlston, et al., 1995), participants were exposed to a booklet that contained photos of persons paired with trait-implying behavioral statements. For example, participants saw a picture of John that was paired with the following statement "I hate animals. Today I was walking to the pool hall and I saw this puppy. So I kicked it out of my way" (Carlston & Skowronki, 1994, p. 855) This statement implied the trait "cruel". Participants were instructed to familiarize themselves with the materials (exposure task). Then participants were distracted with a confusion task, which interfered with remembering the previous materials, and a filler task (to get participants out of a trait-inferring mode). Second, participants were subjected to a memory test in which they had to memorize photos that were paired with traits (learning task). Some photo-trait pairs matched with the photos and trait-implying behavioral sentences to which participants had to familiarize themselves in the exposure task and some pairs were new. A matched pair could be for example a photo of John that was paired with the trait "cruel", and a new pair could be a photo of John that was paired with the trait "clumsy", or a photo of Dan paired with the trait "cruel". If participants drew STIs during the exposure task, then the learning task would serve as a second learning phase for photos that were paired with matched traits (relearning trials) but not for new pairs (learning trials). Thus, for relearning trials, participants learned twice that John was paired with the trait "cruel". Next, participants were exposed to a filler task. Third and finally, participants were exposed to the photos presented in the learning task, and were asked to recall the word that was paired with the photo (cued recall

task). Results indicated that when the photo-trait pairs matched with the photo-behavioral sentences pairs earlier (relearning), participants had better cued recall than when the pairs were new (learning); this is also called a savings effect (Carlston & Skowronski, 1994).

From these results it was concluded that participants had drawn spontaneous trait inferences. However, Brown and Bassili (2002) suggested that this is not necessarily the case. They investigated spontaneous trait transference effects (STTs). STTs are made when a communicator gives a statement about another actor. In this statement a trait-implying behavioral sentence about the actor is given. The implied trait that is inferred to the actor is also transferred to the communicator (although to a weaker extent). The same happens when there is no communicator, but a bystander. For instance, when a statement is made about John being cruel, and Dan is a bystander, then Dan will also be considered to be somewhat cruel. Research by Brown and Bassili (2002) also showed STT effects when the bystanders existed of inanimate objects. For example, they showed that participants inferred that a banana was superstitious. They concluded that participants did not draw STIs, but made spontaneous trait associations (STAs). Associations are connections that are made between two (or more) stimuli in memory (e.g., a trait gets connected to a photo; Bassili, 1989). Inferences go beyond associations and imply that traits that are drawn based on the trait-implying sentences, are inferred as a property of the actor (Bassili, 1989). We suggest that the same could be happening for artificial agents. Therefore, we investigated whether participants really drew trait inferences for artificial agents (and inanimate objects), or whether they only made associations.

To assess ITIs, we employed a quite straightforward measurement paradigm, namely we provided participants with rating scales (based on e.g. Uleman, 1999). Specifically, we showed the same photo-behavioral sentences pairs, as with a relearning paradigm, but after presenting each pair, we directly asked them to indicate to what extent the implied trait could be attributed to the actor in the photo on a seven-point scale (1 = *not at all* to 7 = *completely*). Therefore, participants indicated that they intentionally attributed the implied traits to the presented actors.

In summary, we investigated participants' spontaneous trait inferences and intentional trait inferences of humans, artificial agents, and inanimate objects. We expected that in their automatic responses, participants might respond in social ways to humans, artificial agents, and inanimate objects. Furthermore, we expected that these automatic social responses would consist of STIs for humans, but would consist of STAs for artificial agents and inanimate objects. As for participants' controlled responses, we expected participants to draw ITIs dependent on actor type.

Overview of the Current Research

In the current five studies, we investigated whether traits that are either spontaneously inferred (STIs) or intentionally inferred (ITIs) are different for actors that are either humans, artificial agents (e.g., robots), or inanimate objects (e.g., bricks). Before investigating participants' STIs and ITIs, we first checked whether participants thought humans, artificial agents and inanimate objects were able to perform behavior or possess traits. In Study 1, we assessed whether participants actually categorized human actors, artificial agents and inanimate objects to different categories. In the four following studies, STIs were measured using a relearning paradigm (Carlston & Skowronski, 1994), and ITIs were measured using a trait-rating task. In Study 2, we investigated the core question of the current research: Do participants in their automatic responses, respond in comparable, social ways to other humans, artificial agents, and inanimate objects, while in their controlled responses their responses depend on actor type. Studies 3 to 5 served first of all to replicate the findings of Study 2, that STIs can be activated not only when an actor is human, but also for artificial agents, and inanimate objects.

Also, in Studies 3 to 5, we investigated an additional question: Are the spontaneously activated trait inferences found in Study 2 actual inferences about traits of the actor? For instance, is the person, the robot, or the curtain hitting someone, actually aggressive (STI)? Or, are the spontaneously activated trait inferences found in Study 2 not real inferences about traits of the actor, but mere associations? For instance, are the person, the robot, or

the curtain that is paired with aggressive behavior (hitting), just associated with each other because they were incidentally paired (spontaneous trait association, STA). In other words, would participants actually automatically draw the conclusion that a specific actor (human, robot, or object) has a certain trait, or is that trait merely associated to the actor in their memory? An answer to this question will help us understand the nature of the automatic social responses towards humans, artificial agents and inanimate objects. More specifically, in Studies 3 to 5, we used various techniques to differentiate spontaneous trait *inferences* (STIs) from spontaneous trait *associations* (STAs; see Crawford, Skowronski, Stiff, & Scherer, 2007).

Study 1

Before investigating whether participants spontaneously responded socially to artificial agents, we first examined how participants categorized artificial agents compared to humans and inanimate objects. More concretely, we investigated whether participants believed that artificial agents (and humans and inanimate objects) were capable of possessing traits (more indicative of inferences), and/or were capable of showing behavior (more indicative of associations). In Study 1, participants were asked to read trait-implying behavioral sentences describing behavior of humans, artificial agents or inanimate objects. For each sentence, participants had to categorize whether this was a trait of the actor in the sentence (which would be more indicative of trait inference), whether it was behavior (which would be more indicative of associations), or whether it was neither a trait nor a behavior (neither indicative of inference nor association; in McCarthy & Skowronski, 2011). We expected that participants would report the highest ratings of traits and behavior for humans (compared to neither-trait-nor-behavior ratings), because humans have a personality and can show human behavior. Next, we expected that participants would report the highest ratings of neither-trait-nor-behavior for inanimate objects (compared to ratings of traits or behavior), because inanimate objects do not have human personality traits and cannot show human behavior. And most importantly, based on Chapter 1 and earlier research

(Blascovich, 2002; Nass, & Moon, 2000; Reeves, & Nass, 1996; Smits, 2006), we expected participants would score artificial agents equally to all three categories (traits, behavior, and neither-trait-nor-behavior), because they are seen as a category that is in between humans and inanimate objects.

Method

Participants and Design

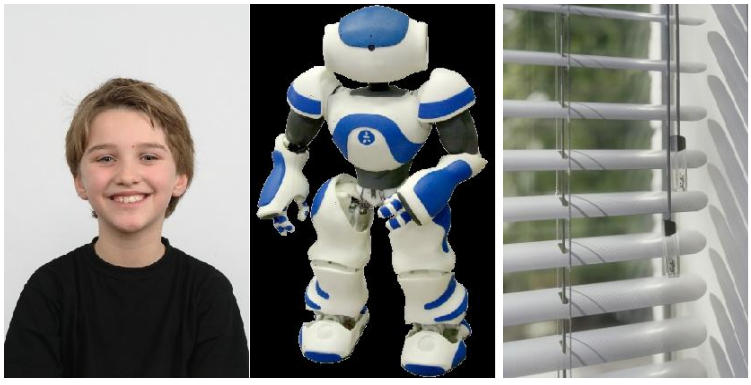
We recruited 45 participants (28 males, 17 females; age *M* = 29.04, *SD* = 16.42), mainly students at Eindhoven University of Technology. Participants were randomly assigned to an actor type (human vs. artificial agent vs. object) between-subjects design. The dependent variables were the percentages of traits rated, behavior rated, and neither rated.

Procedure

Participants were invited to the lab to participate in a study. After they arrived at the laboratory, participants were seated in cubicles behind a desktop computer. Participants were randomly exposed to 12 photo-sentence pairs (Figure 3.1). The sentences consisted of trait-implying behavioral sentences. For participants in the human condition, the actor in the sentences was a human, for participants in the artificial agent condition the actor in the sentences was an artificial agent, and for participants in the object condition the actor in the sentences was an inanimate object. The photos corresponded to the actors in the sentences. In the sentences it was clearly stated whether the actor was human (man, woman, boy, girl), an artificial agent (robot, avatar), or an inanimate object. For example, a sentence about a couch was accompanied with a photo of a couch. For each photo-sentence pair participants had to choose as fast as possible whether this sentence was best categorized as a trait of the actor, as behavior by the actor, or as neither, by clicking on the corresponding button. For example, participants were presented with a photo of a boy/a robot/a couch and the sentence “The boy/the robot/the couch was standing next to the 5 hard-working men and did nothing”. In summary, one-third of the participants only saw

photo-sentence pairs of humans, one-third only saw pairs of artificial agents, and one-third only saw pairs of inanimate objects. Participants then could choose between three buttons; “The boy/the robot/the couch is lazy,” “This is lazy behavior,” or “Neither”. Finally, participants were thanked, paid, and dismissed.

Figure 3.1. Examples of photo-sentence pairs.



The boy / robot / curtain hit the mother in the face

Note: Some participants saw the left photo with the sentence “The boy hit the mother in the face”; some participants saw the middle photo with the sentence “The robot hit the mother in the face”; and some participants saw the right photo with the sentence “The curtain hit the mother in the face”.

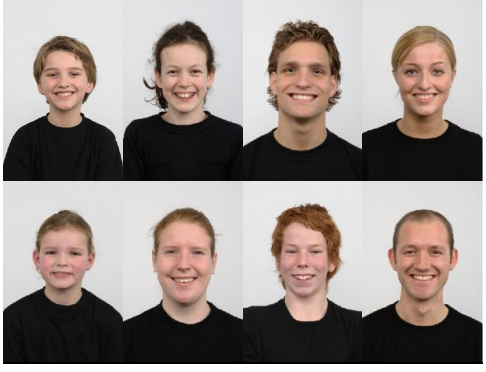
Materials

Trait-implying behavioral sentences. We used twelve short behavioral sentences that implied a trait. One important point to underline, is that the traits used consisted only of *human personality* traits (i.e., not object characteristics).The behavioral descriptions along with their implied traits were mainly based on the short stories that were used in Carlston and Skowronski (1994). However, in this study we only used one sentence instead of a story, and the sentences were adapted in a way to fit all actor types (human, artificial agent and object). In other words, it was made sure that the sentences were both applicable and sounded logical for all actor types. For example, “The boy/the robot/the curtain hit the mother in the face”. The three versions of a sentence were completely identical, except that we replaced the actor in the sentence. By keeping the sentences the same for all actor types we excluded possible confounding of the type of sentences used. A pretest was done to determine which sentences were used as the trait-implying sentences. Forty sentences were

rated by 64 participants (not participating in Study 1) in two tasks. First, participants had to write down the first word that came to mind while reading the sentence. Second, participants had to rate to what extent the implied trait was applicable to the sentence compared with one evaluative consistent trait and one evaluative inconsistent trait (in line with McCarthy & Skowronski, 2011). This was done to investigate whether the implied trait really was specific for the implied behavioral sentence. Twelve sentences were rated as the best trait-implying sentences in both tasks.

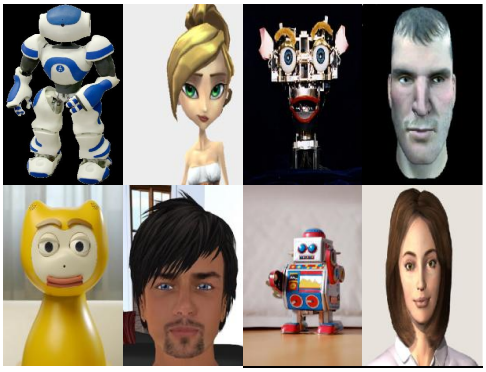
Photos. The human photos (Figure 3.2) consisted of frontal faces used from the Radboud Faces Database (RaFD; Langner, et al., 2010). Both the photos for artificial agents as the photos for inanimate objects were found via a search engine on the Internet. The artificial agent photos (Figure 3.3) consisted of photos of robots, artificial agents, artificial agents, and onscreen characters (one actor per photo). The object photos (Figure 3.4) consisted of photos of inanimate objects like bricks, curtains, glass, and couches (likewise, one actor per photo). We excluded inanimate objects that had visible human-like characteristics (e.g., human-like eyes).

Figure 3.2. Examples of human photos.



Note: These are some of the photos used in our research paradigms.

Figure 3.3. Examples of artificial agent photos.



Note: These are some of the photos used in our research paradigms.

Figure 3.4. Examples of inanimate object photos

Note: These are some of the photos used in our research paradigms.

Results

We hypothesized that participants would report higher ratings of traits and behavior for humans compared to neither-trait-nor-behavior ratings, similar frequency ratings of traits, behavior and neither-trait-nor-behavior ratings for artificial agents, and higher ratings of neither-trait-nor-behavior for inanimate objects compared to ratings of traits or behavior. To investigate how participants categorized humans, artificial agents and inanimate objects, we calculated the percentage of traits rated, the percentage of behavior rated and the percentage of neither rated. These scores were submitted to a 3 (rating: trait vs. behavior vs.

neither) x 3 (actor type: human vs. artificial agent vs. object) within- and between-subjects repeated measures ANOVA. The first factor served as the within-subjects variable. There were no main effects (Rating, $F(2, 41) = .28$, $p = .76$; Actor type, $F(2, 42) = .72$, $p = .49$). Results showed an interaction effect of Rating x Actor Type, $F(4, 84) = 3.53$, $p = .010$, $n_p^2 = .14$. Bonferroni pairwise comparisons were used to investigate participants' ratings for traits, behavior or neither for each actor type. For humans, participants reported higher ratings on traits ($M = 44.45$, 95% CI [30.61 - 58.28]) and behavior ($M = 42.22$, 95% CI [29.22 - 55.22]), than on neither ($M = 13.33$, 95% CI [-.06 - 26.72]), $p = .012$ and $p = .013$ respectively. Ratings for humans on traits and behavior did not differ, $p = .848$. That is, participants chose more traits and behavior than neither when categorizing humans. For artificial agents, there were no differences between traits ($M = 36.11$, 95% CI [22.28 - 49.95]), behavior ($M = 32.22$, 95% CI [19.22 - 45.22]) or neither ($M = 31.67$, 95% CI [18.28 - 45.06]), $p = .738$, $p = .710$, and $p = .960$. That is, participants chose equally for traits, behavior and neither when categorizing artificial agents. For inanimate objects, participants reported higher ratings on neither ($M = 47.22$, 95% CI [33.83 - 60.61]) than traits ($M = 20.56$, 95% CI [6.72 - 34.39]) and marginally higher neither than behavior ($M = 32.22$, 95% CI [19.22 - 45.22]), $p = .030$ and $p = .185$ respectively. Ratings for inanimate objects on traits and behavior did not differ, $p = .317$. That is, participants chose more neither than traits or (marginally) behavior when categorizing inanimate objects.

Discussion

In this study we investigated how participants categorize humans, artificial agents and inanimate objects. In line with expectations, participants reported higher ratings of traits and behavior for humans. That is, participants reported that traits and behavior were more appropriate when the actor in the sentence was human. Also, for inanimate objects, participants reported higher ratings of neither-trait-nor-behavior than of traits and marginally of behavior. More concretely, participants reported that neither-trait-nor-behavior was more

applicable when the actor in the sentence was an object. Finally, participants did not differ on their ratings of traits, behavior or neither-trait-nor-behavior for artificial agents. Participants reported as much traits and behavior as neither-trait-nor-behavior when the actor in the sentence was an artificial agent. These results seemed to suggest that artificial agents were perceived as a category in between humans and inanimate objects. Support for this comes from robotic psychology which showed that participants held different perceptions for robots when they were compared to animals, humans, or inanimate objects (Libin & Libin, 2004). Also, Smits (2006) suggested that new technologies are difficult to categorize, because they fit two categories that are considered to be mutually exclusive. In our case, artificial agents fit into the category of humans, because of their human-like characteristics (like a face), and also into the category of objects, because they are just technology. As mentioned in the Introduction, Joseph Engelberger, the father of robotic, once mentioned that he could not define a robot, but knew one when he saw one (CBC, 2007). However, future research should further investigate whether participants really experience artificial agents as a separate category in-between humans and inanimate objects.

In conclusion it can be said that participants could clearly categorize humans (i.e., possessing traits and are able to show behavior) and inanimate objects (i.e., not possessing traits nor able to show behavior), but had more difficulty in categorizing artificial agents. The implication of these results is that when humans use controlled responses while interacting with artificial agents they still exhibit social responses, although these social responses are weakened (compared to humans). Designers should be aware that when controlled responses are triggered, the effects of their robot on human behavior could be lower than expected. However, another explanation for these results could be that participants' quick, automatic responses were mixed up with their slower, more thoughtful responses. As described in the Introduction, earlier research suggested that on a more automatic level participants respond socially to artificial agents. However, on a more controlled level participants respond to artificial agents as if they were inanimate objects (Reeves & Nass, 1996). In Study 1 participants had to choose as fast as possible, but there were no time

constraints. Perhaps participants chose for traits or behavior when not really thinking about it (more human-like), but chose neither-trait-nor-behavior when they did took some time to think about it (more object-like) when categorizing artificial agents. It could be that these responses got mixed up and as a result participants found it hard to categorize artificial agents. To get a better view of this, in Study 2, participants' automatic, spontaneous and controlled, intentional responses were investigated separately.

Study 2

In Study 1, we observed that it was rather difficult for participants to categorize artificial agents. Furthermore, we wanted to investigate whether the social responses to artificial agents were indeed automatically social and to compare participants' automatic and controlled responses to artificial agents. Therefore, we wanted to investigate participants' spontaneous and intentional responses separately. Earlier research (Reeves & Nass, 1996) suggested that participants responded automatically social to artificial agents as if they were responding to humans. However, Reeves and Nass (1996) did not provide any direct proof that this behavior was indeed automatically social, neither did they directly compare participants' responses to artificial agents with their responses to humans. To our knowledge, this is the first study that directly compared responses to artificial agents with responses to humans and inanimate objects on as well a spontaneous (automatic) as intentional (controlled) level. Study 2 served as an initial test of whether participants spontaneously responded socially to artificial agents (or inanimate objects), compared to human actors. In this study, participants had to perform tasks in a relearning paradigm. In the exposure task, participants had to familiarize themselves with photo-sentence pairs, which consisted of photos from humans, artificial agents and inanimate objects paired with trait-implying[17] behavioral sentences (every pair was presented for 6 s.). In the learning task, participants had to memorize photo-trait pairs (every pair was presented for 6 s.). Some photo-trait pairs corresponded to the photo-sentence pairs in the exposure task (relearning

[17] Important to note is that these were *human personality* traits, and not object characteristics.

trials), and some photo-trait pairs were new (learning trials). Finally, in the cued recall task, they had to recall the words that were previously paired with the photos. Better cued recall on relearning trials than on learning trials suggests that participants drew STIs in the exposure task; this is called a savings effect. After the relearning paradigm, participants performed a trait rating task that measured their ITIs. Based on previous research (Blascovich, 2002; Reeves & Nass, 1996; Epley, et al., 2008; Epley, et al., 2007; Waytz, et al., 2010), we expected that participants would draw STIs about humans and artificial agents and inanimate objects. In contrast, we expected that participants would draw ITIs about humans, to a lesser extent about artificial agents and even lesser about inanimate objects.

Method

Participants and Design

We recruited 80 participants (50 males, 30 females; age M = 24.52, SD = 8.14), mainly students at Eindhoven University of Technology. Participants were randomly assigned to a 2 (learning trial type: relearning vs. learning) x 3 (actor type: human vs. artificial agent vs. object) within-subjects design. The dependent variables were the percentage of correct answers on the cued recall task and the mean ratings scores on the trait rating task.

Procedure

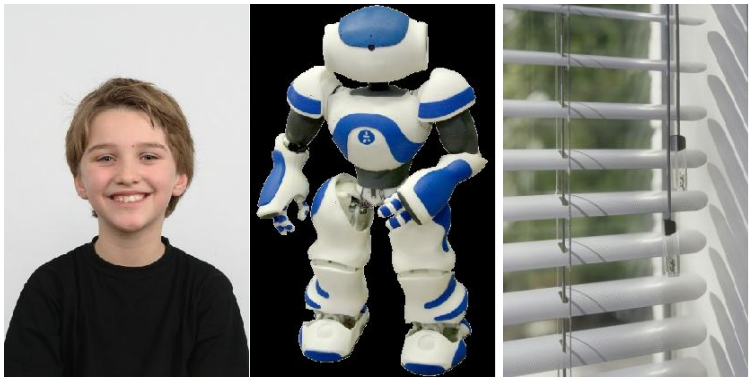
Participants were invited in the lab to participate in a study about memory. When arriving at the laboratory, participants were seated in cubicles behind a desktop computer. Participants were exposed to the relearning paradigm. The relearning paradigm consisted of experimental trials (6 relearning trials and 6 learning trials) and filler trials (18 trials that were neutral or did not reliably elicit a trait). First, participants were exposed to the exposure task which consisted of a computer task that contained 6 relearning photo-sentence pairs[18] (that

[18] Of the 6 relearning trials, participants were presented with 2 human pairs, 2 artificial agent pairs, and 2 object pairs.

were randomly chosen from 12 experimental sentences)[19] and 18 filler photo-sentence pairs (one pair per page). The photos consisted of the different types of actors, namely humans (i.e., males, females), artificial agents (i.e., artificial agents, robots) and inanimate objects (e.g., a brick, a couch). The sentences consisted of behavioral descriptions that implied a trait about the actors in the photo. Participants were instructed to familiarize themselves with the materials and view each photo-sentence pair for 8 s. Next, participants were exposed to a confusion task (see Ham, & Vonk, 2011) to make sure that memory for the photo-sentence pairs was reduced. In this unrelated task, participants had to rate their preference for one of two trait-implying sentences. After the confusion task, participants were exposed to a short unrelated filler task. In this task participants had to rate the valence of bright and dark photos. Next, participants were exposed to the learning task, which consisted of photo-trait pairs. The photo-trait pairs (Figure 3.5) consisted of 12 experimental trials and 18 new filler trials. Six of these experimental trials served as the relearning trials (traits matched with previous behavioral sentences of the exposure task), and the other 6 experimental trials served as the learning trials (new traits). Participants were instructed to look at each photo-trait pair for 6 s. and memorize these pairs for a future recall task (cued recall task). Then, a second unrelated filler task was assessed in which participants had to determine photo or color similarity, which lasted for approximately 5 minutes. In the final task of the relearning paradigm, participants were exposed to the cued recall task in which the photos from the learning task were presented and participants were asked to recall the word that was paired with each photo. After the relearning paradigm, participants were exposed to the trait rating task in which the twelve experimental photo-sentence pairs were again presented. Participants had to rate to what extent a certain trait could be ascribed to the actor type in the photo. Finally, participants were thanked, paid, and dismissed.

[19] In total there were 12 experimental sentences. From these sentences 6 relearning sentences were randomly picked that were shown in the exposure task. The remaining 6 sentences were used as learning photo-trait pairs in the learning task. This was done in order to exclude possible confounding of specific photo-sentence pairs memory.

Figure 3.5. Examples of photo-trait pairs.



Rude Rude Rude

Note: Some participants saw the left photo with the trait “Rude”; some participants saw the middle photo with the trait “Rude”; and some participants saw the right photo with the trait “Rude”.

Materials

Trait-implying behavioral sentences & photos. The same twelve trait-implying behavioral sentences and photos were used as in Study 1.

Trait rating task. To measure intentional trait inferences, we assessed whether the twelve traits were applicable to the actor types (i.e., humans, artificial agents, and inanimate objects). The photo-sentence pairs of the exposure task were again used. For example, the following behavioral sentence was shown: “The boy/robot/curtain hit the mother in the face”. A question was asked for each photo-sentence pair: To what extent does the following trait applies to the human/artificial agent/object in the photo? For example, the following question was asked for a behavioral sentence: “To what extent is the boy/robot/curtain rude?” The items were rated on a Likert scale (1 = *not at all*, 7 = *completely*).

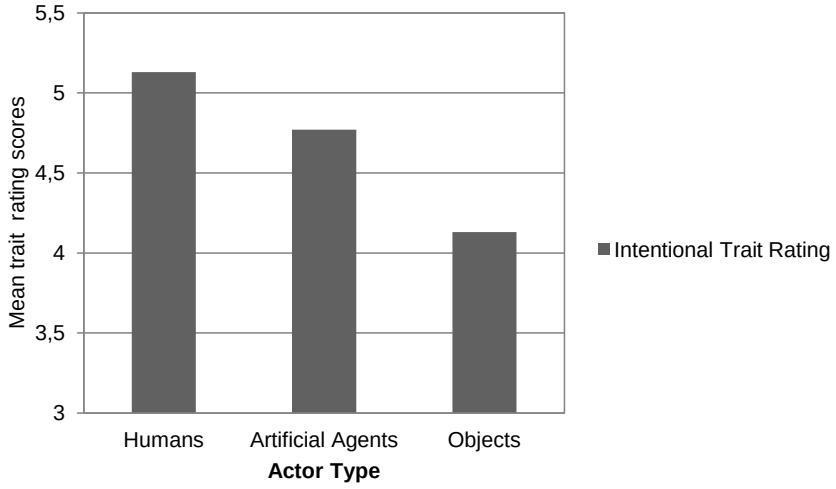
Results

Two participants who indicated suspicion of the nature of the experiment, and one non-native Dutch speaker (who told he did not understand the instructions) were excluded from analyses. The following analyses were conducted on the remaining 77 participants.

