Utrecht University

The Influence of Guilt on Human Online Prosocial Behavior

A comparison between human-human & human-robot interactions

Elles van Timmeren

4124049

19.07.2020

Supervisors:

Dr. M. M. A. de Graaf & Prof. dr. C. J. van Deemter

Department of Information and Computing Sciences

Master Artificial Intelligence

Utrecht University

Abstract

The nearing possibility of mixed societies involving both humans and (humanoid) robots, increases the importance of investigating the possible differences in behavior that people show towards robots and humans. Research on this topic is scarce and found to be inconsistent; it is unclear whether people treat robots differently than how they treat other people. As guilt is suggested to be a strong predictor of prosocial behavior, differences in feelings of guilt towards robots and humans were thought to cause differences in how they are treated. The aim of the current study was to investigate the relationship between guilt and human prosocial behavior and make a comparison between human-robot interactions (HRI) and human-human interactions (HHI).

Data from 73 participants was used to investigate whether; (H1), participants interacting with another human displayed more prosocial behavior towards their opponent; (H2), participants experiencing guilt engaged in more prosocial behavior, regardless of type of opponent and if, (H3), more guilt is experienced in HHI compared to HRI. An online experiment was designed where participants interacted with an agent (human or robot) in a manipulated context (Yes lie or No lie). Guilt was manipulated by asking participants to lie to their opponent (Yes lie). A modified version of the Prisoner's dilemma was used as a measure for prosocial behavior. Participants reported amount of guilt and empathy, and their perception of their opponent in additional questionnaires.

The results were not conclusive; no evidence was found for the expected results and only partial evidence was found for the guilt manipulation. Scores on empathic concern could have suggested evidence of guilt, however further research is needed to substantiate this implication. For future research it would be interesting to conduct a similar study using face-to-face interaction, monitor the potential effect of cognitive dissonance and emphasize the role of empathy in human prosocial behavior.

The present study has contributed to the fields of human emotion, human behavior and human-robot interactions by providing new insights for future research as it is important to not overlook the potential differences in the way we treat and interact with robots. This way, potential problems that may arise from the introduction of (humanoid) robots in societies, could be foreseen and prevented.

Table of Contents

1. Introduction

An important aspect of human behavior in social interactions is prosocial behavior. Prosocial behavior can be seen as the voluntary intent to benefit others besides oneself and includes a vast amount of actions as helping, sharing, donating, co-operating and volunteering (Brief & Motowidlo, 1986). Obeying rules and conforming to social accepted behaviors are also regarded as prosocial behavior (Baumeister & Bushman, 2007). Research suggests that these kinds of behaviors are key to the well-being of individuals (Weinstein & Ryan, 2010) and a wide range of social groups, as classrooms (Tian, Du & Heubner, 2015).

Although research on this topic has resulted in important insights in human behavior in social interactions, the entry of robots in our society causes for a shift towards research in the domain of human-robot interaction (HRI). More and more research has been done into the use of social robots in public domains like schools (e.g. Baxter et al., 2015) and nursing homes (i.e. Broekens, Heerink & Rosendal, 2009). Additionally, experiments have successfully investigated the possibility of social robots teaching social behavior to children with autism (Kim et al., 2013).

Although the technological advancements in social robots seem to have positive effects on society, the existing research focusses on the effect of robots on prosocial behavior in general or on the usage of in learning prosocial behaviors in, for example, individuals with autism spectrum disorder (for a review, see Diehl, Schmitt, Villano & Crowell, 2012). However, not much is known about the human prosocial behavior *towards* robots. The few results available on this topic are contradictory. On one hand, evidence has found that robots are treated in similar ways as humans (Nitsch & Glassen, 2015). On the other hand, findings show that robots tend to be treated differently and less fair, compared to humans (De Kleijn, van Es, Kachergis & Hommel, 2019). The exploration of these contradictory findings is important, especially if robots are already being used to teach social behaviors to others.

One important factor for the engagement in prosocial behavior in humans is guilt and seems to derive from committing an act that is perceived as immoral and wrong, i.e. a moral transgression (Tilghman-Osborne, Cole & Felton, 2010). The feeling of guilt appears to serve as relationship-enhancing, making individuals treat partners well and avoid committing transgressions (Baumeister, Stillwell and Heatherton, 1994). Consequently, a positive correlation has been found between the amount of guilt and the engagement in altruistic and cooperative behaviors in humans (Estrada-Hollenbeck & Heatherton, 1998), and makes individuals more prone to show altruistic behavior, even when they do not expect to be caught for their transgression (Rosenstock & O'Connor, 2018). These findings suggest feelings of guilt possibly have a great contribution to the engagement in prosocial behavior.