ITIs: Trait Rating Scales

First, we wanted to investigate whether participants differed on their intentional trait inferences for humans, artificial agents, and inanimate objects. We expected that participants would draw the strongest ITIs for humans, weaker ITIs for artificial agents, and the weakest ITIs for inanimate objects. We calculated a mean trait rating score with higher scores meaning that participants reported higher scores on the implied trait. In other words, participants drew stronger ITIs (e.g., a boy/robot/curtain is rude). Therefore, the mean trait rating scores were submitted to a single factor (actor type: human vs. artificial agent vs. object) within-subjects repeated measures ANOVA (see Figure 3.6). Results showed an effect of actor type, $F(1.81, 137.25) = 17.85$, $p < .001$, $n_p^2 = .19$. Planned contrasts between the three actor types revealed that participants reported higher intentional trait inferences for humans ($M = 5.13$, $SD = 1.14$) than for artificial agents ($M = 4.77$, $SD = 1.28$), $F(1,76) = 5.81$, $p = .018$, $n_p^2 = .07$, and higher intentional trait inferences for humans than for inanimate objects ($M = 4.13$, $SD = 1.50$), $F(1,76) = 26.33$, $p < .001$, $n_p^2 = .26$. Furthermore, participants made higher intentional trait inferences for artificial agents than for inanimate objects, $F(1,76) = 15.90$, $p < .001$, $n_p^2 = .17$. In sum, participants drew the highest intentional trait inferences for humans, followed by artificial agents, followed by inanimate objects.

Figure 3.6. Mean rating scores in the rating task (ITIs) of Study 2.



Note: A significant effect of actor type was found. The human condition had a higher trait rating score than either the artificial agent condition or object condition. The artificial agent condition had a higher trait rating score than the object condition.

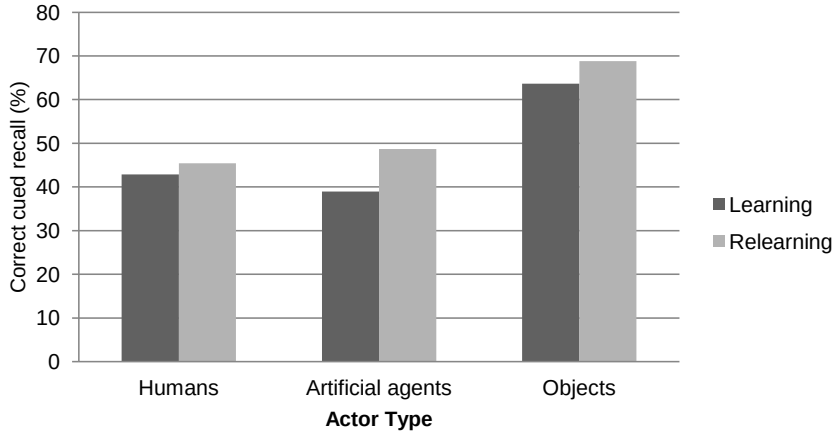
STIs for Humans, Artificial Agents, and Inanimate Objects?

Next, we wanted to investigate whether participants differed on making STIs for humans, artificial agents, and inanimate objects. We expected that participants would draw STIs, independent of actor type. Therefore, the percentage of correctly recalled traits was tabulated for the relearning trials and for the learning trials. The correct cued recall scores were submitted to a 2 (learning trial type: relearning vs. learning) x 3 (actor type: human vs. artificial agent vs. object) within-subjects repeated measures ANOVA (see Figure 3.7). First, we checked whether a savings effect occurred. Results showed a main effect of learning trial type, $F(1, 76) = 4.30$, $p = .042$, $n_p^2 = .05$. Thus the savings effect was significant. That is, the correct cued recall for relearning trials ($M = 54.33\%$, $SD = 38.26$) was higher than for learning trials ($M = 48.49\%$, $SD = 37.10$).

Finally, as expected, the interaction effect of Learning Trial Type x Actor Type was non-significant, $F(2, 152) = .54$, $p = .584$. That is, there was no difference in the magnitude of the savings effect for the different actor types; participants made as many STIs for humans as for artificial agents as for inanimate objects.

Also, results showed a main effect of actor type, $F(2, 152) = 25.44$, $p < .001$, $n_p^2 = .25$. Bonferroni pairwise comparisons of the three actor types indicated that participants had a higher overall cued recall for inanimate objects ($M = 66.23\%$, 95% CI [59.24 - 73.23]) than for humans ($M = 44.16\%$, 95% CI [36.86 - 51.46]) or for artificial agents ($M = 43.83\%$, 95% CI [37.00 - 50.67]), both ps < .001. Participants did not differ on their overall cued recall for humans and artificial agents, $p = 1.00$.

Figure 3.7. Correct cued recall in the relearning paradigm (STIs) of Study 2.



Note: Savings effects were found for all three actor types. No differences between the savings effects were found for the three actor types.

Discussion

In this study we investigated participants' spontaneous and intentional responses to humans, artificial agents and inanimate objects. In line with expectations when participants were directly questioned about their ITIs, they did take actor type into account. That is, the highest ITIs were drawn for humans, followed by artificial agents, followed by inanimate objects. These results are similar to Study 1, that indicated that artificial agents were seen as a category in between humans and objects. In line with previous research (e.g., Carlston & Skowronski, 1994), participants spontaneously drew trait inferences after reading trait-implying sentences. Furthermore, as expected, we found that participants drew STIs for humans, as for artificial agents, as for inanimate objects. It could be argued that the used

sentences are linguistic conventions that make it easier to draw STIs. However, that cannot explain the difference we found between STIs and ITIs.

These results are in line with the CASA paradigm (Reeves & Nass, 1996), that stated that participants automatically respond social to artificial agents, but correct for their social responses in their controlled responses. Also, these results are in line with the theory of anthropomorphism (e.g., Epley, et al., 2007) that stated that participants infer human personality traits to inanimate objects. When designing persuasive agents the highest probability to find effects would be to avoid deliberate thinking about the agents (i.e., triggering controlled responses). We suggest designers to trigger automatic responses by providing persuasive agents with social cues that are sufficient to avoid deliberate thinking.

Somewhat unexpected, we found that participants had an overall better memory for inanimate objects compared to humans and artificial agents. One explanation could be that it was easier to remember things that were dissimilar from each other (like inanimate objects), than to remember things that were similar to each other (like faces of humans, or artificial agents; e.g., Desmarais, Dixon, & Roy, 2007). That is, the inanimate objects in this study were quite different from each other and therefore easier to distinguish from each other and easier to remember. In contrast, the faces of the humans and artificial agents in this study were not that different from each other and therefore more difficult to distinguish from each other and to remember. Another explanation could be that the implied traits in the sentences with the inanimate objects as the actor were seen as more bizarre. In other words, it sounds more logical that a human or a robot is paired with rude than that a curtain is paired with rude.

Concluding, these results suggested that participants had an automatic default to respond socially, but could account for their responses by taking actor type into the equation. However, there are two alternative explanations for the spontaneous social responses to artificial agents and inanimate objects; (1) Because participants also had to draw inferences about humans, it could be that participants were in some kind of inferences-drawing mode, which made them draw inferences of every stimulus that they were presented with (thus

independent of actor type); (2) It could be that the relearning paradigm only measured associations and not inferences per se (see next paragraph). In other words, it could be that participants only made memory associations and did not inferred the traits as a property of the different actors (Bassili, 1989).

Previous research (e.g., Carlston & Skowronski, 2005; Crawford, et al., 2007; McCarthy & Skowronski, 2011; Skowronski, Carlston, Mae, & Crawford, 1998; Todorov & Uleman, 2002; Todorov & Uleman, 2003; Todorov & Uleman, 2004; Uleman & Moskowitz, 1994; see also Uleman, et al., 1996; Wells, Skowronski, Crawford, Scherer, & Carlston, 2011) already strived to distinguish associations and inferences in other paradigms. In these studies, researchers compared STIs with spontaneous trait transferences (STTs). As already explained in the Introduction, STTs are made when a communicator (a person) states a trait-implying behavioral sentence about an actor (another person). The trait that is inferred to the actor is also transferred to the communicator (although in a weaker extent). It is assumed by these authors that STTs are driven by associative processes. Studies that compared STIs with STTs found that there was more evidence for characteristics for attributional processing. That is, a negativity bias (i.e., stronger effects on negative traits than on positive traits) and a halo effect (i.e., general positive evaluation bias) were only found for STIs and not for STTs (Carlston & Skowronski, 2005). Also, a process goal (lie detection) that is known to interfere with attributional processing was used to compare STIs and STTs. Indeed, when participants had to detect whether someone is telling the truth or lying, only the STIs reduced, but the STTs were unaffected (Crawford, et al., 2007).

Another option to disentangle STIs from STAs could be by presenting ambiguous information. For instance when conflicting trait information is presented that leaves it ambiguous whether a person has a certain personality trait, it is more difficult to draw STIs and therefore they would decrease. However, STAs should remain unchanged, because these are only memory connections. For example, when participants first learn that a woman rescued a child from a huge tree, they could infer the woman was brave. However, when they then learn that the child was sitting on the lowest branch of the tree and touching the

grass with her toes, does not make the woman that brave anymore. As a consequence, the previous trait (brave) is invalidated by other information and the STI will get diminished. An STA however is not subjective to ambiguous information and should remain unchanged. Unpublished work by Ham and Skowronski (2007) already showed that STIs were diminished for humans when participants received information that invalidated the previously formed STIs, but left STAs unchanged. That is, when participants first read that "The man lifted a big rock" (which implies strong) and then read that "The rock was made of paper mache", then the inferred trait (strong) was diminished.

When participants were asked directly about their inferences (in Study 2), we saw that they only drew ITIs for humans, and to a lesser extent for artificial agents and inanimate objects. So, participants did experience that humans could possess human personality traits, but artificial agents and inanimate objects not (or at least to a lesser extent). We therefore expected that participants only drew STIs for humans, but made STAs for artificial agents and inanimate objects. In Study 3, we used invalidating sentences to distinguish between STIs and STAs and investigate whether participants indeed made STIs in Study 2 about the different actors, or whether they merely made STAs.

Study 3

In Study 3 we wanted to replicate the findings of Study 2. That is, a savings effect for all three types of actors. Furthermore, we wanted to exclude the alternative explanation that participants were in an inference-making mode, by manipulating type of actor between-subjects. And finally, we wanted to investigate whether participants really drew STIs, or merely made STAs. In this experiment we tried to distinguish STIs from STAs by using an invalidating sentences manipulation. That is, for some sentences in the exposure task (and the trait rating task) we provided participants with invalidating information that should invalidate the original STI. For example, participants first read the trait-implying sentence "The boy hit the mother in the face" (which implies rude), and then receive an invalidating sentence stating "The mother was practicing a scene of her play tonight in which she was in

a fight with someone". In line with previous research (Ham & Skowronski, 2007), we expected that STIs would decrease significantly after receiving the invalidating information, but that STAs would remain unchanged. As already explained above, humans can possess human personality traits, but artificial agents and inanimate objects not (or at least to a lesser extent). Therefore, we expected that trait recall would significantly decrease for humans (STIs), but not for artificial agents and inanimate objects (STAs). Also, we expected that ITIs would significantly decrease for humans, but not for artificial agents and inanimate objects. We first investigated in Study 3a whether the invalidating sentences actually could decrease previously drawn ITIs.

Study 3a

Method

Participants and Design

We recruited 52 participants (22 males, 30 females; age M = 27.06, SD = 11.83), mainly students at Eindhoven University of Technology. Participants were randomly assigned to a 2 (sentences: no invalidating sentences vs. invalidating sentences) x 3 (actor type: human vs. artificial agent vs. object) within- and between-subjects design. The dependent variables were the mean rating scores on the trait rating task.

Procedure

The same trait rating task was used as in Study 2, except for the following adaptations; participants were presented with 24 photo-sentence pairs (instead of 12 pairs); actor type was manipulated as a between-subjects variable; and after half of the trait-implying sentences, an additional invalidating sentence was presented.

Results

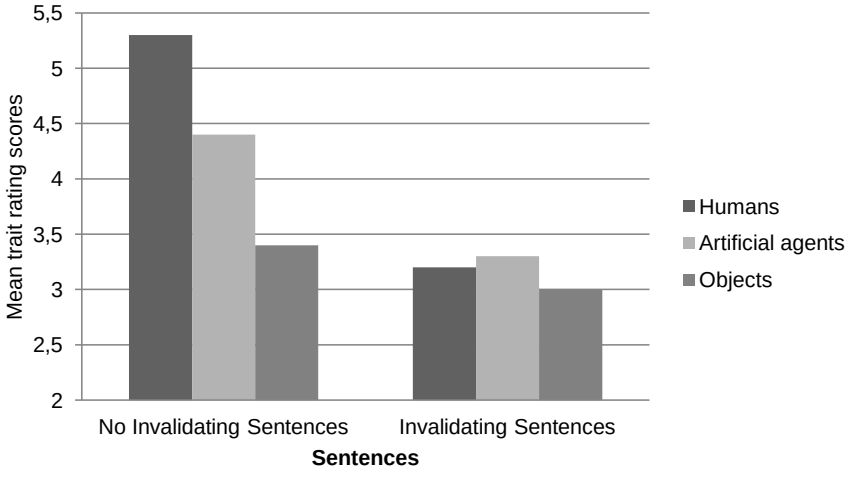
We investigated whether ITIs could be decreased when presenting participants with invalidating information. We expected that when participants were not provided with the invalidating information, they would draw the most ITIs for humans, less for artificial agents,

and even less for inanimate objects (replicating Study 2). In contrast, we expected that when participants were provided with the invalidating sentences, ITIs would decrease, leading to no differences between the types of actors. The mean trait rating scores were submitted to a 2 (sentences: no invaliding sentences vs. invalidating sentences) x 3 (actor type: human vs. artificial agent vs. object) repeated measures ANOVA, with the first factor serving as a within-subjects factor (see Figure 3.8). Results showed that there was a main effect of sentences, $F(1, 49) = 96.08$, $p < .001$, $n_p^2 = .66$. That is, there was a higher rating when there were no invalidating sentences presented ($M = 4.40$, $SD = 1.23$) than when invalidating sentences were presented ($M = 3.17$, $SD = .86$). This implies that the ITIs participants drew significantly decreased when adding invalidating information that invalidated the ITIs.

Second, there was a main effect of actor type, $F(2, 49) = 7.32$, $p = .002$, $n_p^2 = .23$. This effect was validated by an interaction of Sentences x Actor Type, $F(2, 49) = 14.81$, $p < .001$, $n_p^2 = .38$. Bonferroni pairwise comparisons of the three actor types showed that when no invalidating sentences were presented, there were significant differences between all actor types; Ratings of humans ($M = 5.30$, 95% CI [4.83 - 5.76]) were higher than for artificial agents ($M = 4.42$, 95% CI [3.94 - 4.89]), $p = .010$; ratings for humans were higher than for inanimate objects ($M = 3.44$, 95% CI [2.96 - 3.91]), $p < .001$; and ratings for artificial agents were higher than for inanimate objects, $p = .005$. Thus replicating the results of Study 2, we found that participants drew the highest ITIs for humans, lower ITIs for artificial agents, and even lower ITIs for inanimate objects.

When participants were presented with invalidating sentences, the differences between the actor types disappeared. In other words, there was no difference between humans and artificial agents, $p = .846$; no difference between humans and inanimate objects, $p = .437$; and no difference between artificial agents and inanimate objects, $p = .339$. That is, ITIs decreased significantly when presenting the invalidating sentences.

Figure 3.8. Mean rating scores in the rating task (ITIs) of Study 3a.



Note: A significant effect of sentences was found. The no invalidating sentences condition had a higher trait rating score than the invalidating sentences condition. Furthermore, a significant interaction effect of Sentences X Actor type was found. For the no invalidating sentences condition, all actor types differed from each other. For the invalidating sentences condition, none of the actor types differed from each other.

Discussion

In this study we investigated whether invalidating sentences could decrease previously drawn ITIs. Replicating Study 2, when presented with trait-implying sentences participants drew higher ITIs for humans, followed by artificial agents, followed by inanimate objects. However, when presented with trait-implying sentences followed by invalidating information, ITIs decreased for as well humans, artificial agents, as inanimate objects. Consequently, the differences between the actor types disappeared. Concluding, these results suggested that the invalidating sentences were capable of decreasing trait inferences at an intentional level. Therefore, we expected that the invalidating sentences were also able to decrease spontaneous trait inferences at a spontaneous level. In Study 3b we investigated whether STIs could be decreased when presenting invalidating information. We expected that participants made STIs for humans, and that these would decrease when presenting invalidating information. In contrast, we expected that participants made STAs for

artificial agents and inanimate objects, which remained unchanged when presenting invalidating information.

Study 3b

Method

Participants and Design

We recruited 112 participants (69 males, 43 females; age M = 22.28, SD = 4.18), mainly students at Eindhoven University of Technology. Participants were randomly assigned to a 3 (learning trial type: relearning vs. no-relearning vs. learning) x 3 (actor type: human vs. artificial agent vs. object) within- and between-subjects design. The dependent variable was the percentage of correct answers on the cued recall task.

Procedure

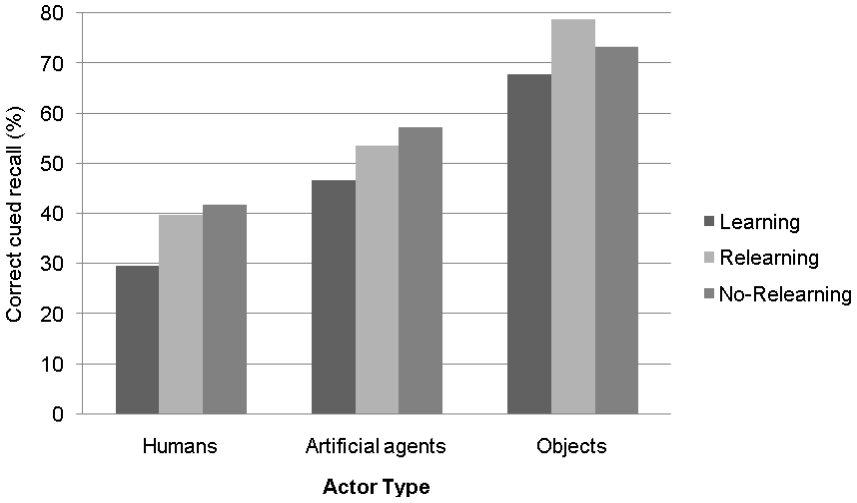
The procedure of Study 3b was similar to that of Study 2 with a few exceptions. First of all, we extended our stimulus material up to 24 experimental sentences. From these sentences, 12 relearning trials were randomly chosen for the exposure task. The remaining twelve sentences were used as learning trials in the learning task. Second, we manipulated actor type as a between-subjects factor. That is, participants were only exposed to photos of humans, artificial agents, *or* inanimate objects. Finally, in the exposure task, we presented additional invalidating sentences after some of the trait-implying sentences. For example, participants were first exposed to the trait-implying sentence “The boy hit the mother in the face” (implies “rude”), and after reading that sentence they were exposed to the invalidating sentence “The mother was practicing a scene of her play tonight in which she was in a fight with someone” (invalidates “rude”).

Results

First, we investigated whether we could replicate the results of Study 2 (i.e., participants drew STIs, independent of type of actor). Second, we investigated whether STIs would decrease when presenting participants with invalidating information. We expected that

participants drew STIs for humans and therefore trait recall should decrease when the invalidating sentence were presented. In contrast, we expected that participants made STAs for artificial agents and inanimate objects and therefore trait recall should remain unchanged when the invalidating sentences were presented. The percentage of correctly recalled traits was tabulated for the relearning trials *without* invalidating information (i.e., relearning), the relearning trials *with* invalidating information (i.e., no-relearning) and for the learning trials. The correct cued recall scores were submitted to a 3 (learning trial type: relearning vs. no-relearning vs. learning) x 3 (actor type: human vs. artificial agent vs. object) repeated measures ANOVA, with the first factor serving as a within-subjects factor. Results showed that there was a main effect of learning trial type, $F(2, 218) = 11.73$, $p < .001$, $n_p^2 = .10$ (see Figure 3.9). As expected, planned contrast analysis[20] showed a higher correct cued recall for relearning trials ($M = 56.85\%$, $SD = 30.14$) than for learning trials ($M = 47.54\%$, $SD = 25.35$), $F(1, 109) = 17.89$, $p < .001$, $n_p^2 = .14$, indicating that the savings effect was significant.

Figure 3.9. Correct cued recall in the relearning paradigm (STIs) of Study 3b.



Note: Savings effects were found for all three actor types. No differences between the savings effects were found for the three actor types. Also, no differences were found between the relearning and no-relearning condition.

[20] Furthermore, there was a higher correct cued recall for no-relearning trials ($M = 56.99\%$, $SD = 28.00$) than for learning trials, $F(1,109) = 19.75$, $p < .001$, $n_p^2 = .15$. That is, participants had a better overall memory for no-relearning trials than learning trials. There was no difference between the correct cued recall for relearning and no-relearning, $F(1, 109) = .00$, $p = .979$.

Thus, we replicated the results of Study 2.

The expected interaction effect of Learning Trial Type x Actor Types was non-significant, $F(4, 218) = .91$, $p = .458$. That is, there was no difference in the magnitude of the savings effect for the different actor types; participants made as many STIs (or STAs) for humans as for artificial agents and inanimate objects, regardless of invalidating information. In other words, the invalidating sentences did not have an effect on the previously drawn STIs (or STAs).

Also, results showed that there was a main effect of actor type, $F(2, 109) = 32.24$, $p < .001$, $n_p^2 = .37$. In line with Study 2, Bonferroni pairwise comparisons of the three actor types indicated that participants had a higher overall cued recall for inanimate objects (M = 73.23%, 95% CI [66.78 - 79.67]) than for humans (M = 37.11%, 95% CI [30.92 - 43.30]) or for artificial agents (M = 52.48%, 95% CI [46.12 - 58.83]), both ps < .001. Not in line with Study 2, results showed that participants had a higher cued recall for artificial agents than for humans, $p = .003$.

Discussion

Replicating the STI results of Study 2, we found that participants who were presented with trait-implying sentences spontaneously drew inferences (or made associations), independent of actor type. In addition, we found that participants had a better overall recall for inanimate objects followed by artificial agents, followed by inanimate objects. In Study 2 we only found a higher overall performance of inanimate objects than humans and artificial agents. However, in Study 3b we also found that participants had a higher overall recall performance of artificial agents than for humans. This could be the result of the extra material we used compared to Study 2 (instead of 6 photo-sentence pairs, we used 12 photo-sentence pairs). That is, the extra material could led robots to stand out more. Next, presenting participants with information that invalidated the trait-implying sentences did not lead to a decrease of STIs for any of the actor types. Although Study 3a showed that ITIs did

decrease, we could not find the same result on STIs. This is in contrast with previous work (Ham & Skowronski, 2007) that found a decrease of STIs for human actors. We expected that participants drew inferences for humans, but made associations for artificial agents or inanimate objects at an automatic level. In other words, we expected that participants did not really infer any traits to artificial agents (or inanimate objects), but only made associations in their memory. However, the invalidating information did not have an effect on the inferred traits (or associations made) for any of the actor types. As a consequence we could not distinguish spontaneous inferences from spontaneous associations in Study 3b.

One possible explanation for our null result is that the invalidating sentences were not strong enough to invalidate the trait-implying sentences. However, we did find an effect at the trait rating task in Study 3a (ITIs diminished). Another option is that our relearning paradigm did not measure STIs at all, but only spontaneous trait associations (STAs). This was already proposed by Brown and Bassili (2002) when using an adapted version of the relearning paradigm. However, more recent work did find differences between STIs and STAs (e.g., Ham & Skowronski, 2007; see also, Carlston & Skowronski, 2005; Crawford, et al., 2007; McCarthy & Skowronski, 2011). Another possible explanation could be that participants were not able to remember the invalidating sentence in the relearning paradigm. Research on negation showed that the rejection of a statement is more effortful (e.g., Gilbert, 1991). In fact, merely comprehending a rejection of a statement can increase the likelihood of considering the idea as true (e.g., Gilbert, Krull & Malone, 1990). This would imply that participants would have a higher cued recall when they were presented with the invalidating sentences, then when they were not presented with the invalidating sentences. However, we did not find that participants had a higher cued recall when they were presented with invalidating information.

Another important finding of earlier research (e.g., Deutsch, Kordts-Freudinger, Gawronski & Strack, 2009; Gilbert, Tafarodi & Malone, 1993) is that memory for negations relies on working memory. In other words, when working memory is taxed, it is very difficult to remember a negation. In our task, participants did not have a lot of time to consider the

invalidating sentences (6 s. per sentence), and they had to familiarize themselves with 30 trait-implying sentences which could have sufficiently taxed working memory. Consequently, this could have led to memory impairment for the invalidating sentences (i.e., negations), and therefore we found no difference when invalidating sentence were presented, and when they were not presented.

Therefore, in Study 4 we extended the presentation times of the sentences (to 10 s. per sentence). In this way, participants were able to rehearse the invalidating sentences for themselves. This could lead to a better memory of the invalidating information.

Study 4

In Study 4, we again wanted to replicate the results of Study 2. Furthermore we wanted to exclude the possibility that participants could not remember the invalidating sentences. Therefore, we replicated the method of Study 3, with the exception that the presentation times of the trait-implying and invalidating sentences were almost twice as long (10 s. instead of 6 s.). In this way, participants had enough time to read the invalidating sentences. We hypothesized that when participants had more time to read and remember the invalidating sentences, then the spontaneous trait inferences would decrease and the spontaneous trait associations would remain the same.

Method

Participants and Design

We recruited 81 participants (20 males, 61 females; age M = 21.41, SD = 4.95), mainly students at Tilburg University. Participants were randomly assigned to a 3 (learning trial type: relearning vs. no-relearning vs. learning) x 3 (actor type: human vs. artificial agent vs. object) within- and between-subjects design. The dependent variables were the percentage of correct answers on the cued recall task and the mean ratings scores on the trait rating task.

Procedure

The procedure of Study 4 was identical to that of Study 3, with the exception that the presentation times of the trait-implying and invalidating sentences were 10 s. instead of 6 s.

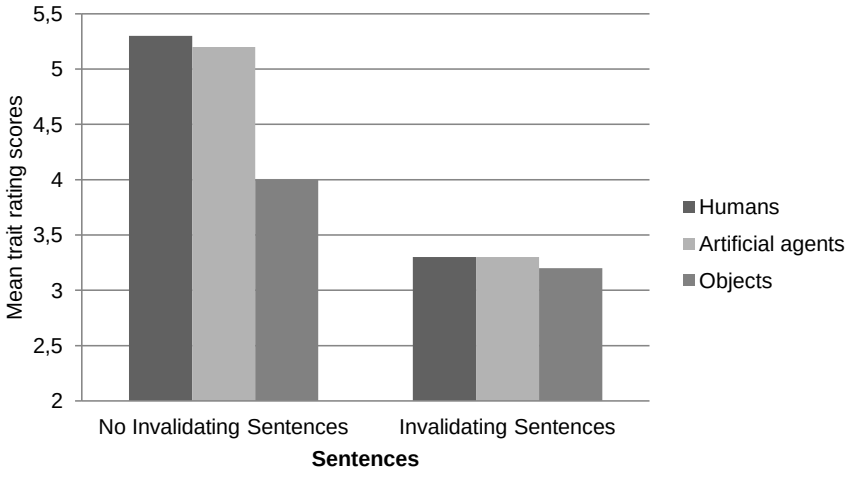
Results

ITIs: Trait Rating Scales

First, we investigated whether ITIs could be decreased when presenting participants with invalidating information, as in Study 3a. We expected to replicate Study 3a's results; strongest ITIs for humans, weaker for artificial agents and weakest for inanimate objects when no invalidating sentences were presented. Furthermore, we expected that when the invalidating sentences were presented, the differences between type of actor would disappear. The mean trait rating scores were submitted to a 2 (sentences: no invaliding sentences vs. invalidating sentences) x 3 (actor type: human vs. artificial agent vs. object) repeated measures ANOVA, with the first factor serving as a within-subjects factor (see Figure 3.10). Results showed that there was a main effect of sentences, $F(1, 78) = 179.35$, $p < .001$, $n_p^2 = .70$. That is, there was a higher rating for no invalidating sentences ($M = 4.86$, $SD = 1.15$) than for invalidating sentences ($M = 3.23$, $SD = .77$). This implies that the ITIs participants drew significantly decreased when adding information that invalidated the ITIs. Second, there was a main effect of actor type, $F(2, 78) = 8.12$, $p = .001$, $n_p^2 = .17$. This effect was validated by an interaction of Sentences x Actor type, $F(2, 78) = 9.49$, $p < .001$, $n_p^2 = .20$. Bonferroni pairwise comparisons of the three actor types showed that when presented with no invalidating sentences, there were significant differences between the actor types; Ratings of humans ($M = 5.31$, 95% CI [4.93 - 5.69]) were higher than for inanimate objects ($M = 4.04$, 95% CI [3.65 - 4.42]), $p < .001$; ratings for artificial agents ($M = 5.23$, 95% CI [4.84 - 5.61]) were higher than for inanimate objects, $p < .001$; however ratings for humans were not higher than for artificial agents, $p = 1.00$. Thus partly replicating the results of Study 3a, we found that participants drew higher ITIs for humans and artificial agents than for

inanimate objects. When participants were presented with invalidating sentences, the differences between the actor types disappeared (M_{humans} = 3.25, 95% CI [2.96 - 3.55] vs. $M_{artificial\ agents}$ = 3.29, 95% CI [3.00 - 3.59] vs. $M_{objects}$ = 3.16, 95% CI [2.86 - 3.45]), all *p*s = 1.00.

Figure 3.10. Mean rating scores in the rating task (ITIs) of Study 4.



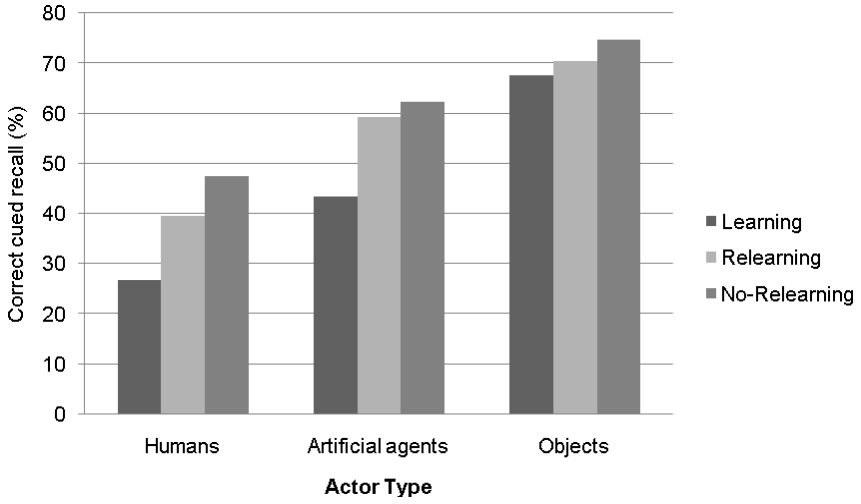
Note: A significant effect of sentences was found. The no invalidating sentences condition had a higher trait rating score than the invalidating sentences condition. Furthermore, a significant interaction effect of Sentences X Actor type was found. For the no invalidating sentences condition, (almost) all actor types differed from each other. For the invalidating sentences condition, none of the actor types differed from each other.

STIs: Relearning Paradigm

Next, we investigated whether we could replicate the results of Study 2 and 3b. We investigated whether these STIs would decrease when presenting participants with invalidating information for a longer period of time. In line with Study 3b, we expected that participants drew STIs for humans (that should decrease when presenting the invalidating sentences) and made STAs for artificial agent and inanimate objects (that should remain unchanged when presenting the invalidating sentences). The percentage of correctly recalled traits was tabulated for the relearning trials *without* invalidating information (i.e., relearning), the relearning trials *with* invalidating information (i.e., no-relearning) and for the learning trials. The correct cued recall scores were submitted to a 3 (learning trial type: relearning vs. no-relearning vs. learning) x 3 (actor type: human vs. artificial agent vs. object)

repeated measures ANOVA, with the first factor serving as a within-subjects factor (see Figure 3.11). Results showed that there was a main effect of learning trial type, $F(2, 156) = 17.19$, $p < .001$, $n_p^2 = .18$. Planned contrast analyses[21] showed a higher correct cued recall for relearning trials ($M = 56.38\%$, $SD = 27.20$) than for learning trials ($M = 45.99\%$, $SD = 23.83$), $F(1, 78) = 14.99$, $p < .001$, $n_p^2 = .16$, indicating that the savings effect was significant. In other words, we replicated the results of Study 2 and 3b.

Figure 3.11. Correct cued recall in the relearning paradigm (STIs) of Study 4.



Note: Savings effects were found for humans, and artificial agents, but not for inanimate objects. No differences between the savings effects were found for the actor types. Also, a marginal difference was found between the relearning and no-relearning condition.

The expected interaction effect of Learning Trial Type x Actor Types was non-significant, $F(4, 156) = 1.63$, $p = .170$. That is, there was no difference in the magnitude of the savings effect for the different actor types; participants made as many STIs (or STAs) for humans as for artificial agents and inanimate objects, regardless of invalidating information. Thus the invalidating sentences did not have an effect on trait recall.

[21] Furthermore, there was a higher correct cued recall for no-relearning trials ($M = 61.52\%$, $SD = 27.72$) than for learning trials, $F(1,78) = 31.31$, $p < .001$, $n_p^2 = .29$. That is, participants had a better overall memory for no-relearning trials than learning trials. Unexpectedly, there was marginally higher correct cued recall for no-relearning trials than for relearning trials, $F(1, 78) = 3.81$, $p = .055$. That is, participants had a marginally better overall memory for no-relearning than for relearning trials.

Also, results showed that there was a main effect of actor type, $F(2, 78) = 23.07$, $p < .001$, $n_p^2 = .37$. Replicating Study 3b, Bonferroni pairwise comparisons of the three actor types indicated that participants had a better overall memory for inanimate objects (M = 70.86%, 95% CI [64.06 - 77.71]) than for humans (M = 37.96%, 95% CI [31.14 - 44.79]), $p < .001$, or for artificial agents (M = 55.04%, 95% CI [48.22 - 61.87]), $p = .005$. Furthermore, participants had a better overall memory for artificial agents than for humans, $p = .002$.

Discussion

Mainly in line with Study 3a, ITIs (when not presented with invalidating information) were highest for humans and artificial agents, and lowest for inanimate objects. Again, ITIs decreased when participants were confronted with invalidating information. Also, replicating the STI results of Study 2, we found that participants who were presented with trait-implying sentences spontaneously drew inferences (or made associations), independent of actor type. In line with Study 3b, we found that participants had a better overall memory for inanimate objects followed by artificial agents, followed by humans. Although results showed that at a controlled level inferences did decrease when presenting the invalidating information, we could not find the same result at an automatic level. This is in contrast with previous research that showed that STIs for humans decreased when presenting invalidating information (Ham, & Skowronski, 2007). Perhaps when spontaneous trait inferences were formed, it was hard to break them down again by invalidating information. A study by Wigboldus, Dijksterhuis and Van Knippenberg (2003) already showed a similar effect with stereotype-inconsistent information. That is, when presenting stereotype-inconsistent behavior *after* the trait-implying sentences, STIs did not decrease. However, when presenting inconsistent stereotype information *before* trait-implying sentences STIs did decrease. For example, when they first primed participants with an actor "*Skinhead/girl*" and then presented participants with the sentence "*X hits the saleswoman*", participants drew weaker STIs when the actor was stereotype-inconsistent (girl) with the implied trait

"aggressive", then when the actor was stereotype-consistent (skinhead). They suggested that when inconsistent information followed trait-implying sentences, the damage was done and it was too late to interfere with the inference process. Furthermore, they suggest that these results served as evidence that STIs occur at encoding. Therefore, in Study 5, we decided to present the invalidating sentences *before* the trait-implying sentences. In this way, participants get some sort of forewarning when considering the trait-implying sentences. We expected that this would lead to a decrease of STIs. However, we expected that STAs would remain unchanged.

Study 5

In Study 5, we again wanted to replicate the results of Study 2. Furthermore we investigated whether it is possible to prevent participants from forming STIs (for humans). Therefore, we replicated the method of Study 4, with the exception that the invalidating sentences were presented *before* the trait-implying sentences.

Method

Participants and Design

We recruited 71 participants (36 males, 35 females; age M = 25.34, SD = 10.95), mainly students at Eindhoven University of Technology. Participants were randomly assigned to a 3 (learning trial type: relearning vs. no-relearning vs. learning) x 3 (actor type: human vs. artificial agent vs. object) within- and between-subjects design. The dependent variables were the percentage of correct answers on the cued recall task and the mean ratings scores on the trait rating task.

Procedure

The procedure of Study 5 was identical to that of Study 4, with the exception that the invalidating sentences were presented *before* the trait-implying sentences (instead of after).

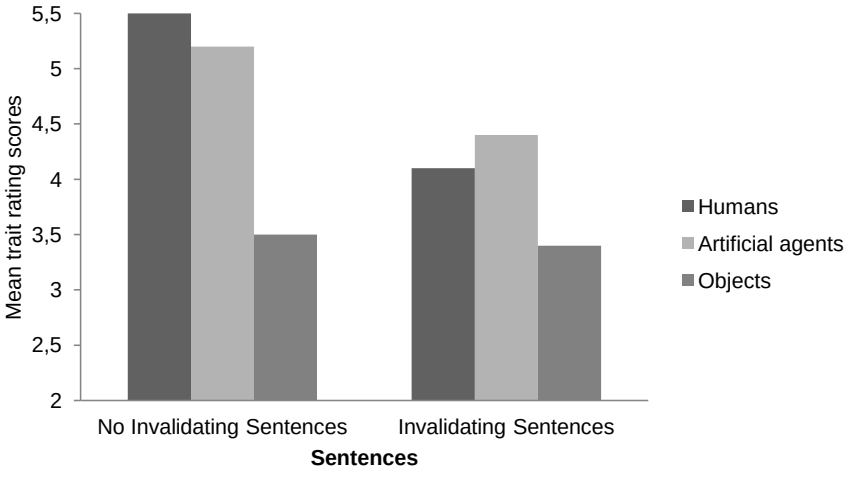
Results

ITIs: Trait Rating Scales

First, we investigated whether ITIs could be decreased when presenting participants with invalidating information. We expected to replicate Study 3a; strongest ITIs for humans, weaker for artificial agents, and weakest for inanimate objects when not presented with invalidating information; however, no differences between the types of actors when presented with invalidating information. The mean trait rating scores were submitted to a 2 (sentences: no invaliding sentences vs. invalidating sentences) x 3 (actor type: human vs. artificial agent vs. object) repeated measures ANOVA, with the first factor serving as a within-subjects factor (see Figure 3.12). Results showed that there was a main effect of sentences, $F(1, 68) = 51.86$, $p < .001$, $n_p^2 = .43$. That is, there was a higher rating for no invalidating sentences ($M = 4.46$, $SD = 1.58$) than for invalidating sentences ($M = 3.84$, $SD = 1.30$). This implies that the ITIs participants drew significantly decreased when adding additional information that invalidated the ITIs. Second, there was a main effect of actor type, $F(2, 68) = 10.45$, $p < .001$, $n_p^2 = .24$. This effect was validated by an interaction of Sentences x Actor type, $F(2, 68) = 14.02$, $p < .001$, $n_p^2 = .29$. Bonferroni pairwise comparisons of the three actor types showed that when presented with no invalidating sentences, there were significant differences between the actor types; Ratings of humans ($M = 5.47$, 95% CI [4.85 - 6.08]) were higher than for inanimate objects ($M = 3.54$, 95% CI [3.10 - 3.99]), $p < .001$; ratings for artificial agents ($M = 5.24$, 95% CI [4.62 - 5.85]) were higher than for inanimate objects, $p < .001$; however ratings for humans were not higher than for artificial agents, $p = 1.00$. Thus partly replicating the results of Study 3a, we found that participants drew higher ITIs for humans and artificial agents than for inanimate objects. When participants were presented with invalidating sentences, the differences between the humans ($M = 4.06$, 95% CI [3.47 - 4.65]) and inanimate objects ($M = 3.44$, 95% CI [3.02 - 3.86]) disappeared, $p = .267$, but the difference between artificial agents ($M = 4.42$, 95% CI [3.83 - 5.00]) and

inanimate objects remained significant, p = .026. Thus we also partly replicated the results of Study 4.

Figure 3.12. Mean rating scores in the rating task (ITIs) of Study 5.

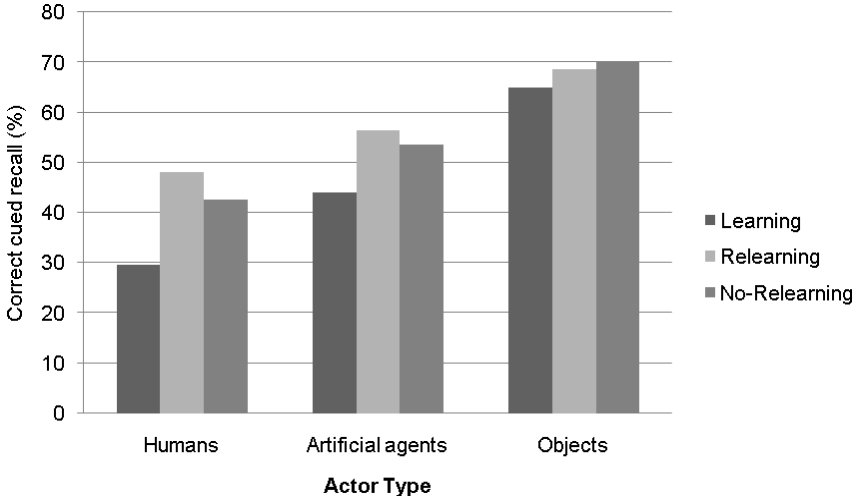


Note: A significant effect of sentences was found. The no invalidating sentences condition had a higher trait rating score than the invalidating sentences condition. Furthermore, a significant interaction effect of sentences X actor type was found. For the no invalidating sentences condition, the humans and artificial agents different from inanimate objects. However, there was no difference between humans and artificial agents. For the invalidating sentences condition, the actor types did not differ from each other, except for artificial agents and inanimate objects.

STIs: Relearning Paradigm

Next, we investigated whether we could replicate the results of Study 2 and whether these STIs would decrease when presenting participants with invalidating information *before* the trait-implying behavioral sentences. We hypothesized that participants drew STIs for humans (that should decrease when presenting invalidating information), but made STAs for artificial agents and inanimate objects (that should remain unchanged when presenting invalidating information). The percentage of correctly recalled traits was tabulated for the relearning trials *without* invalidating information (i.e., relearning), the relearning trials *with* invalidating information (i.e., no-relearning) and for the learning trials. The correct cued recall scores were submitted to a 3 (learning trial type: relearning vs. no-relearning vs. learning) x 3 (actor type: human vs. artificial agent vs. object) repeated measures ANOVA, with the first factor serving as a within-subjects factor (see Figure 3.13).

Figure 3.13. Correct cued recall in the relearning paradigm (STIs) of Study 5.



Note: Savings effects were found for humans, and artificial agents, but not for inanimate objects. No differences between the savings effects were found for the actor types. Also, the expected interaction between Learning trial type X Actor type was not significant.

Results showed that there was a main effect of learning trial type, $F(2, 136) = 8.92$, $p < .001$, $n_p^2 = .12$. Planned contrast analyses[22] showed a higher correct cued recall for relearning trials ($M = 60.33\%$, $SD = 28.08$) than for learning trials ($M = 50.70\%$, $SD = 28.13$), $F(1, 68) = 17.25$, $p < .001$, $n_p^2 = .20$, indicating that the savings effect was significant. That is, we replicated the results of Study 2.

The expected interaction effect of Learning Trial Type x Actor Type was non-significant, $F(4, 136) = 1.33$, $p = .263$. That is, there was no statistical difference in the magnitude of the savings effect for the different actor types; participants made as many STIs for humans as for artificial agents and inanimate objects, even when invalidating information was presented.

Also, results showed that there was a main effect of actor type, $F(2, 68) = 10.04$, $p < .001$, $n_p^2 = .23$. Bonferroni pairwise comparisons of the three actor types indicated that

[22] Furthermore, there was a higher correct cued recall for no-relearning trials ($M = 58.92\%$, $SD = 28.43$) than for learning trials, $F(1, 68) = 11.35$, $p = .001$, $n_p^2 = .14$. That is, participants had a better overall memory for no-relearning trials than learning trials. There was no difference between the correct cued recall for relearning and no-relearning, $F(1, 68) = .54$, $p = .466$. That is, participants did not have a better overall memory for relearning as for no-relearning trials.

participants had a higher overall cued recall for inanimate objects (*M* = 67.86%, 95% CI [60.40 - 75.31]) than for humans (*M* = 40.12%, 95% CI [29.73 - 50.52]), $p < .001$, or for artificial agents (*M* = 51.39%, 95% CI [40.99 - 61.79]), $p = .037$. Humans and artificial agents did not differ from each other, $p = .393$. Thus replicating the results of Study 2.

Discussion

Partly in line with Study 4, ITIs (when not presented with invalidating information) were highest for humans and artificial agents, and lowest for inanimate objects. Again, ITIs decreased when participants were confronted with invalidating information. However, ITIs of artificial agents and inanimate objects remained significantly different from each other when presented with invalidating information. Furthermore, replicating the STI results of the previous three studies, we found that participants who were presented with trait-implying sentences spontaneously drew inferences (or made associations), independent of actor type. Also in line with Study 2, we found that participants had an overall better memory for inanimate objects compared to humans and artificial agents. Next, we expected that participants would draw STIs of humans but made STAs for artificial agents and inanimate objects. Therefore, we checked whether the correct cued recall diminished for humans but remained the same for artificial agents and inanimate objects when invalidating information was presented. Results indicated that cued recall remained the same for humans, artificial agents and inanimate objects. One explanation is that this could be due to our fairly small sample size. Future research should increase the sample size to investigate whether this was indeed the case.

Another explanation could be that we did not measure inferences at all, but associations instead. This was already proposed by Brown and Bassili (2002) in an adapted relearning paradigm. However, other research showed that it was possible to measure STIs and STAs separately with another adaptation of the relearning paradigm (e.g., Carlston, & Skowronski, 2005). Also, Ham and Skowronski (2007) showed that STIs could be decreased while STAs remained the same for human actors. In our studies we did not find any effect of

the invalidating sentences on the cued recall of either humans, artificial agents, or inanimate objects. This could suggest that we did not measure STIs, but STAs instead. In other words, participants did not infer any traits to the humans, artificial agents or objects. They only made associations between the traits and the actors because they were paired with each other. Still, we did find that when participants had to intentionally infer traits to the different actors, they did infer traits to humans, some to artificial agents, and (almost) none to inanimate objects.

In conclusion, results showed that at a controlled level participants did take the actor (human, artificial agent, or inanimate object) into account in their social responses, but at an automatic level they responded socially regardless of the actor. We found that on an intentional level artificial agents were seen as a category in between humans and inanimate objects, which is in line with monster theory (Smits, 2006). In other words, it could be that, just like a monster, participants experienced the artificial agents as fitting into two mutually exclusive categories (i.e., humans and inanimate objects), and therefore found it hard to categorize them.

Furthermore, we can say that participants responded automatically social to artificial agents, which was already proposed by other researchers (e.g., Nass & Moon, 2000; Reeves, & Nass, 1996). We tentatively suggest that associations that were made in the participant's memory led to these social responses to artificial agents. That is, participants did not really infer traits to artificial agents. However, our results did not give any conclusive answer to the underlying process of these automatic social responses. Future research should investigate whether associations led to social responses.

General Discussion

The research in this chapter investigated participant's spontaneously inferred and intentionally inferred trait inferences of humans, artificial agents and inanimate objects. In Study 1, we explored how participants categorized artificial agents. We showed that artificial agents were seen as a category that was in between humans and objects. This is in line with

monster theory that suggested that new technology is difficult to categorize, because it fits two mutually exclusive categories (Smits, 2006). In our case this would mean that artificial agents fit into the category of humans, but also into the category of inanimate objects. As a consequence, it is seen as a category in between humans and inanimate objects. In the other four studies we investigated participants' spontaneous, automatic responses and their intentional, controlled responses, and compared three types of actors (i.e., humans, artificial agents, and inanimate objects). Observing the results at an automatic level, we showed that the spontaneous inferences of participants were social in nature, but participants could control for these responses by taking into account actor type. Our results gave direct proof for previous research (e.g., Reeves & Nass, 1996) that suggested that participants respond automatically social when interacting with artificial agents, but view artificial agents more as inanimate objects when directly asked about this social behavior. Although they still distinguish artificial agents from inanimate objects. We extend this research by showing that at an automatic level, participants respond socially to any type of actor if they were presented with social behavioral cues. Observing the results at a controlled level, we found that participant's rated artificial agents (that were paired with social behavioral cues) as some sort of category that is in between humans and inanimate objects.

Furthermore, in Study 3 to 5, we investigated a possible underlying process for these apparent social responses. We tentatively suggested that humans did not (temporarily) believe that artificial agents possessed personalities, but they formed associations in their memory that led to social responses to artificial agents. Earlier research by Brown and Bassili (2002) suggested that participants did not really respond socially (drawing STIs), but merely formed associations between traits and stimuli in working memory (making STAs). More recent work showed evidence for distinguishing STIs from STAs (e.g., Carlston & Skowronski, 2005; Crawford, et al., 2007; McCarthy & Skowronski, 2011; Skowronski, et al., 1998; Todorov & Uleman, 2002; Todorov & Uleman, 2003; Todorov & Uleman, 2004; Uleman & Moskowitz, 1994; see also Uleman, et al., 1996; Wells, et al., 2011). In this work,

it became clear that STIs were affected by interference of inferential processing, but STAs remained unchanged by these manipulations.

A study by Ham and Skowronski (2007) found that presenting contradictory information about traits, led to an interference of inferential processing. That is, when an invalidating sentence was presented after the trait-implying sentence STIs decreased, but STAs were left unchanged. For example, when they first presented the sentence “The man lifted a rock” (implies strong), but then presented the sentence “The stone was made out of paper mache” (invalidates strong), STIs significantly decreased. Although we found in our studies that ITIs decreased when presenting participants with invalidating information after the trait-implying sentences (see Study 3b and Study 4), we did not find that STIs decreased when presenting participants with invalidating information. A study by Wigboldus and colleagues (2003) found a similar effect. That is, they showed that STIs only decreased when presenting inconsistent stereotypes *before* the trait-implying sentences, but not when presenting these afterwards. Although we found that the means were in the right direction in Study 5, we did not find a similar effect. It could be that our sample size was not large enough to find the effect. Future research should extend the sample size.

Another possibility is that we did not measure STIs, but STAs instead, because STAs would remain unchanged. We did not find an effect of the invalidating sentences, which could imply that we measured STAs. We tentatively suggested that associations in memory lead to social responses to artificial agents (and inanimate objects). However, because we could not distinguish STIs from STAs, we cannot give any conclusive evidence. Most research in person perception is of a cognitive nature. We believe that behavioral measurements are a good way of distinguishing STIs from STAs. In this way, it is possible to observe whether participants really inferred traits to an actor, by observing their future behavior. For example, when participants first learn that an actor is aggressive, would they also keep some distance when encountering this person in the future? If the answer is yes, they probably inferred that the actor is aggressive. If not, they probably made meaningless associations that do not have any consequences for participants’ behavior. More importantly,

would humans also infer traits to artificial agents and keep their distance when encountering them? We believe not. However, future research should investigate whether the underlying mechanism for social responses to artificial agents is an associative process as we tentatively proposed. An implication for designers is that they do not have to design fancy robots to persuade humans. Instead, simple objects that are provided with social cues seem to be enough to trigger social responses. At least, when these social cues trigger automatic responses.

In summary, we found that participants responded automatically social to artificial agents, but controlled for this response by taking into account actor type in their controlled responses. However, it is not clear whether these automatic social responses were based on real trait inferences (STIs) or were merely associations in memory that got formed. We did not find that participants' rating decreased when using successful manipulations (e.g., Ham & Skowronski, 2007; Wigboldus, et al., 2003) that interfered with attributional processing. We tentatively suggested that associations in memory lead to automatic social responses to artificial agents, but we do not have conclusive evidence that supports this suggestion. Future research should investigate the underlying processes of the apparent social responses to artificial agents to get a more clear view about *why* humans seem to respond socially to artificial agents. With our results we can conclude that participants' initial responses to any type of actor were indeed social. On an intentional level, participants experienced the artificial agents as a category in between humans and inanimate objects, which is in line with monster theory about new technology (Smits, 2006). As could be observed in participants' intentional inferences, humans were able to better control for their social responses. It could be that the awareness of the fact that artificial agents are only a piece of technology, led to a better control on social responses. In Chapter 4 we will explore this further.

Chapter 4

Just focus!