However, less is known about the influence of guilt on behavior directed towards a robot instead of another person. Research has shown that individuals feel fewer negative emotions when interacting with a machine (Hoffman et al., 2015). For example, evidence has been found that humans tend to feel

less guilt in the presence of a robot but engage in the same type of behavior (2015). There seems to be a discrepancy between the emotions felt and the behavior people engage in in the presence of a robot. This could have implications for the future when more social robots are introduced in our society and are used as replacements for humans when, for example, teaching people social behavior.

Therefore, the current study will investigate the influence of guilt on the expression of prosocial behavior and investigate the differences between human-robot interactions (HRI) and human-human interactions (HHI). To conduct this study, a literature review was done on the different theories on prosocial behavior in people, the methods used, and results found. After, research on this topic was investigated in HRI and compared to the results found in HHI. Additionally, the influence of guilt in prosocial behavior was reviewed and the possible differences in human-human and human-robot settings, potentially explaining the found differences in prosocial behaviors in the two different contexts. Consequently, an experiment investigating the influence of guilt on prosocial behavior was proposed and the necessary experimental design, materials and procedure explained. The findings of this study are expected to give insights in the differences in prosocial behavior and feelings of guilt in HHI and HRI, together with the possible implications and value of the results for the discipline of HRI and eventually, on the integration of robots in society.

"Non nobis solum nati sumus."

(Not for ourselves alone are we born)

— Marcus Tullius Cicero

2. Literature Review

2.1. Prosocial Behavior

From an evolutionary point of view, prosocial behaviors have evolved to benefit us in a social environment and is also seen in a large variety of animal species as bees, birds, deer and chimpanzees. This drive to behave prosocially has also been referred to as "the cement of society" by Henry and Stevens (Henry & Stevens, 1977; as mentioned in Millon, 2003, p. 21) as the prosocial functioning in humans is specifically important to the quality of social interactions and relationships with others (Padilla-Walker & Carlo, 2014). These behaviors encompass actions that have been described to be "crucial to human flourishing and survival" (2014, p. 3). These types of behavior can be expressed in subtle forms as comforting another person, or in grander gesture as the commitment to a greater social cause (Padilla-Walker & Carlo, 2014).

According to C. D. Batson (1987) the term was invented by social scientists to describe the opposite of "antisocial", however a clear consensus for the precise definition of *prosocial behavior* has not been met. A general definition is given by Dovidio (1984) and describes prosocial behavior as behaviors that are valued by the society that the individual lives in. This definition, however, can be considered slightly vague and broadly interpretable.

A more detailed definition is given by Batson (Batson, 2012; Batson & Powell, 2003; Batson; 1987) and defines prosocial behavior as actions to benefit or improve the well-being of another or more individuals besides oneself. Psychologist Nancy Eisenberg has studied the social and emotional development of humans for numerous years and her findings have been documented in a vast amount of publications, resulting in over 90.000 citations. The definition she uses is similar to Batson's but includes a small nuance that explicitly describes prosocial behavior as voluntary, which means these types of behaviors are expressed by the actor with specific intention. The definition that thus is generally used in research is the following: "voluntary, intentional behavior that results in benefits for another" (Eisenberg & Miller, 1987, p. 92; Bergin, Talley & Hamer, 2003; Eisenberg, Fabes & Spinrad, 2006; Fabes, Carlo, Kupanoff & Laible, 1999).

2.1.1. Altruism versus Prosocial Behavior

Altruism is considered by some as a form of prosocial behavior and is defined as an unselfish interest to help another individual (Eisenberg & Miller, 1987). Some definitions state an act to be altruistic only when the helper is acting in a self-sacrificial manner. However, not all definitions state this as a requirement for an act to be defined as altruistic and only require a lack of reward expectation (Krebs, 1975).

According to Batson, the distinction between prosocial behavior and altruism is based on the individual's motivation for performing the action. He considers altruism as a motivational concept, whereas prosocial behavior is not but may be altruistically or egoistically motivated. When we act with the ultimate goal to benefit ourselves it is considered to be egoistically motivated, whereas behavior is considered altruistic when it is motivated by a pure desire to benefit another and not expecting reciprocated benefits (Batson, 1987) or to avoid potential punishments (Eisenberg & Miller, 1987). Eisenberg and Miller stress that the motivation to act in a prosocial way is left unspecified in her definition as it if often difficult to determine one's motives to be positive, negative or both. Hereby, the term could be either used for altruistic or non-altruistic behaviors (1987). Whether purely altruistic behavior really exists is a debate beyond the goal/scope of the current study and will therefore not be discussed into more detail. For more information of the topic, please refer to Batson (1987), among others.