Controlling social responses by focusing on the technical characteristics of an artificial social agent

Imagine in the near future that John has a household robot that he really likes. He more or less sees it as his buddy. One day, John cannot find his expensive watch. Then it occurs to him that his robot cleaned the room, and the watch might have been sucked up into the vacuum cleaner. Struck by emotions, he shouts and swears at the robot and threatens to shut him down. In response, the robot starts to display a sad face and body posture as if it was pleading to be innocent. Now, John feels really bad about his behavior. When he tells his girlfriend about what happened, she reminds him that his robot is just a piece of technology, without feelings. Her advice helps John to focus on the technical characteristics of the robot. Thereby, John calmed down and was able to control his initial automatic social responses.

Automatic Social Behavior to Artificial Agents

Earlier research showed that humans behave automatically social to artificial agents (see Chapter 2 and 3; Fogg, & Nass, 1997a; Fogg, & Nass, 1997b; Lee, 2008; Lee, & Nass, 1999; Lee, & Nass, 2004; Moon, 2000; Nass, & Moon, 2000; Nass, et al., 1999; Nass, et al., 1997; Nass, Steuer, Tauber, & Reeder, 1993; for an overview see Reeves, & Nass, 1996). Other research suggested that humans anthropomorphize non-human agents (Epley, et al., 2007). That is, anything that exhibited social cues could trigger attributions of human-like properties (e.g., inferring personality traits as shown in Chapter 3) and human mental states (e.g., emotions). However, when participants in these earlier studies were asked to what extent they thought artificial agents warranted social behavior, they answered that they thought artificial agents did not warrant social behavior at all (Reeves & Nass, 1996). But why do humans respond socially when they know these social responses are not appropriate?

According to Nass and Moon (2000) humans respond socially to artificial agents because they are in a mindless state. When in a mindless state, humans have no conscious attention (i.e., are not aware) for the aspects of the situation they are in (Langer, 1992). That is, humans rely on behavioral scripts that are learned in the past about human-human interaction and apply these scripts to human-computer interaction (Nass & Moon, 2000;

Reeves & Nass, 1996). They (temporarily) ignore the fact that they are interacting with a piece of technology. Reeves and Nass (1996) furthermore suggested that these social responses are an automatic default. Also in daily life humans respond socially to artificial social agents. For instance, when your computer malfunctions after you just wrote an important paper you may curse on it, and sometimes you even have the idea that it intentionally malfunctions; or when you beat the computer player in your favorite video game you feel proud that you finally defeated the computer; or when your little sister's Tamagotchi (i.e., virtual pet) is sick, she feels sad. Thus, humans anthropomorphize technological devices when focusing on the social cues that are displayed (Epley, et al., 2007). Although humans know they are not interacting with a social being, they still respond socially. So are humans doomed to respond socially when exposed to artificial social agents? Or is there a way to control for these social responses? We argue that humans are able to control their initial social responses. In this way, they can overcome the social influence techniques of artificial agents. This is important when designing persuasive technology, because designers do not want to coerce or force users to respond in a desired way.

In line with Nass and Moon's (2000) perspective on the social perception of artificial agents, research on human social perception in human-human interaction (see Bargh, & Pratto, 1986; Bargh, 1994) suggests that social perception is primarily a postconscious, automatic response. That is, to initiate a postconscious automatic response it is only required that certain triggers in the environment (e.g., social cues) are noticed (Bargh, 1994). Humans may be aware of the triggers that initiated the (social) response, but cannot report about the consequences of the response (Park, & Catambrone, 2008). In the current chapter, we argue that a comparable cognitive process could exist when humans interact with artificial agents. That is, when humans observe the social cues displayed by the artificial agent they are interacting with, they automatically respond in social ways, but are not able to report about (and even deny) their social behavior (see Nass & Moon, 2000). We suggest that when humans are reminded about the fact that they are interacting with an artificial agent, and not a human, they snap back into reality and treat the artificial agent as a piece of

technology. Thereby, we argue that it might be possible to control these automatic social responses.

Social Inferences

In order to control these initial automatic social responses, we argue that a cognitive selection process can be initiated that can make a selection by either focusing more on the social characteristics of the artificial agents, or by focusing more on the technical characteristics of the artificial agents. When the social characteristics get selected by focusing on the social characteristics of the artificial agent, participants will respond more socially. However, when the technical characteristics get selected by focusing on the technical characteristics of the artificial agent, participants respond less socially. A similar cognitive selection process was proposed in the social inference literature (e.g., Todd, Molden, Ham, & Vonk, 2011).

In human-human interaction, humans make social inferences when observing others' behaviors. For instance when we see Simon stepping repeatedly on his partner's toes during dancing, we infer that Simon is clumsy. Important to our current line of argument, Gilbert, Pelham and Krull (1988b; Gilbert, Krull, & Pelham, 1988a) explain that the social inference model consists of three phases; categorization, characterization, and correction. First, humans categorize the behavior of the observed person ("Simon is stepping a lot on his partner's toes"). Second, they characterize the behavior by inferring a certain trait that could be implied by the behavior ("Simon is a clumsy person"). And finally, when humans have enough cognitive resources, they can correct for situational information ("Simon's shoe laces are tied together by Eddy; maybe Simon is not a clumsy person after all"; Gilbert et al, 1988a, see also Gilbert, et al, 1988b). The same process might be true for interactions with artificial agents. First, the artificial agent's behavior is categorized ("The robot is opening the door for people"). Second, the artificial agent is characterized ("The robot is polite"), and humans respond socially in response. And finally, when humans have enough cognitive resources, they control for the fact that they are interacting with a piece of technology ("It's just a robot") and refrain from responding socially.

Gilbert and colleagues (1988a; 1988b) argued that humans automatically make dispositional inferences when observing another person, but can control for this when taking into account situational inferences. However, more recent work (e.g., Ham & Vonk, 2003; Krull, 1993; Krull & Erickson, 1995; Krull & Dill, 1996; Lupfer, Clark, Church, DePaola, & McDonald, 1995; Lupfer, Clark & Hutcherson, 1990; Todd, et al., 2011) showed that this model was not completely correct. That is, it was suggested that humans can initially make situational inferences as well as dispositional inferences, and if they have enough cognitive capacity they can control for either dispositional information or situational information. In fact, research by Todd and colleagues (2011) suggested that these situational and dispositional inferences were spontaneously activated in co-occurrence. Furthermore, the goals participants had led to either situational inferences or dispositional inferences. For instance, participants made dispositional inferences when they had the goal to rate the personality of the observed person and made situational inferences when they had the goal to rate the situation of the observed person (see also, Krull & Dill, 1996), *but only* when they had enough cognitive resources (see also, Krull, 1993).

We built our line of argumentation on Todd and colleagues (2011) who argued that human social inferences are based on a selection process that selects an accessible behavior interpretation among multiple (e.g., dispositions, situations, intentions, goals, belief) co-occurring interpretations. This selection process can be guided by goals, lay theories about behavior, or context information that boosts one interpretation above others. Now, we argue likewise, that when interacting with artificial agents, it could be that humans activate social inferences as well as technical inferences of the artificial agents. The social cues (context information) in the interaction could guide humans to be more inclined to respond socially (e.g., Bargh, 1994). However, when they are made aware of the technical characteristics (e.g., with explicit questions; as in e.g., Reeves & Nass, 1996), they could be snapped back to reality and treat artificial agents more as the inanimate objects they really are.

So, we argue that although the initial responses to social cues displayed by artificial agents might be automatic social responses, they are able to control these responses. When humans are made aware of the technical characteristics of these artificial agents, we argue that we can initiate a cognitive selection process (c.f., Todd et al., 2011) that leads to non-social responses. By focusing the attention of humans on the technical characteristics of an artificial social agent, they become able to control their social responses. But are humans able to make a cognitive shift to the technical characteristics (by observing social characteristics and technical characteristics, but focusing on technical), or do humans control for their social responses by completely ignoring the social cues (like humans do with a scary movie they do not want to watch)? If humans can cognitively shift between social and technical cues, designers are advised to design sufficient social cues to keep the user's attention to the social cues. However, if humans ignore the social cues when focusing on technical characteristics, then all the work designers go through to provide persuasive agents with social cues, would be meaningless. In Study 2, we investigated these process explanations.

Overview of the Current Research

In the current two studies, we investigated participants' automatic social responses to artificial agents, and whether participants can control these automatic social responses. In two studies, participants had to watch short film clips of an empathy-provoking (i.e., a social influencing technique) female artificial agent. These film clips contained subtitles that described the topics discussed (either empathy-provoking topics or technical topics). The studies were based on findings by Gilbert and colleagues (1988a; 1988b), Krull (1993) and Todd and colleagues (2011). In Study 1, we investigated whether a focus manipulation could help participants to control their automatic social behavior to an artificial agent. Participants were instructed to focus on either social or technical subtitles while watching film clips of an empathy-provoking artificial agent. Earlier research (Todd et al., 2011, see also Krull, 1993) suggested that participants only responded consistent with their focus (or goals) when they were cognitively busy. In other words, when participants are cognitively busy they do not

have time to consider any alternative interpretations then their goal. As a consequence they respond consistent with their focus. For this reason, we adapted the cognitive load manipulation of Gilbert et al. (1988a). That is, we used the manipulation to give participants a focus by showing them either social or technical subtitles that they had to read and simultaneously we used it to give participants a cognitive load by asking them to memorize the subtitles for a later recall test. In Study 2, we investigated whether this focus was cognitive or attentional in nature. In this experiment, we made sure that participants had enough attention to watch both the subtitles and the film clips. Participants had to focus either on technical subtitles beneath the film clips or after the film clips. Because the film clips and subtitles were shown independent of each other (when the subtitles were presented after the film clips), participants had enough time to watch both the film clips and the subtitles.

Study 1

In Study 1, participants had to watch short film clips of an empathy-provoking female artificial agent. The film clips did not contain sound, but subtitles were presented that showed what was communicated. Some participants saw subtitles that represented social information, whereas some participants saw subtitles that represented technical information. Furthermore, some participants were told to remember the order of the topics discussed in the subtitles, and some participants were not. In that way, some participants had to focus on the subtitles (either social or technical) and some participants did not have to focus. We expected that participants would report higher ratings of empathy (high social behavior) towards the artificial agent when presented with the social subtitles, especially when they had to focus on the social subtitles. However, we expected that participants would report lower ratings of empathy (low social behavior) towards the artificial agent but only when they had to focus on the technical subtitles. That is, when participants focused on the technical features they would be able to control for their initial social responses.

Method

Participants and Design

We recruited 80 participants (52 males, 28 females; age M = 23.70, SD = 6.77) who were mainly students at the Eindhoven University of Technology. Participants were randomly assigned to a 2 (topic: social vs. technical) x 2 (focus: no focus vs. focus) between-subjects design. The dependent variable was the mean score on the empathy for artificial agent scale. Furthermore, we checked whether participants remembered the order of the topics in the short film clips.

Materials[23]

Short film clips. We used fourteen film clips of an empathy-provoking female artificial agent. These film clips were made by programming movements of a 3D graphic artificial agent developed by Haptek Corporation (Version 1; Haptek Inc., 2011) in a way that it resembled an empathy-provoking human. Seven of the film clips contained subtitles about social topics, and seven identical film clips contained subtitles about technical (characteristics of the artificial agent) topics (see Table 4.1). Each film clip lasted approximately between 20 and 30 seconds and showed a female artificial agent that was programmed to move like somebody who tells a social experience, including lip movement (see Figure 4.1). However, like in the original study of Gilbert, and colleagues (1988b), two film clips were designed to show neutral experiences in order to minimize suspicion of the nature of the experiment (see Figure 4.1).

[23] We also used the Inclusion of other in the self (IOS) scale (Aron, Aron, & Smollan, 1992) to explore whether participants felt an overlap with the artificial agent and themselves. Furthermore, we measured empathy emotions to check whether participants' felt any empathy emotions when watching the film clips. Next, we used a pro-social behavior task, because when participants feel empathy, they are more inclined to show pro-social behavior. We asked participants to do any extra tasks to help the experimenter, without gaining anything (Twenge, et al. 2007). However, all these scales did not show any differences. Finally, we measured anthropomorphism and found that participants judged the agent to be marginally more anthropomorphic in the social condition (M = 3.96, SD = 1.15) than in the technical condition (M = 3.49, SD = .99), $F(1, 76) = 3.75$, p = .056, η_p^2 = .05.

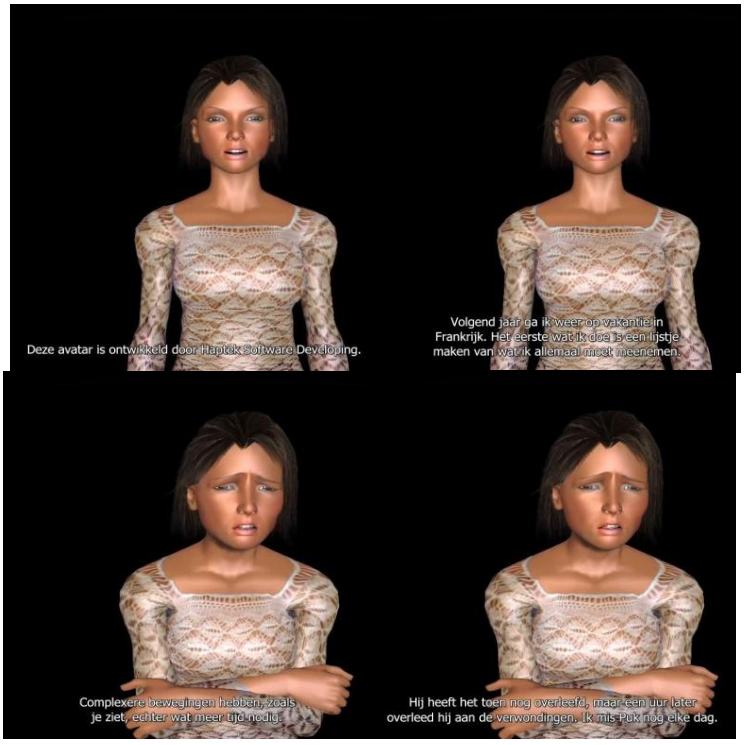
Table 4.1. Topics in the Film Clips and Artificial agent's Behavior.

Technical topics	Social topics	Artificial agent's behavior
Facial expressions	Cheating	Empathy-provoking
Scripts	Car accident	Empathy-provoking
Speed	Dog died	Empathy-provoking
Software developer	Vacation	Neutral
Speech	Fired	Empathy-provoking
Artificial agents	Weather	Neutral
Artificial agent options	Falsely accused of shoplifting	Empathy-provoking

Note. The video material in the film clips of the artificial agent were identical for the technical and social topics. The only thing that differed was the subtitles beneath the film clips that implied technical topics in the technical condition and social topics in the social condition.

Topic order score. To check whether participants focused their attention, we used a recall task for the order of the topics they had to remember. Participants had to watch seven film clips that contained subtitles. In these subtitles conversation topics were made clear. Participants were asked to remember the order of these topics and recall them later. Recall was coded as follows: 2 points for the correct topic on the correct order place; 1 point for the correct meaning on the correct order place, but not the exact topic (e.g., relational problems instead of cheating for the social condition, or something with emotions instead of facial expressions in the technical condition); and 0 points for the incorrect topic. So, participants' scores could range from 0 to 14 points.

*Figure 4.1***.** Snap Shots of the Film Clips.



Note: The left-upper picture is a snap shot of the neutral (technical) film clip with neutral behavior (Software developer).The right-upper picture is a snap shot of the (social) neutral film clip with neutral behavior (Vacation). The left-lower picture is a snap shot of the technical film clip with empathy-provoking behavior (Speed of movements). The right-lower picture is a snap shot of the social film clip with empathy-provoking behavior (Dog died).

Empathy for artificial agent scale. To measure participants' empathy towards the artificial agent, we developed three subscales, in Dutch, consisting of cognitive empathy (e.g., "To what extent do you have the feeling that the character feels uncomfortable?"), emotional convergence (e.g., "To what extent can you empathize with the situation of the character?") and prosocial behavior intention (e.g., "To what extent do you want to help the character?"). These subscales were based on concept descriptions of cognitive empathy (i.e., inferring what another is feeling) and emotional convergence (i.e., experiencing what another is feeling) by Decety and Jackson (2004), and a concept description of prosocial behavior (i.e., benefiting others at the potential costs of the self) by Eisenberg (2008). To prevent participants from thinking thoroughly about their answers, we asked them to answer the questions based on their gut feelings and not think too long about their answers.

Responses were rated at a seven-point Likert scale ranging from 1 (not at all) to 7 (extremely). By averaging these items we constructed a reliable measure for empathy for the artificial agent (α = .826)[24].

Procedure

Participants were invited to the lab to participate in a study about categorization. When arriving at the laboratory, participants were seated in cubicles behind a desktop computer. They were asked to watch seven short film clips as part of an interview in which "experiences" of the artificial agent were told and to answer some questions afterwards. We kept it vague what these questions were about, because we did not want to have participants already focus on either social or technical characteristics of the artificial agent. Instead we wanted participants to focus on the social or technical contents in the subtitles. Therefore, it was made clear that the film clips did not contain sound, but that subtitles were presented that explained the clip's topic. The topics of the film clips became clear in the subtitles (as said, half the participants had to remember the order of these topics). The video material was identical for all participants, except for the subtitles. Participants in the technical topic condition watched film clips in which the subtitles explicitly pointed at the technical characteristics of the artificial agent by discussing technical features (topics were for example "discussing different artificial agent options") and two topics showed neutral behavior and neutral topics (see Table 4.1). The neutral topics were included to minimize suspicion of the nature of the experiment (Gilbert, et al., 1988b). Participants in the social topic condition watched film clips in which the subtitles described the (implied) social characteristics of the artificial agent by discussing social topics (topics were for example "discussing that her dog died") and again two topics concerned neutral behavior and neutral topics (see Table 4.1). Part of the participants were instructed to just watch the film clips while paying attention to the film clips and subtitles (no focus condition). Other participants were instructed to watch the film clips while paying attention to the clips and the subtitles,

[24] Cronbach's alpha of the subscales[24] were α = .702 for cognitive empathy, α = .736 for emotional convergence, and α = .914 for prosocial behavior.

and remember the order of the topics in the film clips (which became clear in the subtitles) for a recall test at the end of the experiment (focus condition). After watching the film clips, participants had to fill in the empathy for artificial agent scale as a measure for social responses. Furthermore they were asked to recall the order of the topics in an open question. These order answers were used to check whether participants really focused on the subtitles. Finally, they were thanked, paid € 5.00 (approximately $ 6.77 at the time of the experiment), and dismissed.

Results

Manipulation Check of Focus

To check whether participants focused more in the focus condition than in the no focus condition, we measured the recall of the order of the topics. We submitted the topic order score to a 2 (topic[25]: social vs. technical) x 2 (focus: no focus vs. focus) between-subjects Univariate ANOVA.[26] Results showed a main effect of focus, $F(1, 76) = 35.99$, $p < .001$, $n_p^2 = .32$. As expected, participants recalled more topics correctly when they were in the focus condition ($M = 8.61$, $SD = 4.18$) than in the no focus condition ($M = 4.03$, $SD = 3.14$), indicating that our manipulation of focus was successful.[27]

Controlling Social Responses?

We expected that participants who watched the social subtitles would show higher ratings of social responses (measured by empathy for the artificial agent) than participants who watched the technical subtitles. In other words, participants could differentiate between social and technical subtitles and this would show in their responses. Furthermore, we expected that participants who watched the film clips would show high ratings of social

[25] To check whether our manipulation of topic was successful we asked an open question. Participants had to indicate their experiences after watching the film clips with subtitles. We observed that participants described the social aspects of the artificial agent more when in the social condition, but described the technical aspects of the artificial agent more when in the technical condition.

[26] Results also showed a significant main effect of topic, $F(1, 76) = 16.56$, $p < .001$, $n_p^2 = .18$. That is, in general, participants recalled more topics correctly when they were in the social condition ($M = 7.98$, $SD = 4.42$) than in the technical condition ($M = 4.78$, $SD = 3.68$).

[27] The interaction effect of the manipulation check was not significant, $F(1, 76) = 2.38$, $p = .127$.

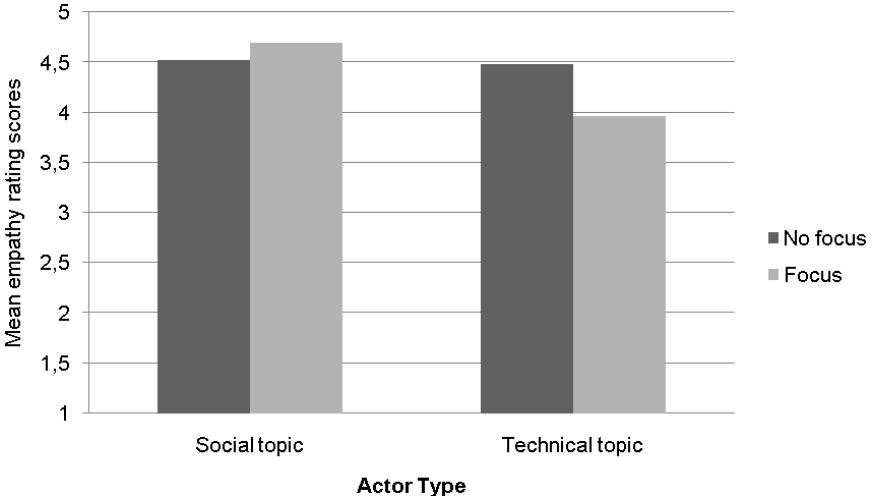
responses, regardless of whether they had to really focus on the social subtitles, or did not have to focus. Also, we expected that participants who watched the film clips and did not have to focus on the technical subtitles, they would still show social responses. However, when participants were asked to watch the film clips with the technical subtitles and really had to focus on these technical subtitles, they would show a decrease in the ratings of social responses. That is, the last group of participants were able to control for their social responses by focusing on the technical characteristics of the artificial agent.

To analyze this we calculated the mean scores of the empathy scale. These scores were submitted to a 3 (subscale: cognitive empathy vs. emotional convergence vs. pro-social intention) x 2 (topic: social vs. technical) x 2 (focus: no focus vs. focus) within- and between-subjects repeated measures ANOVA.[28] The first factor served as the within-subjects variable. Results showed a main effect of topic, $F(1, 76) = 4.57$, $p = .036$, $n_p^2 = .06$. That is, participants gave higher ratings of empathy for the artificial agent when they were in the social topic condition ($M = 4.62$, $SD = .94$) than in the technical topic condition ($M = 4.22$, $SD = 1.20$). In other words, participants gave higher ratings of social responses when watching the film clips with social than with technical subtitles.

Most importantly, the expected interaction between topic and focus was found, $F(1, 76) = 3.68$, $p = .029$ (one-tailed), $n_p^2 = .05$ (see Figure 4.2).

[28] Results also showed a main effect of subscale, $F(2, 152) = 50.39$, $p < .001$, $n_p^2 = .40$, and an interaction between subscale and topic, $F(2, 152) = 10.51$, $p < .001$, $n_p^2 = .12$. There was no main effect of focus, $F < 1$, $p > .05$.

Figure 4.2. Mean empathy rating scores in Study 1.



Note: There was a main effect of topic, but no main effect of focus. Most important, the expected interaction between Topic X Focus was significant. Participants showed social responses in all conditions, but their social responses decreased when they had to focus on the technical subtitles.

Bonferroni pairwise comparisons showed that in the social topic condition participants' empathy for artificial agent ratings did not differ between the focus condition (M = 4.70, 95% CI [4.35 - 5.05]) and the no focus condition (M = 4.52, 95% CI [4.15 - 4.89]), p = .500. That is, empathy for the artificial agent did not increase when participants focused on the (implied) social characteristics of the artificial agent. In the technical topic condition there was an effect of focus. That is, participants expressed more empathy for the artificial agent in the no focus condition (M = 4.48, 95% CI [4.12 - 4.85]) than in the focus condition (M = 3.96, 95% CI [3.60 - 4.32]), p = .045. As expected, participants ascribed more empathy for the artificial agent in the technical condition when they did not have to focus than when they did have to focus on the technical characteristics of the artificial agent. This suggests that participants were able to control for their social responses by focusing on the technical characteristics of the artificial agent. No other effects were significant.

Discussion

In this study we investigated whether it was possible to control the automatic social responses to artificial agents by focusing participants' attention on the technical characteristics of the artificial agent. Results showed that when participants watched an empathy-provoking artificial agent with *social* subtitles, the same high amount of empathy for artificial agent ratings was found, regardless of focus. This non-significant difference might be explained by a ceiling effect. That is, the maximum amount of empathy for an artificial agent was already reached when participants did not have to focus on the (implied) social characteristics of the artificial agent and therefore empathy did not increase significantly when they did have to focus. Also, as expected, participants gave lower ratings of empathy for the artificial agent when they focused on the technical characteristics of the artificial agent than when they did not focus on this information. In other words, participants were able to control for their social responses by focusing on the technical characteristics of the artificial agent. This is in line with previous research (e.g., Krull, 1993; Todd et al., 2011) that showed that when participants had a goal to focus on the dispositional characteristics (here, social characteristics) of a woman (here, artificial agent), they mainly rated the implied dispositional characteristics (here, social characteristics) as the cause of the behavior of the woman (here, artificial agent). However, when they had the goal to focus on the situational characteristics (here, technical characteristics), they mainly rated the situation (here, technical characteristics) as the cause of the behavior of the woman (here, artificial agent).

In summary, our results suggested that participants were able to control their initial social responses to an artificial agent by focusing on the technical characteristics of the artificial agent. However, this study could not exclude the possibility that participants completely ignored the social cues in the film clips, and only concentrated on the technical subtitles. Thus, it could be that participants controlled their social responses by simply not looking at (the social cues of) the artificial agent. This would suggest that participants made use of an *attentional* focus. If this is the case, then the direct sensorical experience of an artificial agent leads them to respond socially and people can only overcome their social

responses by not looking at the artificial agent. Another explanation could be that participants did observe the social cues as well as the technical cues of the artificial agent, but then *cognitively* focused on the technical characteristics. This cognitive focus led them to control their social responses. If this is the case, participants are not doomed to respond socially when observing an artificial agent, but are able to control their social responses by focusing more on its technical characteristics.

To get a more definite answer on whether participants made use of an attentional or cognitive focus, we manipulated the presentation time of the subtitles in Study 2. That is, we either presented the subtitles beneath the film clips, as in Study 1, or we presented them after the film clips. By presenting the subtitles after the film clips, we made sure that participants did not completely ignore the film clips (as would be the case with an attentional focus); because they first had to watch the film clip and afterwards had to watch the subtitles of the film clip. Thus, in Study 2 we investigated whether participants made use of a cognitive focus (i.e., observing social and technical characteristics, but concentrating on the technical characteristics), or of an attentional focus (i.e., ignoring the social characteristics completely, and only paying attention to the technical characteristics).

Study 2

In Study 2, participants again had to watch short film clips of an empathy-provoking female artificial agent (see Figure 4.1). We wanted to replicate the findings of Study 1 that participants reported empathy for this agent (showing social responses) when not focusing on the technical characteristics, and that they reported lower empathy for this agent when they did focus on the technical characteristics (controlling their social responses). Furthermore, we wanted to exclude the possibility that this effect was due to an attentional focus (instead of a cognitive focus). In this study we only used technical subtitles. Some participants saw the subtitles beneath each film clip, whereas other participants saw the subtitles after each film clip. We presented subtitles afterwards in order to prevent participants from ignoring the social cues of the artificial agent and only paying attention to

the technical subtitles. In line with the previous study, we expected that participants would report lower ratings of empathy when they had to focus (than when they were not instructed to focus) on technical subtitles beneath the film clips. Furthermore, we expected that participants used a cognitive focus. This hypothesis would be supported when participants paid attention to both the social cues and the technical cues of the artificial agent, but cognitively controlled their social responses by focusing on the technical cues. However, if participants attentionally controlled their social responses, they ignored the social cues and only focused on the technical characteristics of the artificial agent. We manipulated the presentation of the subtitles to investigate whether participants made an attentional or a cognitive focus. When subtitles were presented beneath the film clips, as in Study 1, participants could focus their attention completely on the subtitles, and ignore the social cues (attentional focus). However, when the subtitles were presented after the film clips they could only focus on the subtitles *after* watching the film clips. In this way, participants *had* to observe the social cues of the artificial agent[29].

More concretely, if empathy ratings remained low (in both focus conditions) regardless when the subtitles were presented (i.e., beneath or after the film clips), this would indicate that participants cognitively controlled their social responses. That is, it did not matter whether they paid attention to both the social cues and the technical cues of the artificial agent, they still gave lower empathy ratings when focusing on the technical characteristics. However, if the empathy ratings were higher when the subtitles were presented after the film clips than when the subtitles were presented beneath the film clip, this would indicate that participants were able to control their social responses by using an attentional focus. That, is participants ignored the social cues and focused their attention completely on the technical characteristics of the artificial agent when the subtitles were presented beneath the film clips. However, when the subtitles were presented after the film clips, participants did not have the chance to completely ignore the social cues of the

[29] Of course it was still possible to ignore the artificial agent when the subtitles were presented afterwards. We tried to prevent this by telling participants they should pay attention to both the film clips and the subtitles, and we asked participants afterwards whether they paid attention to the film clips and the subtitles.

artificial agent and so this should influence their degree of social responses. We expected that participants used a cognitive focus to control for their automatic social responses.

Method

Participants and Design

We recruited 71 participants (36 males, 35 females; age *M* = 25.34, *SD* = 10.95), mainly students at Eindhoven University of Technology. Participants were randomly assigned to a 2 (focus: no focus vs. focus) x 2 (subtitles: beneath vs. after) between-subjects design. The dependent variables were again the mean scores on the empathy for artificial agent scale. Furthermore, we checked whether participants remembered the order of the topics in the short film clips. Also, we asked them to what extent they paid attention to the film clips and the subtitles to check whether they did not completely ignore the film clips.

Materials

We used the same materials as in Study 1, except for the subtitles of the short film clips. That is, we only used film clips with technical subtitles and the subtitles were either presented beneath each film clip, or after each film clip. We again used the empathy for artificial agent scale. By averaging these items we constructed a reliable measure for the empathy for artificial agent scale was (α = .86).

We also added two questions to check to what extent participants paid attention. We asked one question to check their attention to the film clips (attention to film clips score) and one question to check their attention to the subtitles (attention to subtitles score), both on a seven-point Likert scale with (1) *not at all* to (7) *extremely*.

Procedure

The procedure was identical to that of Study 1, with the exception that the film clips all contained technical subtitles that were either presented beneath each film clip or after each film clip. The subtitles were presented after the film clips to make sure that participants did not completely ignore the film clips (as would be with an attentional focus) but paid attention to both the film clips as the subtitles (as would be with a cognitive focus). Also, we

asked participants afterwards to what extent they paid attention to the film clips and subtitles (attention scores).

Results

Attention to Film Clips and Subtitles

To check whether participants paid attention to both the film clips and the subtitles,[30] and did not just ignore the film clips, we submitted the attention scores to a One-Sample T-test that compared the attention scores with the midpoint of the attention scale. Results showed that participants paid more than average attention compared to the midpoint of the attention scale (*M*s > 3.5) to both the film clips ($M = 5.42$, $SD = .79$), $t(71) = 20.59$, $p < .001$, and the subtitles ($M = 5.79$, $SD = .79$), $t(71) = 24.37$, $p < .001$. As expected, participants did not ignore the film clips as they would when having an attentional focus. This result gave a first indication that participants made use of a cognitive focus.

Manipulation Check of Focus

To check whether participants focused more in the focus condition than in the no focus condition, we calculated a score for the recall of the order of the topics (as in Study 1). We submitted this score to a 2 (focus: no focus vs. focus) x 2 (subtitles: beneath vs. after) between-subjects Univariate ANOVA. Results showed a main effect of focus, $F(1, 67) = 5.14$, $p = .027$, $n_p^2 = .07$. As expected, participants recalled more topics correctly when they were in the focus condition ($M = 4.97$, $SD = 4.03$) than in the no focus condition ($M = 3.20$, $SD = 2.45$). So, our manipulation of focus seemed to be successful.

Attentional Focus or Cognitive Focus?

In line with Study 1, we expected that participants who were instructed to focus on the technical subtitles would show a decrease in the ratings of social behavior compared to participants who were not instructed to focus. Furthermore, we wanted to investigate

[30] We also, submitted the attention scores to a 2 (question: film clips vs. subtitles text) x 2 (focus: no focus vs. focus) x 2 (subtitles: beneath vs. after) within- and between-subjects Repeated measures ANOVA. Results showed that there was a main effect of question, $F(1, 67) = 9.57$, $p = .003$, $n_p^2 = .13$. That is, participants paid more attention to the subtitles text ($M = 5.79$, $SD = .79$) than to the film clips ($M = 5.42$, $SD = .79$) in general.

whether participants completely ignored the film clips (attentional focus), or observed both the film clips and subtitles and in response controlled for their social responses (cognitive focus). We expected that participants made use of a cognitive focus, rather than an attentional focus when controlling their social responses. Therefore, we expected that the effects were not influenced by whether the subtitles were presented beneath or after the film clips. That is, we expected that there was no interaction between focus and subtitles, because this would indicate that the presentation of the subtitles did not matter. Thus participants did not ignore the film clips when the subtitles were presented beneath the film clips, but cognitively focused on the technical characteristics of the artificial agent, and consequently controlled their social responses.

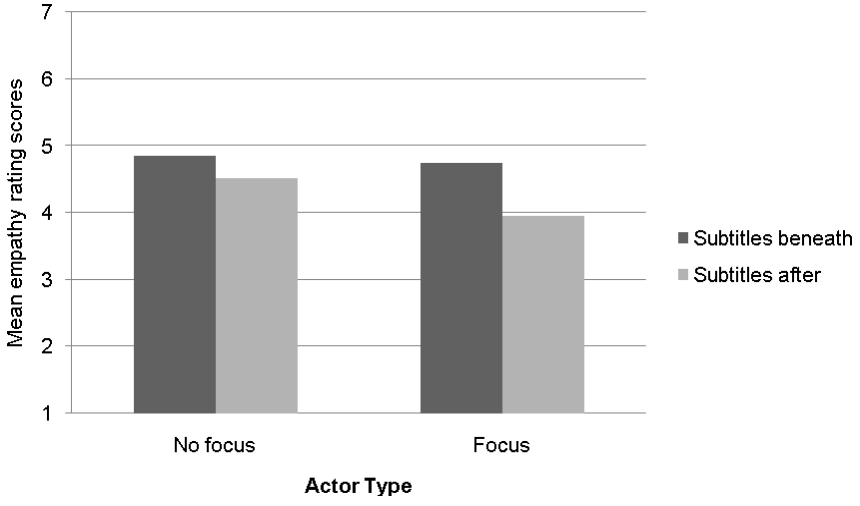
To analyze whether participants experienced empathy for the artificial agent in response to the short film clips, we calculated the mean scores for the empathy for artificial agent scale. These scores were submitted to a 3 (subscale: cognitive empathy vs. emotional convergence vs. pro-social intention) x 2 (focus: no focus vs. focus) x 2 (subtitles: beneath vs. after) within- and between-subjects repeated measures ANOVA.[31] The first factor served as the within-subjects variable. Replicating Study 1, there was a (marginal) main effect of focus, $F(1, 67) = 2.70$, $p = .053$ (one-tailed), $n_p^2 = .04$. That is, participants gave higher ratings of the empathy for artificial agent when they were in the no focus condition ($M = 4.68$, $SD = 1.18$) than in the focus condition ($M = 4.35$, $SD = 1.29$). So, when confronted with the technical characteristics of the artificial agent, participants decreased their social responses when they strongly focused on the technical characteristics.

In addition, the main effect was not qualified by a significant interaction between subtitles and focus, $F(1, 67) = 1.27$, $p = .265$. That is, when confronted with technical subtitles, participants reported less empathy for the artificial agent when they focused on the

[31] Results also showed a main effect of subscale, $F(1.75, 117.47) = 58.80$, $p < .001$, $n_p^2 = .47$, and a marginal significant interaction between subscale and topic, $F(1.75, 117.47) = 2.50$, $p = .093$, $n_p^2 = .04$. Also, there was a main effect of subtitles, $F(1, 67) = 7.98$, $p = .006$, $n_p^2 = .11$. That is, participants gave higher ratings of empathy for artificial agent when they were in the subtitles beneath condition ($M = 4.79$, $SD = 1.01$) than in the subtitles after condition ($M = 4.24$, $SD = 1.34$).

technical characteristics of the artificial agent, regardless of whether the subtitles were presented beneath or after the film clips. In other words, it did not matter whether the subtitles were presented beneath or after the film clips, focusing on the technical characteristics had led to less empathy for the artificial agent compared to not focusing. This result suggested that participants made use of a cognitive focus (see Figure 4.3).

Figure 4.3. Mean empathy rating scores in Study 2.



Note: There was a main effect of focus, and a main effect of subtitles. Most important, these effects were not qualified by a significant interaction between focus and subtitles. In other words, participants could cognitively focus on the technical cues and consequently diminish social responses.

Discussion

In this study we investigated whether participants could control their automatic social responses to an artificial agent by either an attentional focus or a cognitive focus. Results showed that participants paid more than average attention (compared to the midpoint of the attention scale) to both the subtitles as to the film clips. This implied that participants did not completely ignore the film clips, as would be explained by an attentional focus. Instead, they observed both social and technical cues and then cognitively focused on technical cues. Thereby, diminishing their social responses. In line with Study 1, we found that when participants were presented with an empathy-provoking artificial agent and they were instructed to focus on the technical characteristics of the artificial agent, they reported less

empathy for the artificial agent than when they were not instructed to focus on the technical characteristics.[32] There was no difference between the focus conditions when the subtitles were presented beneath the film clips or when presented after the film clips. In line with our previous findings, we could say that it did not matter whether the subtitles were presented beneath or after the film clips, focus led to less empathy than no focus. This again implied that participants' used a cognitive focus to control their social responses, and not an attentional one.

In summary, participants were able to control for their social responses by focusing on the technical characteristics of the artificial agents (taking into account both social and technical cues). Furthermore, these results indicated that participants made use of a cognitive focus when controlling for their social responses. So, these findings suggested that participants do not have to ignore the artificial agent to prevent them from responding socially. Instead, they can observe the social artificial agent and then focus on its technical characteristics to control for their social responses. An implication for designers of persuasive agents that want users to respond socially is that they should try to avoid users to focus on the technical cues of the persuasive agent.

General Discussion

Previous research suggested that human responses to artificial agents were socially in nature. At the same time it was shown that when explicitly asked about their social responses, humans denied responding socially and indicated social responses to artificial agents were not warranted (e.g., Nass & Moon, 2000; Reeves & Nass, 1996). It was however not clear whether humans were doomed to respond socially after experiencing social cues, or whether it was possible to control for these social responses. The current research used a paradigm from the social inferences literature that was proven to be successful in tapping participants' automatic responses and also investigated whether they could control for these automatic responses (e.g., Gilbert et al., 1988a; Krull, 1993).

[32] Note that this was a marginal effect ($p = .051$, one-sided).

In two studies we demonstrated that participants can control for their automatic social responses to artificial agents. When confronted with an artificial agent, participants can control their social responses by focusing on the technical characteristics of the artificial agent. Furthermore, Study 2 showed that this was due to a cognitive focus and not an attentional one. That is, participants observed both the social characteristics as the technical characteristics of the artificial agent, but cognitively focused on the technical characteristics which led them to control their social responses (i.e., lower empathy for artificial agent ratings). Thus, humans do not have to look away, like with a scary movie they do not want to watch, to avoid the "social illusion". Like with a movie, humans can control their responses by strongly focusing on the fact that it is not real (here, the technical characteristics). The results of the current work were in line with previous studies that investigated the activation of dispositional and situational inferences (e.g., Gilbert et al, 1988a, Gilbert, et al, 1988b; Ham & Vonk, 2003; Krull, 1993; Krull & Erickson, 1995; Krull & Dill, 1996; Lupfer, et al., 1995; Lupfer, et al., 1990; Todd, et al., 2011). These studies showed that when humans had a goal to make situational inferences, they followed this goal when they did not have enough cognitive capacity. Hence, when focusing on (i.e., making humans aware about) situational characteristics (technical characteristics in the current research) while being busy led humans to control their initial automatic dispositional responses (social responses in the current research).

Previous research (e.g., Gilbert et al., 1988a; Krull & Erickson, 1995) spoke of revision or correction processes, we avoided the use of these words in our terminology. We suggest using "controlling" initial automatic responses, because we do not have empirical evidence of the underlying process. On the one hand, it could be that participants in our studies initially responded socially but corrected for their social responses by taking into account the technical characteristics. On the other hand, it could be that participants did not initially responded socially, but prevented the social responses from being activated in the first place. Although both processes sound plausible, in Study 2 we found that participants paid attention to both the social characteristics as the technical characteristics of the artificial

agent, which could imply that there was an initial activation of social responses. Also, other studies showed that multiple behavior interpretations were initially activated when observing a person (Ham & Vonk, 2003; Krull, 1993; Krull & Dill, 1996; Todd, et al., 2011). It could be that humans did activate the social responses, but selected only the technical information when rating the artificial agent. Recent research speculated that some sort of selection process guides which interpretation humans select among many interpretations when rating the observed person (Todd, et al., 2011). Further research should investigate these aspects of the underlying process and give a more definite answer to this question.

Although our results are in line with previous research, there are some drawbacks. First, we used a one-way form of communication. That is, participants had to watch film clips of an artificial agent. There was no real interaction with the artificial agent. In the studies of Reeves & Nass (1996) a real interaction with a computer (or artificial agent) was used. It could be that it is even more difficult to control for automatic social responses when interacting with an artificial agent. Some recent work already showed that it was really difficult to control social responses when interacting with an artificial agent (Ham, & Midden, 2010; Midden, & Ham, 2009; Midden, & Ham, 2012). We expect that participants are able to do so, but have to be reminded about the fact that they are interacting with a machine. For instance, a study could be conducted in which participants interact with a robot and are given the goal to figure out the precise machinery within the robot. In this way, participants have to focus on the technical characteristics of the robot that could help them overcome their social responses. However, it should be noted that humans often have certain expectations when interacting with an artificial agent that seems to be working independently. When these expectations are not met, they may become aware of the technical characteristics of the artificial agent and therefore do not respond socially anymore (e.g., Mori, 1970).

Second, we did not compare the responses of participants to an artificial agent with their responses to a human directly. Therefore, we cannot exclude the possibility that humans respond differently to artificial agents than to humans. However, our results were in

line with previous research that investigated participants' responses to humans (e.g., Ham & Vonk, 2003; Krull, 1993; Krull & Dill, 1996; Todd, et al., 2011). Also, our manipulation is not feasible with human actors. That is, we manipulated whether participants had to focus on the social characteristics or the technical characteristics of an artificial agent. Because humans do not have these technical characteristics it is hard to compare this in a direct experiment. Also, in Chapter 3 we showed in a direct comparison that participants responded the same to humans as to artificial agents (and inanimate objects).

Coming back to the example in the Introduction, John can indeed control for his social responses. When he focuses strongly on the technical characteristics of his robot he will be able to control for his social feelings towards the robot while interacting with it. Important to mention is that previous research (Todd, et al., 2011) suggested that this will only work if he is cognitively busy.

Chapter 5

General Discussion

Making artificial agents effective persuaders, requires a better understanding of the nature of human social responses to artificial agents. Earlier research (e.g., Blascovich, 2002; Epley, et al., 2007; Reeves & Nass, 1996) suggested that humans exhibit automatic social responses to artificial agents. Interestingly, it was also shown that when asked explicitly about their relation to artificial agents, humans typically respond that they do not warrant social treatment (e.g., Reeves & Nass, 1996). Apparently, with respect to social relations to artificial agents, automatic responses and controlled responses can be incongruent.

Therefore, we asked ourselves in Chapter 1: What is it that makes humans exhibit a social response to artificial agents? Hitherto, there has been little empirical research that investigated this question directly. The aim of the current research was to investigate the nature of social responses to artificial social agents in greater depth and to determine to what extent these social responses are indeed automatic. Furthermore, we aimed to use the acquired knowledge to formulate recommendations for the design of effective persuasive agents. Based on earlier research and theorizing that explored human social responses to artificial agents (as described in Chapter 1), we predicted that humans exhibit automatic social responses to artificial agents. Therefore, in Chapter 2 we started our research by demonstrating that participants in an experimental study exhibit a social response to a robotic agent that used direct language (i.e., demands rather than requests). Based on this finding we investigated the automatic nature of the social response (Chapter 3). This research was the first to directly investigate the automaticity of social responses to artificial agents. In Chapter 3, in line with our predictions, our results showed that social responses to artificial agents were mainly automatic. Furthermore, we found evidence that, following Bargh's framework of the four components of automaticity (Bargh, 1994), human's automatic social responses to artificial agents can be characterized as unintentional and as requiring low cognitive effort. Chapter 4 confirmed and extended the automatic nature of these social responses by showing that they can only be controlled with considerable cognitive effort. In the following sections, we will briefly review our main research findings. Subsequently, we

will discuss our findings in light of related work. Finally, we will discuss limitations of our work and discuss implications for persuasive technology.

Overview of Research Findings

In Chapter 2 we formulated three main objectives. The first objective was to demonstrate that humans exhibit social responses when interacting with an artificial agent. The second objective was to investigate whether persuasion by an artificial agent could backfire (e.g., lead to less persuasion) in a similar way as is the case with human persuaders. Finally, the third objective was to explore factors that could enhance social responses to artificial agents. We combined these objectives by creating a robotic agent that tried to persuade humans to save energy. In two studies we tried to induce psychological reactance as a social response to a robotic agent that used direct language (prompting reactance) versus one that used more indirect language (barely prompting reactance). Confirming our hypotheses, the participants in our experiment became psychologically reactant (i.e., behaved socially) when they felt threatened in their autonomy (i.e., freedom of choice). Thus, in line with Reeves and Nass (1996), we demonstrated that humans exhibit social responses to an artificial agent.

We also explored two factors that could enhance the observed social responses (i.e., social agency and goal sharing). Both factors indirectly implied that the artificial agent possessed intentionality, which is a uniquely human characteristic (Epley et al., 2007). In other words, in our studies the factors social agency and goal sharing would imply that the artificial agent's responses were intentional. However, manipulating the social agency of the agent or manipulating whether or not the agent shared the participant's goals did not affect the social responses to the robotic agent. That is, it did not matter whether the participants believed that the agent possessed human features (like having social agency or having goals; Epley et al., 2007). Instead, in line with Reeves and Nass (1996), it seemed that simple social cues (e.g., a face, a voice, or human-like behavior) were sufficient for triggering social responses, suggesting that the underlying processes are automatic (see also Vossen

et al., 2010). The nature of the automatic social responses was further investigated in Chapter 3.

The research reported in Chapter 3 had three main objectives. The first objective was to find evidence for the hypothesis that human social responses to artificial agents are automatic. Earlier research (e.g., Nass & Moon, 2000) suggested that human automatic and controlled responses to artificial agents are different. Following speculations by Reeves and Nass (1996), the second objective was to find evidence for the hypothesis that human automatic responses to artificial agents are similar to human automatic responses to humans. In addition, we investigated the hypothesis that controlled responses to artificial agents were more similar to their responses to animate objects (see Chapter 3). The final objective was to study a possible underlying process of automatic social responses to artificial agents. In five studies we measured spontaneous trait inferences (automatic responses) and controlled trait inferences (controlled responses) as the social responses to humans, artificial agents, and inanimate objects. As shown by studies in social cognition, humans tend to draw spontaneous or controlled trait inferences when they read behavioral sentences that imply human personality traits (for an overview, see Uleman et al., 2008). More specifically, Uleman and colleagues (2008) suggested that spontaneous trait inferences are automatic on all four components of automaticity (see also Uleman, et al., 1996). That is, humans draw spontaneous trait inferences unintentionally, they are not aware of drawing spontaneous trait inferences, humans can (almost) not control themselves drawing spontaneous trait inferences, and spontaneous trait inferences are drawn efficiently (e.g., do not require much cognitive capacity, also see Bargh, 1994). As predicted, our results showed that human social responses to artificial agents occurred automatically. Furthermore, as predicted, their automatic responses to artificial agents were similar to their responses to other humans. Extending our hypothesis, we found that also the automatic responses of participants to inanimate objects were similar to their responses to humans. That is, at an automatic level, we found that human responses were social to anything that exhibited social cues (i.e., other humans, artificial social agents, and inanimate objects).

Interestingly, our results suggested that humans also responded socially to artificial agents at a more controlled level. However, human controlled social responses were weaker than their automatic social responses. A possible explanation is that when participants had sufficient time to elaborate on their responses to artificial agents, they realized that artificial agents were only objects and, hence, showed a weaker response. As we predicted, participants' controlled responses to artificial agents were more similar to their responses to inanimate objects (Study 2-5, Chapter 3). Still, we argue that because the artificial agents exhibited social cues, participant's social responses did not completely disappear (as they did for inanimate objects, Study 2-5, Chapter 3). Further examination suggested that participants viewed artificial agents as a category in between humans and objects (see also Study 1, Chapter 3). Summarizing, at a controlled level, current results suggested that participants viewed artificial agents as a category in between humans and inanimate objects.

To pursue our third objective in Chapter 3, we investigated the process that underlies beneath human automatic social responses to artificial agents. Based on earlier literature (see Chapter 3), we argued that humans do not necessarily infer that artificial agents possess human features (e.g., a personality), but merely make memory associations (associative processing), which could also explain automatic social responses to agents. For instance, when one reads that a robot hits a mother in the face, this does not immediately leads to the conclusion that the robot has a rude personality, but the robot may be associated with the word "rude" in memory. In other words, humans might not make spontaneous trait *inferences*, but spontaneous trait *associations*. In Study 4 and 5 (Chapter 3), we used a method that interferes with inferences processes, but leaves associative processes unchanged. We hypothesized that participants would draw spontaneous trait inferences for humans, but would make associations for artificial agents and inanimate objects. Unfortunately, we could not find a conclusive answer, because we could not sufficiently differentiate between inferences and associations. In Chapter 3 we discussed several possible explanations for our results. The most plausible explanation for our results is that participants made associations (and not inferences) in memory for all categories,

humans, artificial agents and inanimate objects. Brown and Bassili (2002) already suggested that participants made associations and not inferences when they used a similar paradigm to investigate whether humans made associations or drew inferences about humans and inanimate objects. Still, more recent research found that participants drew inferences about human actors (as described in Chapter 3). To our knowledge, the current studies were the first to investigate whether participants made associations or drew inferences about artificial agents. Future work should give more clear-cut answers on the differentiation between inferential and associative processes as basis for social responses to artificial agents.

In Chapter 4, we had one main objective, which was to investigate whether it was possible for humans to control for their automatic social responses to artificial agents. Controllability is one of the components of automaticity (Bargh, 1994). Research in Chapter 4 suggested that participants could hardly control their social responses to a virtual female agent when they observed its social behavior cues. However, supporting our hypothesis, participants were able to control their social responses by strongly focusing on the technical features of the virtual agent. In other words, humans were not doomed to respond socially (as could be expected for completely automatic behavior) to an artificial social agent. Further examination (Study 2) revealed that it was not necessary for participants to completely ignore the social cues of the artificial agent at the perceptual level (i.e., not looking at it directly). Instead, they were able to *cognitively* control for their social responses. That is, they could observe both the social and technical cues of the agent, but cognitively focus on the technical cues and thereby significantly reduce their social responses. In line with Todd and colleagues (2011), we suggested that humans use a selection mechanism when observing artificial agents that selects whether they respond socially or not. This mechanism can be guided by social cues of the agent that promote social responses, or can be guided by technical cues of the agent and thereby inhibit social responses. It could be the case that humans have a tendency to pay attention mainly to the social cues of the artificial agent and therefore respond socially in response (Chapter 2-4). However, our studies show that in

principle they do not have to respond socially if they are able to cognitively focus on the technical cues (Chapter 4).

So, based on the results of our research we gained a better understanding of how humans respond to artificial agents. In the following paragraphs we will discuss our results in light of theories about social responses to artificial agents.

Discussion

In Chapter 1 we discussed several important theories that suggest why humans respond socially to artificial agents. In line with Epley and colleagues (2007), all studies in the current dissertation suggested that humans responded socially to artificial agents. Epley and colleagues (2007) explained that humans respond socially to artificial agents because they anthropomorphize these agents. That is, they proposed that humans infer uniquely human characteristics (e.g., personality, intentionality) about these agents (see for a more detailed discussion, Chapter 1). Our results partly support the theory of anthropomorphism proposed by Epley and colleagues (2007). At an automatic level, we showed that participants responded socially to artificial agents. That is, our research suggested that participants inferred these agents to have human characteristics (like a personality) based on their behavior. However, because we could not exclude the possibility that participants did not infer personality traits, but made associations instead (see Chapter 3), we cannot give a definite answer to the question whether participants anthropomorphized non-human agents at an automatic level. In Chapter 2 we observed that manipulating human characteristics (like social agency or goal sharing) hardly affected participant's social responses. This could imply that the participants did not believe that the robot in this study possessed human characteristics. So, at an automatic level, we could not find conclusive evidence that participants inferred that the agents possessed human characteristics. On the other hand, at a controlled level, we do have evidence that indicates that they anthropomorphized artificial agents. That is, when explicitly asked whether an artificial agent possessed certain personality traits, participants reported positively and thus responded socially (Chapter 3).

Future research should examine to what extent humans really infer that an artificial agent possesses human characteristics, or whether they are only triggered to respond socially. An important step could be to link participants' cognitions about artificial agents with participants' actual behavior when interacting with these agents. For example, it could be investigated whether humans would take a detour when they encounter a *static* robot, but to which they do attribute aggressiveness based on earlier experiences (see also "Limitations and Future Research" below). Important to note is that researchers should be aware that participants respond differently to measures that tap different types of responses. Earlier research in social cognition, showed that human cognitions as measured by direct and indirect measures do not always match (e.g., Gawronski & Payne, 2010). When participants' responses are measured by a direct measure, then participants have the opportunity to consciously think about their responses. Research has shown that when participants do not have this opportunity, they respond quite differently (Gawrosnki & Payne, 2010). We argue that the same happens when participants' social responses to artificial agents (or even inanimate objects) are measured. That is, when participants can think consciously about how to respond to artificial agents, they know that these agents do not warrant social treatment and therefore do not report social responses at a direct measure. However, if you ask participants more indirectly about their social responses to artificial agents, they do report social responses. This mismatch has been studied by Ruijten, Ham and Midden (2012) who investigated anthropomorphism with an indirect measure by using an implicit association test (IAT). Their preliminary results suggested that participants do anthropomorphize when interacting with artificial agents or when measured by an indirect measure (in which they do not think consciously about their responses), but their answers showed little evidence of anthropomorphizing when measured by a direct measure (Ruijten, et al., 2012). In other words, when participants do not have the opportunity to consciously think about their social responses, then their automatic and controlled responses to artificial agents are rather similar. Other work has shown similar results (see also Eyssel, et al., 2012).

Our results were also in line with Reeves and Nass' (1996) media equation hypothesis which explored participants' social responses to artificial agents. They suggested that humans respond automatically socially when interacting with artificial agents. Our empirical research is the first that actually provided direct support for this suggestion by using measures that were exclusively developed to assess automatic responses and thereby allow to disentangle controlled and automatic responses. We showed that participant's social responses to artificial agents are indeed mainly automatic. Furthermore, our research extended the media equation hypothesis on several aspects.

First, Reeves and Nass (1996) only observed responses to artificial agents (not responses to humans) in their research. They claimed that participant's social responses to these agents were similar to their responses to humans. Our research supported their claim but also suggested that humans do not only respond socially towards artificial agents, but to objects as well. That is, they showed automatic social responses to humans, artificial agents (ranging from a squared box to human-like agents) and inanimate objects. One could suggest that a human-like agent would trigger more automatic social responses than a squared-box embodied agent. Although we did not compare responses towards different kinds of artificial agents (i.e., from squared box ranging to human-like) with each other, we did find that humans responded socially even to inanimate objects. Therefore, we assumed that our results can be generalized to anything that exhibits social cues. In other words, human automatic social responses to artificial agents, as well as to inanimate objects, seem to be similar to their automatic social responses to other humans.

Furthermore, Nass and Moon (2000) suggested that humans are aware of artificial agents not being social and that humans will not respond socially at a controlled level. They supported their claim by asking participants directly how they would behave to artificial agents. The participants reported that they would not behave socially, even though results showed that they had shown social responses to artificial agents previously. In contrast with their suggestions, our research showed that participants also responded socially to artificial agents at a controlled level, that is, when explicitly (but indirectly) asked. Furthermore our

analyses showed that participants viewed artificial agents as a category in between humans and inanimate objects (Chapter 3). This outcome is in line with recent theorizing about new technologies (Smits, 2006). That is, Smits (2006) explained that just like monsters, new technologies simultaneously fit two categories that are mutually exclusive, and, as a consequence participants find it hard to explicitly categorize new technologies. Applied to our findings, artificial social agents simultaneously fit the category of human and the category of inanimate objects. We argue that this resulted in a new category in between humans and objects and therefore participants' social responses did not completely disappear at a controlled level (similar to monster assimilation; see Smits, 2006). In summary, we conclude that participants' social responses to artificial agents are mainly automatic, but also at a controlled level participants exhibit social responses to artificial agents.

Finally, Reeves and Nass (1996) suggested that humans cannot control their social responses when interacting with artificial agents. Our research confirmed that it was indeed hard for humans to control their social responses to artificial agents (see also Midden & Ham, 2012). However, in Chapter 4, we demonstrated that it *was* possible for humans to control their social responses. That is, they were not doomed to respond socially when observing an artificial social agent. Further examination showed that to refrain from responding socially, it was not necessary for participants to completely ignore the artificial agent at the perceptual level. Instead, they could cognitively control for their social responses. That is, they could observe both the social and technical cues of the agent, and by cognitively focusing on the technical cues, participants could significantly reduce their social responses. In line with Todd and colleagues (2011), we suggested that humans used a selection mechanism when interacting with artificial agents that determines whether they respond socially or not. We argue that this mechanism might be guided by social cues of the agent that lead to social responses, or guided by technical cues of the agent that not lead to social responses. It could be the case that humans have a tendency to pay attention mainly to the social cues of the artificial agent and therefore respond socially in response (as seen

in Chapter 2-4). However, our studies show that in principle they do not have to respond socially if they are able to cognitively focus on the technical cues (Chapter 4). This result seems in line with the notion of the uncanny valley (Mori, 1970): humans stop responding socially when they notice that the artificial agent is not a human being.

A final theory that we want to discuss in light of our results is the theoretical model of social influence (Blascovich, 2002). Blascovich (2002) proposed that cues of human agency and/or behavioral realism are necessary to provoke social responses to artificial agents (see Chapter 1). As seen in Chapter 2, making participants believe that the artificial agents have some kind of agency barely affected their social responses to artificial agents. That is, manipulating the agent's social agency did not influence how participants responded to the agent. Because Blascovich (2002) used another concept of agency (i.e., the extent to which humans believe that a real human is behind the behavior of the artificial agent), we cannot exclude the possibility that a manipulation of the concept of agency could enhance social responses. Still, earlier research did not find any effects of this concept of agency on the social responses to artificial agents (see Chapter 1).

All in all, our research showed that the presence of for example a face (Chapter 2, Study 1 and 2), voice (Chapter 2, Study 2), and/or behavioral cues (Chapter 2-4) led humans to automatically respond socially towards artificial agents. That is, we suggest that simple social cues are sufficient to trigger social responses to artificial agents. But what can we say about the automaticity of these social responses?

In Chapters 2 to 4 we presented evidence suggesting that humans responded automatically socially to artificial agents. As already discussed in Chapter 1, automaticity consists of different components (e.g., Bargh, 1994). How can our findings be understood in light of these different components of automaticity? In other words, we pose the question of whether humans are aware of their social responses to artificial agents, and whether these social responses are efficient, intentional, and/or controllable? We found some indirect support for all components and direct evidence for the controllability of social responses to artificial agents. In Chapter 3, we used a paradigm from social cognition research to

investigate participants' automatic responses. As described previously, Uleman and colleagues (2008) suggested that spontaneous trait inferences are automatic on all four components of automaticity; awareness, efficiency, intentionality and controllability. It should be noted that Bargh (1994) distinguished three forms of awareness: (1) individuals can be aware of the stimulus material; (2) individuals can be aware of the way a stimulus is interpreted; and (3) individuals can be aware of the influence of the stimulus on their behavior. In our studies participants had to be aware of the stimulus material for it to be effective. Due to our instructions (Study 2-5, Chapter 3) they were not aware of how the stimulus material could be interpreted, nor were they aware of the effect of the stimulus material on their behavior. That is, we did not instruct participants to draw (spontaneous) trait inferences. We merely asked them to study the stimulus materials (behavioral sentences paired with either humans, artificial agents, or inanimate objects). And, unbeknownst to the participants, they spontaneously and efficiently drew trait inferences. Furthermore, participants had little control over drawing spontaneous trait inferences. Previous research also showed that humans could not remember drawing trait inferences (Uleman, et al., 1996). Thus, participants were not aware of how to interpret the social behavior cues (i.e., the stimulus materials), nor of responding socially in response (i.e., the effect of the stimulus materials). In summary, this analysis suggests that we used a research paradigm in which social responses to artificial agents are automatic on all four components of automaticity.

In the context of persuasion it is important that humans are able to control their responses. For that reason we particularly focused on controllability of social responses to artificial agents. That is, we used a research paradigm to find out whether it is possible to control for social responses to artificial agents. Our research in Chapter 4 showed that participants can control for their social responses to artificial agents by strongly, cognitively focusing on the technical characteristics of the artificial agents. So, although it can be difficult, they seem to be able to control for their social responses to artificial agents.

In summary, we found indirect support that human social responses to artificial agents were automatic in the sense that social responses were spontaneous (unintentional)

and efficient, and that they were unaware of their social responses. Furthermore, we found direct support that human social responses to artificial agents were to some extent controllable.

Limitations and Directions for Future Research

In the current research we compared participants' controlled and automatic responses to artificial agents. We found that participants could be brought into the illusion that artificial agents are social beings. In response, they respond socially to the artificial agents. However, humans could also get pulled out of this illusion (e.g., if a robot is not properly working), or pull themselves out of the illusion by focusing on the technical features of the artificial agents. In the current research we measured the different components of automaticity by using a paradigm that was established to be automatic on all four components of automaticity. Furthermore, we focused on the automaticity of social responses and the controllability of these responses. Future research could focus more on direct evidence of the other components of automaticity (i.e., awareness, intentionality and efficiency). For example, there are tasks in implicit social cognition developed to test the different components of automaticity (for an overview, Gawronski & Payne, 2010). Other research methodologies for assessing automaticity and its components have been developed that could investigate this in greater depth. Also, the process dissociation procedure of Jacoby (1991), or the quadruple process model of Conrey, Sherman, Gawronski, Hugenberg and Groom (2005) may be used to determine what specific parts of the process are automatic and what parts are controlled. Future research could investigate the other components of automaticity directly and thereby further disentangle the automaticity of human social responses to artificial agents.

In most of our studies we investigated how participants respond when observing artificial social agents. In our research, participants did not take part in an interaction with artificial social agents (except for Study 2 in Chapter 2). We view our research as a first step to a better understanding of how humans respond to artificial social agents. Future research could investigate whether our results are applicable to realistic human-agent interactions.

For example, experiments could be designed in which participants interact with artificial agents that are preprogrammed with certain social behaviors. Even a step further, artificial social agents could be designed that are adaptive to human responses. For example, when an artificial agent wants to persuade humans to conserve energy, but "notices" that they are agitated, the agent could decide not trying to persuade them at that moment. In this way, it could be examined whether our results are generalizable to situations in which humans are interacting with artificial agents.

Another limitation is that we measured participants' cognitive responses, but not their actual behavior to artificial agents. We mainly used questionnaires and cognitive response measures. Except for most studies in Chapter 3, we used questionnaires to determine whether participants exhibited social responses towards artificial agents. As with all questionnaires, these measures depend on introspection, and cognitive responses have fairly low correlations with behavioral responses (e.g., Gawronski & Payne, 2010; Nisbett & Wilson, 1977). Likewise, in Chapter 3, we used a cognitive response measure to investigate whether participants inferred traits to artificial agents (i.e., our measure of social responses). This measure could tell something about human cognitions to artificial agents. With this information, we could predict their responses to artificial agents. However, future research should investigate whether our results extend to actual behavior. As we previously suggested, a study could be designed that links human social cognitions about artificial social agents with their actual behavior towards these agents.

Finally, we studied the short-term effects of human responses on artificial agents. When considering the effects of persuasive agents on human behavior, long-term effects should be studied as well. Although we did not study the long-term effects, we could speculate about important factors that could contribute to the success of persuasive agents. For example, we found that the use of language (Chapter 2) is important to consider. If humans feel threatened in their autonomy to choose, for example to save energy, it might well be that they consume even more energy than usual, just to show they have autonomy. Work that studied the long-term effects of relationships between humans and artificial agents

designed relational agents that are incorporated with factors that are important in human-human relationships (Bickmore & Picard, 2005). For example, Bickmore and Picard (2005) describe that, in the context of persuasion, it is important that an interpersonal closeness between the human and the agent exists. A recent paper about the long-term interaction effects between humans and robots gave an overview of state of the art research (Leite, Martinho, & Paiva, 2013). Leite and colleagues (2013) concluded that because of the diversity in studies it was hard to draw any firm conclusions. However, they did provide some guidelines for designing successful human-robot relationships on the long term. For example, adapt the robot's appearance and behaviors to its use: animal-inspired robots elicit care-taking responses from humans (Leite, et al., 2013). Although no conclusions about persuasive effects were given, we believe that these social robots could be used as persuasive agents. Our results could be the building blocks for these successful social persuasive agents (discussed in more detail in the following paragraph). Still, more research is needed to study whether our short-term effects on human social responses to artificial agents can be replicated for the long term.

Utilizations and Implications

The results of our studies have implications for the development of persuasive methods and human-agent interaction. Utilizing our insights may lead to effective persuasive tools for modifying behavioral patterns and to natural interaction of humans with computational devices. In the following three sections we will discuss the possible utilization of our findings in the context of reducing energy consumption in the household and address implications in a broader sense.

Persuading members of a household to save energy. The research described in this dissertation was performed in the context of a project aimed at reducing energy consumption in the household. The envisaged application comprises persuasive agents that interact with the members of a household. The agents could take the form of a physical robot or of an animated interactive face or body on a display. Our research mainly focused on the conditions under which humans respond socially to such agents. A social response signifies

a social bond and thereby provides a basis for persuasion (see also Chapter 2). The results of our experiments provide guidelines for the development of persuasive agents in the household setting and generalize to the broader context of agents in buildings or agents on mobile phones and other devices.

Our research yielded two encouraging results for the use of artificial agents in social interaction with humans. First, humans automatically respond socially to agents. This implies that evoking a social response is relatively easy and does not necessarily require advanced intelligent systems to suggest agency (e.g., intentionality or personality). Second, artificial agents endowed with simple social cues that suggest humanness (e.g., a face, a voice, or simple behavioral cues) suffice and could be used for establishing a social connection with humans. Indeed, future research should investigate whether more complex, long-term responses to artificial agents would require more advanced systems.

The experimental findings also point at two limitations of the use of artificial agents. The first limitation is that also automatic negative responses can be elicited. That is, when participants feel threatened in their autonomy by persuasive attempts of an artificial agent, they could experience psychological reactance and do the opposite of the desired behavior. Second, the social response is an automatic process that seems to assume that the artificial agent is a full-fledged social partner. Too much emphasis on the fact that the agent is an artificial entity that is controlled by an algorithm may disrupt the automatic modus, causing a controlled modus in which the human may not respond socially to the agent.

Taken together, these results lead to the following three recommendations for the creation of persuasive agents both for specific applications (e.g., reducing energy consumption in the household) and for agent-based applications in a broader sense (e.g., personalized agents in handheld devices). For the design of an artificial agent the recommendations are: (1) rely on simple social cues for achieving social responses from humans (Chapter 3), (2) keep the social cues of the agent modest as to avoid its limitations from becoming apparent (Chapter 4), and (3) ensure that the persuasive messages are

subtle and non-imposing (e.g., by using non-directive language) to prevent reactance (Chapter 2).

Artificial agents as future social interaction partners. Our results on the social effect of artificial agents could have implications for the future of our technological and social environment that exceed the domain of persuasive communication. The human susceptibility to artificial agents exhibiting simple social cues, leads to the question to what extent artificial agents can replace humans as social interaction partners?

Over the past decades several motion pictures implicitly answered this question positively (e.g., "A.I." by Steven Spielberg, 2001; "I, robot" by Alex Proyas, 2004; or "Robot & Frank" by Jake Schreier, 2012). In "I, robot", the lead character Dell Spooner (Will Smith) initially does not trust the robot Sonny but later does (Proyas, 2004). Similarly, in Robot & Frank, Frank eventually sees his robot as his buddy (Schreier, 2012). So, as far as (science) fiction movies are concerned, in the future, humans will have social relations with robots.

The results reported in this dissertation are globally in line with this view on the future. That is, humans respond automatically socially to artificial agents displaying social cues. In his analysis of future human-robot relations, David Levy (2007) suggested that these relations could even become intimate. He claimed that, although the techniques to simulate human emotions, including love, are still in their infancy, eventually *every* aspect of an intimate relationship can be artificially realized, including their physical appearance.

Levy (2007) argued that the social consequences of these developments will be far-reaching. He predicted that humans will fall in love with artificial agents and have sexual relationships with them (Levy, 2007). In contrast to Levy's futuristic visions, our results pertain to more mundane interactions between humans and artificial agents. These interactions could support humans to change their behavior in a persuasive setting. We could imagine a near future in which artificial agents help us to accomplish certain goals (i.e., to save energy, to drive safely, or to adopt or maintain a healthier life style). Our results imply that simple social cues may suffice to induce social responses in humans that may support behavioral change. For example, an implicit way to stimulate desired behavior

change could be accomplished by an artificial agent that starts smiling when the thermostat is turned down or calories are burnt. Another, more explicit way to stimulate desired behavior change could be by using artificial agents that give humans feedback, instructions or advices. In their study of agent-based feedback, Ham and Midden (2010) showed that especially negative feedback from artificial agents was effective in changing behavior (see also, Midden & Ham, 2008, Vossen, Ham, & Midden, 2009). Refining this result, in Chapter 2, we observed that it is important to keep instructions concrete and low-threatening. In this way, the chance that users feel threatened in their autonomy of choice and become psychologically reactant will be minimized. For instance, when humans do the laundry, the artificial agent could better instruct users by using more open and non-threatening language, for instance by saying, "You *could* set the temperature at 30 °C" (concrete, low-threatening instruction), than by saying, "You *have to* set the temperature at 30 °C" (concrete, high-threatening instruction). Instead of instructions, artificial agents could also give prompts by showing the desired behavior. For example, when the artificial agent notices that a human is about to do the laundry, the agent that is displayed on a tablet computer could set its own virtual washing machine to a temperature of 30 °C.

We suggest that artificial agents should adapt to the users' state of mind to maximize the chance of successful persuasion. For example, when an artificial agent notices that a user is frustrated, it may postpone its persuasive attempt, until the frustration has passed (e.g., Hone, 2006). Related to this, we speculate that persuasive attempts of an artificial agent should be non-intrusive. A non-intrusive agent is easy to ignore and is not annoying.

Beyond the domain of persuasion, artificial social agents could be used for social bonding. For example, humans can built up a relationship with an artificial agent for a more efficient behavior change. That is, we speculate that the effectiveness of an artificial agent could be higher when a social bond exists between the artificial agent and its user. Support for this comes from Cialdini (2009) who described that humans that we know and like (i.e., have a social bond with) can be more persuasive. Future research could examine whether these responses are similar when humans interact with artificial agents. Also, artificial social

agents could be used for humans that have trouble with social interaction. As already described in the introduction, persons who feel lonely, could benefit from social interactions with artificial agents. A specific user group might benefit from interaction with artificial agents. Recent studies suggested that humans with autism disorders prefer communicating via artificial agents (e.g., a computer) over communicating to humans directly (e.g., Dautenhahn & Werry, 2004; Huskens, Verschuur, Gillesen, Didden, & Barakova, 2013; Jordan, King, Hellersteth, Wirén, & Mulligan, 2013; Robins, Amirabdollahian & Dautenhahn, 2013). Humans with autism disorders may also benefit more from social interactions with artificial agents than non-autistic individuals.

Ethical considerations. As a final note on social bonding with persuasive artificial agents, we briefly address the ethical aspects of using technology to persuade humans to change their behavior. Obviously, manipulating humans to exhibit energy-saving behavior is beneficial to society. However, as with all technology, persuasive technology can be abused to control human behavior. One may object to the type of persuasive technology studied in this dissertation because it is a concealed form of behavior modification. In Chapter 1 we already addressed that artificial agents become ubiquitous. Human persuaders address groups to change their behavior, but artificial agents can be programmed to be almost anywhere, anytime and therefore can reach a far greater audience. Because humans exhibit social responses to artificial agents, this could be potentially dangerous when used for the wrong reasons (e.g., Berdichevsky & Neuenschwander, 1999). Berdichevsky and Neuenschwander (1999) therefore made a list of ethical principles when designing persuasive technology. From an ethical viewpoint it is imperative that humans are not deceived or coerced by persuasive technology to do something they rather would not want to do (e.g., Berdichevsky, & Neuenschwander, 1999; Spahn, 2012; Verbeek, 2006). Moreover, there are responsibility and privacy issues to consider. Considering the first issue, when the effect of persuasive technology results in harm, who is held responsible? For example, when a persuasive technology is designed to drive more energy-efficient, and because of this technology the driver gets into a fatal accident. Who is then held

responsible? The driver, the persuasive technology, the company that sold the persuasive technology, the programmers of the persuasive technology, the designers of the persuasive technology, or the politicians who stimulated the use of persuasive technology? This is a difficult issue to deal with (Berdichevsky & Neuenschwander, 1999). Also, Berdichevsky and Neuenschwander (1999) pointed out that the privacy of the users of persuasive technology can be in danger. That is, persuasive technology could monitor the energy consumption of a household and energy companies could use this information in their advantage. Or worse, criminals could hack into the system and use this information to determine whether household members are present or absent.

In our studies, we found that participants automatically responded socially when observing, or interacting with artificial social agents. This suggests that mere interaction with agents suffices for persuasive technology to be effective. Fortunately, we found that humans are not doomed to respond socially to persuasive agents. By cognitively focusing on the technical characteristics of the artificial agent, the social bond is reduced. Focusing on the artificial nature of artificial agents allows humans to resist social influence techniques of artificial social agents.

Finally, the development of artificial social agents that are helpful for humans clearly is an intricate endeavor that should take into account the human responses to these agents. The current research attempted to contribute to this goal by exploring the scope of these responses, by introducing new methodological approaches and by offering a closer look into the complex underlying psychological mechanisms.

References

Aron, A., Aron, E. N., & Smollan, D. (1992). Inclusion of other in the self scale and the structure of interpersonal closeness. *Journal of Personality and Social Psychology, 63*, 596–612. doi:10.1037//0022-3514.63.4.596

Bandura, A. (2001). Social cognitive theory: An agentic perspective. In S. T. Fiske (Ed.), *Annual Review of Psychology* (Vol. 52, pp. 1-26). Palo Alto: Annual Reviews, Inc. doi:10.1146/annurev.psych.52.1.1

Bargh, J. A. (1984). Automatic information processing of social information. In R. S. Wyer, & T. K . Srull (Eds.), *Handbook of Social Cognition* (pp. 1-43). Hillsdale: Erlbaum.

Bargh, J. A. (1990). Auto-motives: Preconscious determinants of social interaction. In E. T. Higgins, & R. M. Sorrentino (Eds.), *Handbook of motivation & cognition: Foundations of social behavior* (pp. 93-130). New York: Guilford.

Bargh, J. A. (1994). The four horsemen of automaticity: Awareness, efficiency, intention, and control in social cognition. In R. S. Wyer, Jr., & T. K. Srull (Eds.), *Handbook of Social Cognition* (pp. 1-40). Hillsdale: Erlbaum.

Bargh, J. A., & Pratto, F. (1986). Individual construct accessibility and perceptual selection. *Journal of Experimental Social Psychology, 22,* 293-311. doi:10.1016/0022-1031(86)90016-8

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic and statistical considerations. *Journal of Personality and Social Psychology, 51*, 1173-1182. doi:10.1037/0022-3514.51.6.1173

Bassili, J. N. (1989). Traits as action categories versus traits as person attributes in social cognition. In J. N. Bassili (Ed.), *On-line cognition in person perception* (pp. 61−89). Hillsdale, NJ: Erlbaum.

Bemelmans, R., Gelderblom, G. J., Jonker, P, & De Witte, L. (2012). Socially assistive

robots in elderly care: A systematic review into effects and effectiveness. *Journal of the American Medical Directors Association, 9,* 114-120. doi:10.1016/j.jamda.2010.10.002

Berdichevsky, D. & Neuenschwander, E. (1999, May). Toward an ethics of persuasive technology. *Communications of the ACM* (pp.51-58). New York, NY: ACM. doi:10.1145/301353.301410

Bickmore, T. W., & Picard, R. W. (2005). Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction, 12*, 293-327. doi:10.1145/1067860.1067867

Biocca, F., Harms, C., & Burgoon, J. K. (2003). Towards a more robust theory and measure of social presence: Review and suggested criteria. *Presence, 12*, 456-480. doi:10.1162/105474603322761270

Blascovich, J. (2002). Theoretical model of social influence for increasing the utility of collaborative virtual environments. *Proceedings of the 4th International Conference on Collaborative Virtual Environments* (pp. 25-30), New York, NY, USA: ACM Press. doi:10.1145/571878.571883

Brehm, J. W. (1966). *A theory of psychological reactance*. New York, NY, USA: Academic Press Inc.

Brehm, S. S., & Brehm, J. W. (1981). *Psychological reactance A theory of freedom and control.* New York, NY, USA: Academic Press.

Brehm, J. W., Stires, L. K., Sensenig, J., & Shaban, J. (1966). The attractiveness of an eliminated choice alternative. *Journal of Experimental Social Psychology, 2*, 301-313. doi:10.1016/0022-1031(66)90086-2

Broekens, J., Heerink, M., & Rosendal, H. (2009). Assistive social robots in elderly care: A review. *Gerontechnology, 8*, 94-103. doi:10.4017/gt.2009.08.02.002.00

Brown, R. D., & Bassili, J. N. (2002). Spontaneous trait associations and the case of the superstitious banana. *Journal of Experimental Social Psychology, 38*, 87-92. doi:10.1006/jesp.2001.1486.

Buller, D. B., Borland, R., & Burgoon, M. (1998). Impact of behavioral intention on effectiveness of message features. Evidence from the family sun safety project. *Human Communication Research, 24*, 433-453. doi:10.1111/j.1468-2958.1998.tb00424.x

Burgoon, M., Alvaro, E., Grandpre, J., & Voulodakis, M. (2002). Revisiting the theory of psychological reactance. Communicating threats to attitudinal freedom. In J. P. Dillard, & M. Pfau (Eds.) *The persuasion handbook: Developments in theory and practice* (pp. 213-232), Thousand Oaks, CA, USA: Sage Publications, Inc.

Carlston, D. E., & Skowronski, J. J. (1994). Savings in the relearning of trait information as evidence for spontaneous inference generation. *Journal of Personality and Social Psychology, 66*, 840-856. doi:10.1037//0022-3514.66.5.840

Carlston, D. E., Skowronski, J. J., & Sparks, C. (1995). Savings in relearning: II. On the formation of behavior-based trait associations and inferences. *Journal of Personality and Social Psychology, 69*, 420-436. doi:10.1037//0022-3514.69.3.429.

Carlston, D. E., & Skowronski, J. J. (2005). Linking versus thinking: Evidence for the different associative and attributional bases of spontaneous trait transference and spontaneous trait inference. *Journal of Personality and Social Psychology, 89,* 884 898. doi:10.1037/0022-3514.89.6.884.

CBC (2007). *What is a robot?* CBC News, In depth. Retrieved 10-02-2012, from http://www.cbc.ca/technology/technologyblog/2007/07/your_view_how_would_you_define.html

Chartrand, T. L., Dalton, A. M., & Fitzsimons, G. J. (2007). Nonconscious relationship reactance: When significant others prime opposing goals. *Journal of Experimental Social Psychology, 43,* 719-726. doi:10.1016/j.jesp.2006.08.003

Cialdini, R. B. (1993). *Influence: Science and practice*. New York, NY, USA: Harper Collins

Cialdini, R. B. (2009). *Influence: The psychology of persuasion.* New York, NY, USA: Harper Collins

Crawford, M. T., Skowronski, J. J., Stiff, C., & Scherer, C. R. (2007). Interfering with

inferential, but not associative, processes underlying spontaneous trait inference. *Personality and Social Psychology Bulletin, 33*, 677-690. doi:10.1177/0146167206298567.

Daft, R. L., & Lengel, R. H. (1986). Information richness: A new approach to managerial behavior and organizational design. In: L.L. Cummings & B.M. Staw (Eds.), *Research in organizational behavior* (Vol. 6, pp. 191-233), Homewood, IL, USA: JAI Press.

Dautenhahn, K., & Werry, I. (2004). Towards interactive robots in autism therapy: Background, motivation and challenges. *Pragmatics and Cognition, 12*, 1-35. doi:http://dx.doi.org/10.1075/pc.12.1.03dau

Decety, J., & Jackson, P. L. (2004). The functional architecture of human empathy. *Behavioral and Cognitive Neuroscience Reviews, 3*, 71-100. doi:10.1177/1534582304267187

Dehn, D. M. & Van Mulken, S. (2000). The impact of animated interface agents: A review of empirical research. *International Journal of Human-Computer Studies, 52,* 1-22. doi:10.1006/ijhc.1999.0325

Desmarais, G., Dixon, M. J., & Roy, E. A. (2007). A role for action knowledge in visual object identification. *Memory & Cognition, 35*, 1712-1723. doi:10.3758/BF03193504.

Deutsch, R., Kordts-Freudinger, R., Gawronski, B., & Strack, F. (2009). Fast and fragile: A new look at the automaticity of negation processing. *Experimental Psychology, 56*,434-446. doi: 10.1027/1618-3169.56.6.434.

Dillard, J. P., Shen, L. (2005). On the nature of reactance and its role in persuasive health communication. *Communication Monographs, 72*, 144-168. doi:10.1080/03637750500111815

Eisenberg, N. (2008) Empathy-related responding and prosocial behaviour. In G. Bock & J. Goode (Eds.), *Empathy and Fairness: Novartis Foundation Symposium 278* (pp. 71-80). Chichester, UK: John Wiley & Sons, Ltd,. doi: 10.1002/9780470030585.ch6

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of

anthropomorphism, *Psychological Review, 114*, 864-886. doi:10.1037/0033-295X.114.4.864

Epley, N., Waytz, A., Akalis, S., & Cacioppo, J. T. (2008). When we need a human: Motivational determinants of anthropomorphism, *Social Cognition, 26*, 143-155. doi:10.1521/soco.2008.26.2.143

Eyssel, F., & Hegel, F. (2012). (S)he's got the look: Gender-stereotyping of social robots. *Journal of Applied Social Psychology, 42,* 2213-2230. doi: 10.1111/j.1559-1816.2012.00937.x

Eyssel, F., & Reich, N. (2013, March). Loneliness makes the heart grow fonder (of robots) – On the effects of loneliness on psychological anthropomorphism. *Proceedings of the 8th ACM/IEEE International Conference on Human-Robot Interaction* (pp. 121-122). Piscataway, NJ: IEEE Service Center. doi:10.1109/HRI.2013.6483531

Eyssel, F., Kuchenbrandt, D., Hegel, F., & De Ruiter, L. (2012, September). Activating elicited agent knowledge: How robot and user features shape the perception of social robots. *The 21st IEEE International Symposium on Robot and Human Interactive Communication* (pp. 851-857). Red Hook, NY: Curran Associates Inc. doi:10.1109/ROMAN.2012.6343858

Field, A. (2009). *Discovering statistics using SPSS*. Thousand Oaks, CA, USA: Sage Publications Inc.

Finkel, S. E., Guterbock, T. M., & Borg, M. J. (1991). Race-of-interviewer effects in a preelection poll: Virginia 1989. *The Public Opinion Quarterly, 55,* 313-330. doi:10.1086/269264

Fitzsimons, G. J., & Lehmann, D. R. (2004). Reactance to recommendations: When unsolicited advice yields contrary responses. *Marketing Science, 23*, 82-94. doi:10.1287/mksc.1030.0033

Fogg, B. J. (2002). Computers as persuasive social actors. In B. J. Fogg (Ed.), *Persuasive*

technology: Using computers to change what we think and do (pp. 89-120). San Francisco, CA, USA: Morgan Kaufmann Publishers. doi:10.1016/B978-155860643-2/50007-X

Fogg, B. J. (2003). The functional triad: Computers in persuasive roles, *Persuasive technology: Using computers to change what we think and do* (pp. 23-30). San Francisco, CA: Morgan Kaufmann Publishers.

Fogg, B. J., & Nass, C. (1997a, March). How users reciprocate to computers: An experiment that demonstrates behavior change. *Proceedings of the Human Factors in Computing Systems Conference* (pp. 331-332). New York, NY: ACM. doi:10.1145/1120212.1120419

Fogg, B. J., & Nass, C. (1997b). Silicon Sycophants: The effects of computers that flatter. *International Journal of Human-Computer Studies, 46,* 551-561. doi:10.1006/ijhc.1996.0104

Fong, T., Nourbakhsh, I., & Dautenhahn, K. (2003). A survey of socially interactive robots. *Robotics and Autonomous Systems, 42,* 143-166. doi:10.1016/S0921-8890(02)00372-X

Gawronski, B., & Payne, B. K. (2010). *Handbook of implicit social cognition: Measurement, theory, and applications*. New York, NY: Guilford Press.

Gilbert, D. T. (1991). How mental systems believe. *American Psychologist, 46*, 107-119. doi:10.1037//0003-066X.46.2.107.

Gilbert, D. T., Krull, D. S., & Malone, P. S. (1990). Unbelieving the unbelievable: Some problems in the rejection of false information. *Journal of Personality and Social Psychology, 59*, 601-613. doi:10.1037//0022-3514.59.4.601.

Gilbert, D. T., Krull, D. S., & Pelham, B. W. (1988a). On cognitive busyness. When person perceivers meet persons perceived. *Journal of Personality of Social Psychology, 54,* 733-740. doi:10.1037//0022-3514.54.5.733

Gilbert, D. T., Pelham, B. W., & Krull, D. S. (1988b). Of thoughts unspoken. Social inference and the self-regulation of behavior. *Journal of Personality and Social Psychology, 55*, 685-694. doi:10.1037//0022-3514.55.5.685

Gilbert, D. T., Tafarodi, R. W., & Malone, P. S. (1993). You can't believe everything you read. *Journal of Personality and Social Psychology, 65*, 221-233. doi:10.1037//00223514.65.2.221

Grandpre, J., Alvaro, E. M., Burgoon, M., Miller, C. H., & Hall, J. R. (2003). Adolescent reactance and anti-smoking campaigns: A theoretical approach. *Health Communication, 15,* 349-366. doi:10.1207/S15327027HC1503_6

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science, 315*, 619. doi:10.1126/science.1134475

Guadagno, R. E., Blascovich, J., Bailenson, J. N., & McCall, C. (2007). Virtual humans and persuasion: The effects of agency and behavioral realism. *Media Psychology, 10*, 1-22. doi:10.108/15213260701300865

Ham, J., & Midden, C. (2010). A persuasive robotic agent to save energy: The influence of social feedback, feedback valence and task similarity on energy conservation behavior. In S. S. Ge, H. Li, J.-J. Cabibihan, and Y. K. Tan (Eds.). *Lecture Notes in Computer Science* (pp.335-344). Berlin Heidelberg, Germany: Springer-Verlag. doi:10.1007/978-3-642-17248-9_35

Ham, J., & Vonk, R. (2003). Smart and easy: Co-occurring activation of spontaneous trait inferences and spontaneous situation inferences. *Journal of Experimental Social Psychology, 39,* 434-447. doi:10.1016/S0022-1031(03)00033-7

Ham, J., & Vonk, R. (2011). Impressions of impression management: Evidence of spontaneous suspicion of ulterior motivation. *Journal of Experimental Social Psychology*, *47*, 466-471. doi:10.1016/j.jesp.2010.12.008

Haptek Inc. (2011). Haptek Player (Version 1) [Computer software]. Santa Cruz, CA: Haptek Inc. Retrieved from http://www.haptek.com.

Hernandez, M., & Iyengar, S. S. (2001). What drives whom? A cultural perspective on human agency. *Social Cognition, 19*, 269-294. doi:10.1521/soco.19.3.269.21468

Huskens, B., Verschuur, R., Gillesen, J., Didden, R., & Barakova, E. (2013). Promoting question-asking in school-aged children with autism spectrum disorders: Effectiveness of a robot intervention compared to a human-trainer intervention. Accepted in *Developmental Neurorehabilitation*, 1-12. doi:10.3109/17518423.2012.739212

IJsselsteijn, W., De Kort, Y., Midden, C., Eggen, B., & Van den Hoven, E. (2006). Persuasive technology for human well-being: Setting the scene. In W. A. IJsselsteijn, Y. A. W. De Kort, C. Midden, B. Eggen, and E. Van den Hogen (Eds.). *Lecture Notes in Computer Sciences* (pp. 1-5). Berlin Heidelberg, Germany: Springer-Verlag. doi:10.1007/11755494_1

Johnson, D., & Gardner, J. (2007). The mediation equation and team formation: Further evidence for experience as a ,oderator, *International Journal Human-Computer Studies, 65*, 111-124. doi:10.1016/j.ijhcs.2006.08.007

Johnson, D., & Gardner, J. (2009). Exploring mindlessness as an explanation for the media equation: A study of stereotyping in computer tutorials. *Personal and Ubiquitous Computing, 13*, 151-163. doi:10.1007/s00779-007-0193-9

Johnson, D., Gardner, J., Wiles, J. (2004). Experience as a moderator of the media equation: The impact of flattery and praise, *International Journal of Human-Computer Studies, 61,* 237-258. doi:10.1016/j.ijhcs.2003.12.008

Jordan, K., King, M., Hellersteth, S., Wirén, A., & Mulligan, H. (2013). Feasibility of using a humanoid robot for enhancing attention and social skills in adolescents with autism spectrum disorder. Accepted article in *International Journal of Rehabilitation Research*, 1-7. doi: 10.1097/MRR.0b013e32835d0b43

Kashima, Y., Kashima, E., Chiu, C.-Y., Farsides, T., Gelfand, M., Hong, Y.-Y., Kim, U., Strack, F., Werth, L., Yuki, M., Yzerbyt, V. (2005). Culture, essentialism, and agency:

Are individuals universally believed to be more real entities than groups? *European Journal of Social Psychology, 35*, 147-169. doi:10.1002/ejsp.237

Kidd, C., D., Breazeal, C. (2005). Comparison of social presence in robots and animated characters. *Proceedings of Human-Computer Interaction*

Kiesler, S., Powers, A., Fussel, S. R., & Torrey, C. (2008). Anthropomorphic interactions with a robot and robot-like agent. *Social Cognition, 26,* 169-181. doi:10.1521/soco.2008.26.2.169

King, W. J., & Ohya, J. (1996, April). The representation of agents: Anthropomorphism, agency, and intelligence. *Proceedings of the Conference on Human Factors in Computing Systems Conference* (pp. 289-290). New York, NY: ACM. doi:10.1145/257089.257326

Koda, T., & Maes, P. (1996, November). Agents with faces: The effect of personification. *Proceedings of the 5th IEEE International Workshop on Robot and Human Communication* (pp. 189-194). doi:10.1.1.30.8285

Krull, D. S. (1993). Does the grist change the mill? The effect of the perceiver's inferential goal on the process of social inferences. *Personality and Social Psychology Bulletin, 19,* 340-348. doi:10.1177/0146167293193011

Krull, D. S. (1996). On thinking first and responding fast: Flexibility in social inference processes. *Personality and Social Psychology Bulletin, 22,* 949-959. doi:10.1177/0146167296229008

Krull, D. S., & Erickson, D. J. (1995). Inferential Hopscotch: How people draw social inferences from behavior. *Current Directions in Psychological Science, 4,* 35-38. doi:10.1111/1467-8721.ep10770986

Langer, E.J. (1992). Matters of mind: Mindfulness/mindlessness in perspective, *Consciousness and Cognition, 1,* 289-205. doi:10.1016/1053-8100(92)90066-J

Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D. H. J., Hawk, S. T., & Van Knippenberg, A. (2010). Presentation and validation of the Radboud Faces Database. *Cognition and Emotion, 24*, 1377-1388*.* doi:10.1080/02699930903485076.

Lee, E.-J, & Nass, C. (1999, May). Effects of the form of representation and the number of computer agents on conformity. *Proceedings of the Human Factors in Computing Systems Conference* (pp. 238-239). New York, NY: ACM. doi:10.1145/632716.632864

Lee, E.-J. (2008). What triggers social responses to flattering computers? Experimental tests of anthropomorphism and mindlessness explanations. *Communication Research, 37,* 191-214. doi:10.1177/0093650209356389

Lee, K. M., & Nass, C. (2004). The multiple source effect and synthesized speech. Doubly-disembodied language as a conceptual framework. *Human Communication Research, 30*, 182-207. doi:10.1093/hcr/30.2.182

Lester. J. C., Converse, S. A., Kahler, S. E., Barlow, S. T., Stone, B. A., & Bhogal, R. S. (1997, March). The persona effect: Affective impact of animated pedagogical agents. *Proceedings of the 15th ACM Conference on Human Factors in Computing Systems* (pp.359-366). New York, NY: ACM. doi:10.1145/258549.258797

Levy, D. (2007). *Love and sex with robots: The evolution of human-robot relationships*. New York, NY, USA: Harper Collins

Libin, A. V., & Libin, E. V. (2004). Robotic psychology. In C. D. Spielberger (Ed.), *Encyclopedia of Applied Psychology* (Vol. 3, pp. 295-298). Oxford, UK: Elsevier.

Looije, R., Cnossen, F., & Neerincx, M. A. (2006, September). Incorporating guidelines for health assistance into a socially intelligent robot. *Proceedings of the 15th IEEE International Symposium on Robot and Human Interactive Communication* (pp. 515-520). doi:10.1109/ROMAN.2006.314441

Looije, R., Neerincx, M. A., & Cnossen, F. (2010). Persuasive robotic assistant for health self-management of older adults: Design and evaluation of social behaviors. *International Journal of Human-Computer Studies, 68,* 386-397. doi:10.1016j.ijhcs.2009.08.007

Louwerse, M. M., Graesser, A. C., Lu, S., & Mitchell, H. H. (2005) Social cues in animated conversational agents. *Applied Cognitive Psychology, 19*, 693-704. doi:10.1002/acp.1117

Lupfer, M. B., Clark, L. F., Church, M., DePaola, S. J., & McDonald, C. D. (1995). *Do people make situational as well as trait inferences spontaneously?* Unpublished manuscript, University of Memphis

Lupfer, M. B., Clark, L. F., Hutcherson, H. W. (1990). Impact of context on spontaneous trait and situational attributions. *Journal of Personality and Social Psychology, 58,* 239-249. doi:10.1037/0022-3514.58.2.239

Mayer, R. E., Sobko, K., & Mautone, P. D. (2003). Social cues in multimedia learning: Role of speaker's voice. *Journal of Educational Psychology, 95*, 419-425. doi:10.1037/0022-0663.95.2.419

McCalley, L. T., & Midden. C. (2002). Energy conservation through product-integrated feedback: The roles of goal-setting and social orientation. *Journal of Economic Psychology, 23,* 589-603. doi:10.1016/S0167-4870(02)00119-8

McCarthy, R. J., & Skowronski, J. J. (2011). What will Phil do next? Spontaneously inferred traits influence predictions of behavior. *Journal of Experimental Social Psychology, 47*, 321-332. doi:10.1016/j.jesp.2010.10.015.

Midden, C.J.H. & Ham, J.R.C. (2008, April). The persuasive effects of positive and negative social feedback from an embodied agent on energy conservation behavior. *Proceedings of the British Society for the Study of Artificial Intelligence and the Simulation of Behaviour* (pp. 9-13).

Midden, C., & Ham, J. (2009, April). Using negative and positive social feedback from a robotic agent to save energy. *Proceedings of the 4th International Conference on Persuasive Technology* (Article No 12). New York, NY: ACM doi:10.1145/1541948.1541966

Midden, C., & Ham, J. (2012). The illusion of agency: The influence of the agency of an

artificial agent on its persuasive power. In M. Bang, & E. Ragnemalm (Eds.), *Persuasive Technology: Design for Health and Safety* (pp.90-99). Heidelberg, Germany: Springer Berlin. doi:10.1007/978-3-642-31037-9_8

Miller, G. R., & Baseheart, J. (1969). Source trustworthiness, opinionated statements, and response to persuasive communication. *Speech Monographs, 36,* 1-7. doi:10.1080/03637756909375602

Miller, C. H., Lane, L. T., Deatrick, L. M., Young, A. M., & Potts, K. A. (2007). Psychological reactance and promotional health messages: The effects of controlling language, lexical concreteness, and the restoration of freedom. *Human Communication Research, 33,* 219-240. doi:10.1111/j.1468-2958.2007.00297.x

Mittelmark, M. B. (1999). The psychology of social influence and health public policy. *Preventive Medicine, 29*, S24-S29. doi:10.1006/pmed.1998.0468

Moon, Y. (2000). Intimate exchanges: Using computers to elicit self-disclosure from consumers. *Journal of Consumer Research, 26,* 323-339. doi:10.1086/209566

Moors, A., & De Houwer, J. (2006). Automaticity: A theoretical and conceptual analysis. *Psychological Bulletin, 132,* 297-326. doi:10.1037/0033-2909.132.2.297

Moreno, R., Mayer, R. E., Spires, H.A., & Lester, J. C. (2001). The case for social agency in computer-based teaching: Do students learn more deeply when they interact with animated pedagogical agents? *Cognition and Instruction, 19,* 177-213. doi:10.1207/S1532690XCI1902_02

Mori (1970). The uncanny valley. *Energy, 7*, 33-35.

Nass, C. (2004). Etiquette Quality: Exhibitions and expectations of computer politeness, *Communications of the ACM, 47,* 35-37.

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers, *Journal of Social Issues, 56*, 81-103. doi:10.1111/0022-4537.00153

Nass, C., Moon, Y., & Carney, P. (1999). Are people polite to computers? Responses to computer-based interviewing systems, *Journal of Applied Social Psychology, 29* (5), 1093-1110. doi:10.1111/j.1559-1816.1999.tb00142.x

Nass, C., Moon, Y., Green, N. (1997). Are machines gender neutral? Gender-stereotypic responses to computes to voices. *Journal of Applied Social Psychology, 27,* 864-876. doi:10.1111/j.1559-1816.1997.tb00275.x

Nass, C., Steuer, J., Tauber, E., & Reeder, H. (1993, April). Anthropomorphism, agency, and ethopoeia: Computers as social actors. *Proceedings of the Human Factors in Computing Systems Conference* (pp. 111-112). New York, NY: ACM. doi:10.1145/259964.260137

Nisbett, R. E., Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 8,* 231–259. doi:10.1037/0033-295X.84.3.231

Nolan, J. M., Schultz, P. W., Cialdini, R. B., Goldstejn, N. J., & Griskevicius, V. (2008). Normative social influence is underdetected. *Personality and Social Psychology Bulletin, 34,* 913-923. doi:10.1177/0146167208316691

Norman, G. (2010). Likert scale, level of measurement and the "laws" of statistics. *Advances in Health Sciences Education, 15,* 625-632. doi:10.1007/s10459-010-9222-y

Park, S., & Catrambrone, R. (2008). Social responses to virtual humans: Automatic over-reliance on the "human" category. In H. Prendinger, J. Lester, and M. Ishizuka (Eds.). *Lecture Notes in Computer Science* (pp. 530-532). Berlin Heidelberg, Germany: Springer-Verlag. doi:10.1007/978-3-540-85483-8_74

Petty, R. E., & Cacioppo, J. T. (1984). Source factors and the elaboration likelihood model of persuasion. *Advances in Consumer Research, 11,* 668-672.

Quick, B. L., & Considine, J. R. (2008). Examining the use of forceful language when designing exercise persuasive messages for adults: A test of conceptualizing reactance arousal as a two-step process. *Health Communication*. *23,* 483-491. doi:10.1080/10410230802342150

Quick, B. L., & Stephenson, M. T. (2007a). Further evidence that psychological reactance can be modeled as a combination of anger and negative cognitions. *Communication Research, 34,* 255-276. doi:10.1177/0093650207300427

Quick, B. L., & Stephenson, M. T. (2007b). The Reactance Restoration Scale (RRS): A measure of direct and indirect restoration. *Communication Research Reports, 24,* 131-138. doi:10.1080/08824090701304840

Quick, B. L., & Stephenson, M. T. (2008). Examining the role of trait reactance and sensation seeking on perceived threat, state reactance, and reactance restoration. *Human Communication Research, 34,* 448-476. doi:10.1111/j.1468-2954.2008.00328.x

Rains, S. A., & Turner, M. M. (2007). Psychological reactance and persuasive health communication: A test and extension of the intertwined model. *Human Communication Research, 33,* 241–269. doi:10.1111/j.1468-2958.2007.00298.x

Reeves, B., & Nass, C. (1996). *The media equation: How people treat computers, television, and new media like real people and places.* New York, NY, USA: Cambridge University Press.

Reinhart, A. M., Marshall, H. M., Feeley, T. H., & Tutzauer, F. (2007). The persuasive effects of message framing in organ donation: The mediating role of psychological reactance. *Communication Monographs, 74,* 229-255. doi:10.1080/03637750701397098

Riether, N., Hegel, F., Wrede, B., & Horstmann, G. (2012, March). Social facilitation with social robots? *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction* (pp. 41-48). New York, NY: ACM. doi:10.1145/2157689.2157697

Robins, B., Amirabdollahian, F., & Dautenhahn, K. (2013, February). Investigating child-robot tactile interactions: A taxonomical classification of tactile behavior of children with autism towards a humanoid robot. *Proceedings of the sixth international conferences on Advances in Computer-Human Interactions* (pp. 89-94). Nice, France: IARIA. ISBN: 978-1-61208-250-9

Roubroeks, M., Ham, J., & Midden, C. (2011). When artificial social agents try to persuade

people: The role of social agency on the occurrence of psychological reactance. *International Journal of Social Robotics, 3,* 155-165

Roubroeks, M., Ham, J., & Midden, C. (2010). The dominant robot: Threatening robots cause psychological reactance, especially when they have incongruent goals. In T. Ploug, P. Hasle, & H. Oinas-Kukkonen (Eds.). *Lecture Notes in Computer Science* (pp. 174-184). Berlin Heidelberg, Germany: Springer-Verlag. doi:10.1007/9783-642-13226-1_18

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist, 55,* 68-78. doi:10.1037//0003-066X.55.1.68

Saerbeck, M., Schut, T., Bartneck, C., & Janse, M. D. (2010, April). Expressive robots in education. Varying the degree of social supportive behavior of a robotic tutor. *Proceedings of the 28th ACM Conference on Human Factors in Computing Systems* (pp.1613-1622). New York, NY: ACM. doi:10.1145/1753326.1753567

Schultz, P. W., Nolan, J. M., Cialdini, R. B., Goldsteijn, N. J., & Griskevicius, V. (2007). The constructive, destructive, and reconstructive power of social norms. *Psychological Science, 18,* 42-434. doi:10.1111/j.1467-9280.2007.01917.x

Short, J. A., Williams, E., & Christie, B. (1976). *The social psychology of telecommunications.* New York, NY, USA: John Wiley & Sons.

Silvia, P. J. (2005). Deflecting reactance: The role of similarity in increasing compliance and reducing resistance. *Basic and Applied Social Psychology, 27,* 277-284. doi:10.1207/s15324834basp2703_9

Smith, C. T., De Houwer, J., & Nosek, B. A. (2013). Consider the source. Persuasion of implicit evaluation is moderated by source credibility. *Personality and Social Psychology Bulletin, 39,* 193-205. doi:10.1177/0146167212472374

Smits, M. (2006). Taming monsters: The cultural domestication of new technology. *Technology in Society, 28*, 489-504. doi:10.1016/j.techsoc.2006.09.008

Spahn, A. (2012). And lead us (not) into persuasion…? Persuasive technology and the ethics of communication. *Science and Engineering Ethics, 18,* 633-650. doi:10.1007/s11948-011-9278-y

Skowronski, J. J., Carlston, D. E., Mae, L., & Crawford, M. T. (1998). Spontaneous trait transference: Communicators take on the qualities they describe in others. *Journal of Personality and Social Psychology, 74*, 837-848. doi:10.1037//0022-3514.74.4.837

Takeuchi, A., & Nagao, K. (1993). Communicative facial displays as a new conversational modality. *Proceedings of the Conference on Human Factors in Computing Systems* (pp. 187-193). New York, NY: ACM. doi:10.1145/169059.169156

Todd, A. R., Molden, D. C., Ham, J., & Vonk, R. (2011). The automatic and co-occurring activation of multiple social inferences. *Journal of Experimental Social Psychology, 47*, 37-49. doi:10.1016/j.jesp.2010.08.006

Todorov, A., & Uleman, J. S. (2002). Spontaneous trait inferences are bound to actors' faces: Evidence from a false recognition paradigm. *Journal of Personality and Social Psychology, 83,* 1051-1065. doi:10.1037//0022-3514.83.5.1051

Todorov, A., & Uleman, J. S. (2003). The efficiency of binding spontaneous trait inferences to actors' faces. *Journal of Experimental Social Psychology, 39,* 549-562. doi:10.1016/S0022-1031(03)00059-3

Todorov, A., & Uleman, J. S. (2004). The person reference process in spontaneous trait inferences. *Journal of Personality and Social Psychology, 87,* 482-493. doi:10.1037/0022-3514.87.4.482

Twenge, J. M., Baumeister, R. F., DeWall, C. N., Ciarocco, N. J., & Bartels, J. M. (2007). Social exclusion decreases prosocial behavior. *Journal of Personality and Social Psychology, 92*, 56-66. doi:10.1037/0022-3514.02.1.56

Uleman, J.S. (1999). Spontaneous versus intentional inferences in impression formation. In S. Chaiken & Y. Trope (Eds.). *Dual-process theories in social psychology* (pp. 141-160). New York: Guilford.

Uleman, J. S., Moskowitz, G. B. (1994). Unintended effects of goals on unintended inferences. *Journal of Personality and Social Psychology, 66*, 490-501. doi:10.1037//0022-3514.66.3.490

Uleman, J. S., Newman, L. S., & Moskowitz, G. B. (1996). People as flexible interpreters: Evidence and issues from spontaneous trait inferences. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 28, pp. 211-279). San Diego, CA, USA: Academic Press. doi:10.1016/S0065-2601(08)60239-7

Uleman, J. S., Saribay, S. A., & Gonzalez, C. M. (2008). Spontaneous inferences, implicit impressions, and implicit theories. *The Annual Review of Psychology, 59,* 329-360. doi:10.1146/annurev.psych.59.103006.093707.

Van der Ploeg, H. M., Van Buuren, E. T., & Van Brummelen, P. (1988). The role of anger in hypertension. *Psychotherapy and Psychosomatics, 43*, 186-193. doi:10.1159/000287878

Verbeek, P. P. (2006, May). Persuasive technology and moral responsibility. *Proceedings of Persuasive 2006* (pp. 1-15).

Von der Pütten, A. M., Krämer, N. C., Gratch, J., & Kang, S.-H. (2010). It doesn't matter what you are! Explaining social effects of agents and avatars. *Computers in Human Behavior, 26,* 1641-1650. doi:10.1016/j.chb.2010.06.012

Vossen, S., Ham, J., & Midden, C. (2009, April). Social influence of a persuasive agent: The role of agent embodiment and evaluative feedback. *Proceedings of the 4th International Conference on Persuasive Technology* (Article No 46). New York, NY: ACM. doi: 10.1145/1541948.1542007

Vossen, S., Ham, J., & MIdden, C. (2010). What makes social feedback from a robot work? Disentangling the effect of speech, physical appearance and evaluation. In T. Ploug, P. Hasle, & Oinas-Kukkonen, H. (Eds.). *Proceedings of the 5th International Conference on Persuasive Technology* (pp.52-57). Heidelberg, Germany: Springer-Verlag Berlin

Wada, K., Shibata, T., Saito, T., Sakamoto, K, & Tanie, K. (2005, April). Psychological and

social effects of one year robot assisted activity on elderly people at a health service facility for the aged. *Proceedings of International Conference on Robotics and Automation* (pp. 2785-2790). doi:10.1109/ROBOT.2005.1570535

Waytz, A., Cacioppo, J. T., & Epley, N. (2010). Who sees human? The stability and importance of individual differences in anthropomorphism. *Perspectives on Psychological Science, 5*, 219-232. doi:10.1037/0033-295X.114.4.864

Wegner, D. M., Sparrow, B., & Winerman, L. (2004). Vicarious agency: Experiencing control over the movements of others. *Journal of Personality and Social Psychology, 86,* 838-848. doi:10.1037/0022-3514.86.6.838

Wells, B. M., Skowronski, J. J., Crawford, M. T., Scherer, C. R., & Carlston, D. E. (2011). Inference making and linking both require thinking: Spontaneous trait inference and spontaneous trait transference both rely on working memory capacity. *Journal of Experimental Social Psychology, 47,* 1116-1126. doi:10.1016/j.jesp.2011.05.013

White, R. W. (1959). Motivation reconsidered: The concept of competence. *Psychological Review, 66,* 297-331. doi:10.1037/h0040934

Wigboldus, D. H. J., Dijksterhuis, A., & Van Knippenberg, A. (2003). When stereotypes get in the way: Stereotypes obstruct stereotype-inconsistent trait inferences. *Journal of Personality and Social Psychology, 84*, 470-484. doi:10.1037/0022-3514.84.3.470

Wojciszke, B., Abele, A. E., & Baryla, W. (2009). Two dimensions of interpersonal attitudes: Liking depends on communion, respect depends on agency. *European Journal of Social Psychology, 39*, 973-990. doi:10.1002/ejsp.595

Yoo, K.-H., Gretzel, U., & Zanker, M. (2013). Source factors. In *Persuasive recommender systems. Conceptual background and implications* (pp. 9-19). New York, NY: Springer. doi:10.1007/978-1-4614-4702-3

Zajonc, R. B. (1965). Social facilitation. *Science, 149,* 269-274. doi:10.1126/science.149.3681.269

Summary

Technology that is designed to persuade people to change their behavior or attitude is called persuasive technology. This dissertation focused on persuasive technology that operates as an artificial social agent. Technical systems, like robots or virtual agents can be experienced as social agents when they exhibit social cues that imply humanlike characteristics, like for example speech. We investigated how humans respond to these artificial social agents. Earlier research showed that when social cues were used in an interaction with artificial social agents, humans exhibited social responses. However, when explicitly asked about these social responses, they showed awareness that social responses to artificial social agents were inappropriate and even denied acting socially towards these agents.

In the current dissertation we studied human reactions to artificial agents in an attempt to uncover why artificial agents can be effective as social agents. We analyzed the underlying mechanisms of social responses to artificial social agents by means of research paradigms used in human-human interaction studies of social influence and person perception. First we demonstrated social responses to a robotic agent and investigated factors that could influence these social responses. Second, we systematically compared automatic and controlled responses to humans, artificial social agents, and inanimate objects. And finally, we investigated whether it was possible for users to control these automatic responses.

More specifically, in Chapter 2 we demonstrated social responses to a robotic agent and showed that these also occur in the domain of negative responses. That is, when a robotic agent used autonomy-restricting language, participants experienced psychological reactance (i.e., feelings of anger and negative cognitions) and a tendency to restore their autonomy. These social responses are similar to how humans respond to other humans, as was demonstrated by earlier research. Furthermore, our results suggested that manipulating social agency or goal sharing (factors that could enhance social responses to artificial agents) did not intensify social responses.

In Chapter 3 we compared participant's automatic and controlled responses, using a paradigm that measures spontaneous trait inferences (STIs). People draw STIs when they observe another person's behavior, and unintentionally infer human personality traits. In contrast, trait inferences that are drawn intentionally, are called intentional trait inferences (ITIs). In our research, STIs represented participant's automatic social responses and ITIs represented participant's controlled social responses. We directly compared responses to humans, with responses to artificial social agents, and responses to inanimate objects. First of all, we found that participants drew STIs for humans, artificial social agents and even objects when these were paired with social cues. In other words, humans seem to exhibit social responses to anything that exhibits social cues. Second, at a controlled level, participants drew the strongest ITIs for humans, weaker ITIs for artificial social agents, and the weakest ITIs for inanimate objects, suggesting that artificial social agents are seen as a category in-between humans and objects. Furthermore, we assessed the nature of these automatic responses, and compared associations in memory to actual spontaneous inferences (that contain a logical, causal link between behavior and inferred trait). However, we could not find evidence for our assumption that participants merely made associations in memory regarding artificial social agents (and objects), instead of really inferring human traits. Future research should further investigate the possible associative versus inferential mechanisms of these automatic social responses.

In contrast to earlier claims, we showed in Chapter 4 that humans are not doomed to exhibit social responses. Instead, humans were able to control their automatic social responses to a certain extent. In two studies, participants were exposed to an artificial agent that both had social and technical characteristics. When participants had to focus on the technical characteristics of an artificial social agent, they were able to control for their social responses. We furthermore showed that participants did not need to completely ignore the social cues to control for their social responses (attentional focus), but could observe the social cues and then cognitively focus on the technical characteristics of the artificial social agent (cognitive focus).

In conclusion, our results extended earlier research by showing that participants automatically exhibited social responses to anything exhibiting social cues (not only robots, computers, and avatars, but also inanimate objects). Also, our results suggested that participants' controlled responses indicated that artificial social agents are difficult to categorize at a controlled level. That is, participants categorized artificial social agents neither as humans nor as inanimate objects, but rather as a category in-between humans and inanimate objects. Finally, our results suggested that participants' can control their *automatic* social responses, at least temporarily, by focusing on the technical characteristics of the artificial social agents. In the final chapter implications are discussed for the understanding of interactions between humans and artificial social agents in general, and for designers of persuasive artificial social agents in particular.

Samenvatting

Persuasieve technologie is technologie die ontworpen wordt met als doel mensen te verleiden of overtuigen hun gedrag of attitude te veranderen. Dit proefschrift richt zich op de bestudering van sociale agenten. Robots of door de computer gegenereerde (virtuele) karakters kunnen ervaren worden als sociale agenten indien de suggestie wordt gewekt dat ze menselijke eigenschappen bezitten. Deze suggestie kan worden gewekt door middel van *social cues.* Voorbeelden van *social cues* zijn: een gezicht*,* menselijke gedrag*,* en een stem. We onderzochten hoe mensen op deze *social cues* reageren. Uit eerder onderzoek bleek dat proefpersonen die deelnamen aan een experiment een sociale reactie vertoonden wanneer ze geconfronteerd werden met de *social cues* van een kunstmatige agent. De proefpersonen waren zich bewust van het feit dat sociale reacties niet vereist zijn in interacties met kunstmatige sociale agenten en ontkenden zelfs dat ze sociaal gereageerd hadden tegen deze sociale agenten.

In het huidige proefschrift bestudeerden we menselijke reacties op kunstmatige agenten, met de bedoeling te ontdekken, waarom kunstmatige agenten effectief kunnen zijn als sociale agenten. We analyseerden de onderliggende mechanismen van sociale reacties op kunstmatige sociale agenten door middel van onderzoeksmethoden die gebruikt worden in mens-mens interactiestudies op het gebied van sociale beïnvloeding en persoonsperceptie. Onze bijdrage bestond uit drie delen. Ten eerste demonstreerden we menselijke sociale reacties bij interactie met een robot en bestudeerden we de factoren die deze sociale reacties zouden kunnen beïnvloeden. Ten tweede vergeleken we systematisch de automatische en gecontroleerde reacties van mensen op (1) andere mensen, (2) kunstmatige sociale agenten en (3) levenloze objecten. Ten derde onderzochten we of het mogelijk was voor gebruikers om deze automatische reacties bewust te sturen. Hieronder geven we per hoofdstuk een meer gedetailleerde beschrijving van ons onderzoek.

In Hoofdstuk 2 demonstreerden we dat mensen sociale reacties laten zien wanneer ze interacteren met een robot, zelfs wanneer de interactie negatief is. Meer specifiek, wanneer een robot autonomiebelemmerende taal gebruikt, dan ervaren proefpersonen

psychologische reactantie, in de vorm van gevoelens van boosheid en negatieve gedachten. Deze sociale reacties zijn vergelijkbaar met de wijze waarop mensen reageren op andere mensen. Daarnaast suggereerden onze resultaten dat het manipuleren van de sterkte van de door de robot gebruikte *social cues* nauwelijks effect had op de geobserveerde menselijke sociale reacties. Ook het manipuleren van de mate van overeenkomst van doelen van de robot en doelen van de proefpersoon versterkte de sociale reacties niet.

In Hoofdstuk 3 vergeleken we de automatische en gecontroleerde reacties van proefpersonen door een methode te gebruiken die spontane gevolgtrekkingen (*Spontaneous Trait Inferences*; STIs) meet. Mensen maken STIs wanneer ze het gedrag van anderen observeren en hen spontaan (niet-intentioneel) menselijke persoonlijkheidseigenschappen toeschrijven. Gevolgtrekkingen die wel intentioneel worden gemaakt worden intentionele gevolgtrekkingen genoemd (*Intentional Trait Inferences*; ITIs). In ons onderzoek waren de automatische sociale reacties van proefpersonen STIs en de gecontroleerde sociale reacties van proefpersonen ITIs. We vergeleken beide typen van reacties op (1) mensen, (2) kunstmatige sociale agenten en (3) levenloze objecten. Onze eerste bevinding was dat voor automatische reacties, proefpersonen STIs maakten voor mensen, kunstmatige sociale agenten en zelfs voor levenloze objecten, wanneer deze gepaard gingen met sociale cues. Anders gezegd: mensen laten sociale reacties zien bij alles dat gepaard gaat met sociale cues. Onze tweede bevinding was dat, voor gecontroleerde reacties, proefpersonen de sterkste ITIs maakten voor mensen, zwakkere ITIs voor kunstmatige sociale agenten en de zwakste ITIs voor levenloze objecten. Dit resultaat suggereert dat sociale agenten gezien worden als een categorie tussen de categorieën “mensen” en “objecten” in. We vonden geen ondersteuning voor de mogelijke verklaring dat proefpersonen enkel associaties maakten in hun geheugen ten opzichte van kunstmatige sociale agenten (en objecten), in plaats van dat ze echt menselijke persoonlijkheidseigenschappen aan de kunstmatige sociale agenten (en objecten) toeschrijven.

In Hoofdstuk 4 lieten we zien dat, in tegenstelling tot claims uit ander onderzoek, mensen *niet* gedoemd zijn om sociaal te reageren. Mensen bleken in staat om hun

automatische sociale reacties tot op zekere hoogte onder controle te hebben. In twee studies werden proefpersonen blootgesteld aan een kunstmatige agent die zowel sociale als niet-sociale (technische) eigenschappen had. Wanneer proefpersonen hun aandacht richtten op de technische eigenschappen van de kunstmatige sociale agent, waren ze in staat hun sociale reacties onder controle te houden. We lieten daarnaast zien dat het voor de proefpersonen niet noodzakelijk was om de *social cues* compleet te negeren om hun sociale reacties onder controle te houden (aandachtsfocus). We vonden dat proefpersonen de *social cues* gewoon konden observeren, maar dan bewust hun aandacht konden richten op de technische eigenschappen van de kunstmatige sociale agent (cognitieve focus). Door deze cognitieve focus konden de proefpersonen hun sociale reacties op kunstmatige sociale agenten onder controle houden.

Onze resultaten verbreden eerdere onderzoeksbevindingen door te laten zien dat proefpersonen automatische sociale reacties laten zien bij alles dat *social cues* bevat, dus niet alleen robots, computers en virtuele karakters, maar ook levenloze objecten. Daarnaast laten (bewuste) reacties van proefpersonen zien dat kunstmatige sociale agenten moeilijk te categoriseren zijn op een gecontroleerd niveau. Anders gezegd, proefpersonen categoriseren kunstmatige sociale agenten niet als mensen, noch als levenloze objecten, maar eerder als een categorie tussen de categorie “mensen” en de categorie “levenloze objecten” in. Ten slotte suggereerden onze resultaten dat proefpersonen hun *automatische* sociale reacties, ten minste tijdelijk, kunnen controleren door hun aandacht te richten op de technische, niet sociale eigenschappen van de kunstmatige sociale agenten. In het laatste hoofdstuk bespreken we implicaties voor het begrip van interacties tussen mensen en kunstmatige sociale agenten in het algemeen, en meer specifiek de ontwikkeling van persuasieve kunstmatige sociale agenten.

Acknowledgements

After graduating at Maastricht University, I knew that I wanted to do a PhD project. I was very glad to start at the Eindhoven University of Technology. My project was build around a multidisciplinary team. Meetings with these team members always gave me inspiration for my experiments. Therefore, I would like to thank all the members of this team, past and current. Suzanne, Wilco, Caixia, Peter R., Ruud, Eric, Cees, Jaap, Ad, Peter B., and joining members, thank you so much for this inspiration. During my PhD project I was offered many opportunities to develop my research skills; I got the opportunity to go to several conferences, national and international; I went to summer school, spring school, winter school; and I took several courses. Therefore, A special thanks goes to Agentschap.nl and the faculty of Human-Technology Interaction who subsidized this project and made it possible for me to have this wonderful experience and deliver this well-written dissertation.

I met some very good international researchers and national researchers, also at our own university. Thanks go to Thijs Verwijmeren and Gloria Jímenez-Moya who cooperated in a study we came up with at summer school. Thanks go to Sanne Nauts, whose enthusiasm for research is contagious even when results of experiments were disappointed. Also thanks go to Daniël Lakens who always tried to help with whatever my questions were. Furthermore, I thank all wonderful colleagues at the Human-Technology Interaction group who made my time as a PhD student unforgettable: Femke, Karin, Peter, Caixia, Shengnan, Daniël, Dik, Ellen, Anita, Dirk, Suzanne, Leon, Elena, Wijnand, Yvonne, Cees, Jaap, Raymond, Ania, Marcin, Jim, Daan, Antal, Chris, Armin, Uwe, Gerrit, Ron, Wouter, Martijn, Mieke, Joris, André, Martin, Jan-Roelof, and everybody else I might forgot to mention. Especially Frank, my office mate who helped me develop some assertiveness. I often had to ask him to please be quiet, but the last months he had to ask me to be quiet as well.

Also thanks go to my committee members who gave my dissertation the finishing touch. Thank you, Vanessa Evers, Friederike Eyssel and Panos Markopoulos for taking the time to read my dissertation and give useful feedback. I mentioned my promoters and co-promotor already, but special thanks go to Cees Midden, Jaap Ham and Eric Postma for

helping me becoming the researcher I am today. Thank you so much for supporting me when I was in the flow, but also when I got stuck. I am convinced that we worked together in developing well-designed studies, giving good presentations, and publishing decent research.

Last, but certainly not least, I want to thank my family. They were always interested in the results of my experiments. And especially my boyfriend, Jeroen, who stood by me during the whole PhD process. He supported me and motivated me to be critical in my work and get the best out of it. Thank you so much, without you, I am sure this dissertation would not be of this high quality.

Curriculum Vitae

Maaike Roubroeks was born on 09 November 1984, in Vlodrop, the Netherlands.
In 2003, she received her VWO diploma from the Bisschoppelijk College Schöndeln in Roermond. In the same year she started the Psychology program at the University of Maastricht. In 2006 she received the Bachelor of Science diploma on Cognitive Psychology and in 2007 she obtained the Master of Science degree on Experimental Health Psychology.

In May 2008 she started a PhD project at Eindhoven University of Technology at Eindhoven, of which the results are presented in this dissertation. During her time as a PhD student she presented her work at a number of national and international conferences. In 2009 she received the Best Student Paper Award of the Persuasive 2009 conference. She also published several papers in conference proceedings and in a journal (see below).

Publications

Conference publications

Roubroeks, M., Ham, J., & Midden, C. (2012). Brutale Mensen, Robots, en Objecten: De Invloed van het Type Actor op Spontane en Intentionele Gevolgtrekkingen. In N. van de Ven, M. Baas, L. van Dillen, D. Lakens, A.M. Lokhorst, & M. Strick (Eds.), Jaarboek Sociale Psychologie 2011 (pp. 189-193). Groningen: ASPO pers.

Roubroeks, M., Ham, J., & Midden, C. (2012, January). *Investigating the Media Equation: The Influence of Actor Agency on Spontaneous and Intentional Trait Inferences.* Poster accepted at the meeting of the Society of Personality and Social Psychology, San Diego, CA, USA.

Roubroeks, M. (2011, November). *Designing smart agents for energy conservation*. Poster presented at the exhibition of home automation and smart living, Eindhoven, the Netherlands.

Roubroeks, M., Ham, J., & Midden, C. (2011, August). *Rude humans, robots, and objects: The influence of actor agency on spontaneous and controlled trait inferences.*

Manuscript presented at the ESCON Transfer of Knowledge Conference, Sligo, Ireland.

Roubroeks, M., Ham, J., & Midden, C. (2011, June). *Investigating the media equation hypothesis: Do we really see computer agents as human-like?* Paper presented at Persuasive 2011, Columbus, OH, USA.

Roubroeks, M., Ham, J., & Midden, C. (2011, January). *Rude humans, robots, and objects: Influence of actor agency on intentional and spontaneous trait inferences.* Poster presented at the meeting of the Society of Personality and Social Psychology, San Antonio, TX, USA.

Roubroeks, M., & Ham, J. (2010, November). *Energie besparen in de woning door middel van gedragsverandering*. Poster presented at the exhibition of home automation and smart living, Eindhoven, the Netherlands

Roubroeks, M., Ham, J., & Midden, C. (2010). The dominant robot: Threatening robots cause psychological reactance, especially when they have incongruent goals. In T. Ploug, P. Hasle, & H. Onias-Kukkonen (Eds.), *Lecture Notes in Computer Science: Vol. 6137. Persuasive Technology* (pp. 174-184). Berlin, Germany: Springer-Verlag. doi:10.1007/978-3-642-13226-1_18.

Roubroeks, M., Midden, C., & Ham, J. (2009). Does it make a difference who tells you what to do? Exploring the effect of social agency on psychological reactance. In S. Chatterjee, & P. Dev (Eds.), *Proceedings of the 4th International Conference on Persuasive Technology* (Art. 15). New York, NY: ACM. doi:10.1145/1541948.1541970

Roubroeks, M., Ham, J., & Midden, C. (2009, June). *Persuasive agents and the occurrence of psychological reactance as a result of restricting communication*. Manuscript presented at Human-Robot Personal Relationships 2009, Leiden, the Netherlands.

Roubroeks, M., Ham, J., & Midden, C. (2009, September). *Does it make a difference who*

tells you to conserve energy? Exploring the effect of social agency on psychological reactance. Manuscript presented at Environmental Psychology 2009, Zürich, Switzerland.

Journal publication

Roubroeks, M., Ham, J., & Midden, C. (2011). When artificial social agents try to persuade people: The role of social agency on the occurrence of psychological reactance. *International Journal of Social Robotics*. Advanced online publication. doi:10.1007/s12369-010-0088-1.

Appendix

Table A.1. Items of the Perceived Threat to Autonomy measure of Chapter 2, Study 1.

Item	English version (α = .87)	Dutch version (α = .90)
1	The advice restricted my autonomy to choose how I wanted to do the laundry.	Het advies beperkte mijn vrijheid om te kiezen hoe ik de was wilde doen.
2	The advice tried to manipulate me.	Het advies probeerde me te manipuleren.
3	The advice tried to make a decision for me.	Het advies probeerde een beslissing voor me te nemen.
4	The advice tried to pressure me.	Het advies probeerde druk op me uit te oefenen.

Note. The original items were from Dillard and Shen (2005) and were adapted to the context of our experiment. For example, the original item 1 was "The message threatened my freedom to choose". The alpha of the English version is from the original paper of Dillard and Shen (2005).

Table A.2. Items of the Feelings of Anger measure of Chapter 2, Study 1.

Item	English version (α = .94)	Dutch version (α = .83)
1	I was irritated	Ik was geïrriteerd
2	I was angry	Ik was boos
3	I was annoyed	Ik ergerde me
4	I was aggravated	Ik stoorde me aan iets

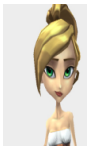
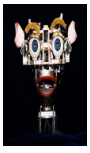
Note. The original items were from Dillard and Shen (2005) and were adapted to the context of our experiment. For example, in the original items participants were asked whether they felt "irritation, anger, annoyance, and aggravation". The alpha of the English version is from the original paper of Dillard and Shen (2005).

Table A.3. Items of the Restoration Intentions measure of Chapter 2, Study 1.

Item	English version (α = .93 – .97)	Dutch version (α = .83)
1	Right now, I am [motivated/determined/encouraged/inspired] to save energy when doing the laundry	Op dit moment, ben ik [gemotiveerd/van plan/aangemoedigd/geïnspireerd] om de volgende keer dat ik de was doe energie te besparen.
2	Right now, I am [motivated/determined/encouraged/inspired] to be around others that save energy when doing the laundry	Op dit moment, ben ik [gemotiveerd/van plan/aangemoedigd/geïnspireerd] om met andere mensen om te gaan die energie besparen als ze de was doen
3	Right now I am [motivated/determined/encouraged/inspired] to do something totally energy consuming	Op dit moment, ben ik [gemotiveerd/van plan/aangemoedigd/geïnspireerd] om iets te doen wat extra veel energie verbruikt.

Note. The original items were from Quick and Stephenson (2007b) and were adapted to the context of our experiment. The verbs between brackets were assessed on 7-point Likert scales (See Chapter 2, Study 1). The alpha of the English version is from the original paper of Quick and Stephenson (2007b).

Table B.1. Items of the Exposure Task of Study 1 of Chapter 3.

Item	Materials
1	   De vrouw/de avatar/de aanwijsstok hielp de docent om zijn presentatie te verduidelijken.
2	   De jongen/de robot/de bank stond naast de 5 hardwerkende mannen en deed niks.
3	   Het meisje/de robot/de broek wees de vrouw erop dat ze was afgevallen.
4	   Het jongetje/de avatar/de boekenkast bewaarde alle boeken op alfabetische volgorde.
5	   Het jongetje/de robot/de luxaflexgordijn sloeg in het gezicht van de moeder.
6	   Het meisje/de avatar/de algoritmische formule zorgde voor een oplossing van een complex probleem.
7	  

De vrouw/de robotde boot was al de hele wereld rond geweest.

8



De man/de robot/de rekenmachine deed precies hetzelfde als alle anderen.

9



De man/de robot/de auto gaf de jongen een lift.

10



De vrouw/de avatar/de ladder bevrijdde het kind uit de enorm hoge boom.

11



De man/de robot/de vulkaan ging als een razende tekeer.

12



De man/de robot/de vlieger vertikte het om hoger te gaan dan één meter.

ISBN: 978-90-8891-878-0

Eindhoven University of Technology
Department of Industrial Engineering & Innovation Sciences