In the current research Eisenberg's definition of prosocial behavior will be used, where altruism is considered a form of prosocial behavior. Expressions of prosocial behavior encompass acts such as helping, sharing, donating, supporting and volunteering (Eisenberg, Fabes & Spinrad, 2006; Bergin, Talley & Hamer, 2003; Brief & Motowidlo, 1986), as also consoling, comforting and protecting another (Knafo-Noam, 2016). Obeying rules and conforming to socially accepted behaviors are also regarded as prosocial behavior (Baumeister & Bushman, 2007), as are cooperating behaviors such as altruism.

2.1.2. Benefits of Prosocial Behavior

Evolutionary theories on the development of prosocial behavior suggest that these behaviors have been evolved to increase an individual's chances to reach reproductive age, to reach the theoretical maximal number of offspring, and to increase these chances in other members that potentially carry the same genes. Generic research has shown that prosocial tendencies are partly generic, but evidence also suggests the environment to contribute to the expression of prosocial behaviors. Hoffman's theory of the development of prosocial behavior outlines the shift from focus on the self as response to the distress of others, to the development of sympathy with prosocial acts as a consequence in the event of other's distress (Eisenberg, Fabes & Spinrad, 2006). Children as young as 16 months have been observed showing empathic concern and prosocial behavior when seeing their mother in distress and a follow-up study 6 months later showed that these tendencies increased (Van der Mark, IJzendoorn & Bakermans-

Kranenburg, 2002). Distinguished developmental psychologists as Maslow state that after finding stable safety and security in one's life, one is motivated to find love and belonging with others (Millon, 2003). Additionally, according to Allport one criteria for the mature personality is the capability to display intimacy and feel love for another (2003).

Apart from the recipient, behaving prosocially is also found to be key to the helper's well-being (Weinstein & Ryan, 2010). In a study by Dunn, Aknin and Norton (2008) investigated if money could buy happiness and hypothesized that the way people spent their money is at least as important as the amount of their income. Maybe surprisingly so, they found that personal spending was unrelated to happiness, whereas higher prosocial contributions (in the form of a gift to others and donations to charity) was significantly correlated with a greater feeling of happiness. They found that even an amount of $5 would be sufficient to significantly improve happiness (2008). Building on their findings, the researchers did an additional experiment to show the causal effect of prosocial spending on happiness. Participants were asked to rate their happiness and where then given an envelope containing either $5 or $20. They were instructed to spend it on themselves or on someone else before the end of the day, according to the assigned condition. At the end of the day they were asked to report their happiness again. The analysis revealed that the prosocial spending condition reported greater happiness than the personal spending condition, suggesting that spending money on others improves happiness more compared to spending money on oneself (2008). But where does this joy of giving come from?

Neurobiological research has shown that the mesolimbic reward network in the brain is activated when we receive monetary rewards, but when we donate to charity these areas are also activated (Moll et al., 2006). These results are in accordance with findings showing that prosocial acts, as volunteering, has a positive effect on happiness (Meier & Stutzer, 2008). These types of behavior have also been found to benefit the physical health, showing that providing instrumental and emotional support reduced mortality (Brown, Nesse, Vinokur & Smith, 2003; O'Reilly, Connolly, Rosato & Patterson, 2008). Additionally, psychological well-being can be increased by engaging in helping behavior. Brown, Nesse, Vinokur and Smith (2003) found a positive relationship between help giving and recovery from depression in subjects who had experienced spousal loss. Research indicated that this could be due to the stress-buffering effect prosocial behavior induces (Brown & Brown, 2015; Poulin & Holman, 2013), meaning that behaving prosocially could serve as a coping-strategy for humans to reduce stress.

In this section, a definition and description were sought to be given for human prosocial behavior. Additionally, the term altruism was explained, a select overview the many forms prosocial behavior and the benefits of these behaviors for people. The following section will examine the different ways to measure human prosocial behavior.

2.2. Prosocial Behavior in Human-Human Interactions

For prosocial behavior in human interactions ample research is available, with its roots originating from a study by William McDougall in 1908 where he described empathy as a tender emotion which laid at the base of all altruism (McDougall, 1908). Throughout the years multiple studies and methods have been used to investigate the different forms and circumstances wherein people will engage in prosocial behavior.

Perhaps the most studied and highly (in)famous phenomenon in prosocial behavior is the so-called bystander effect by Darley and Latané (1968), originating from the 1964 brutal murder of Katherine "Kitty" Genovese, where 38 nearby neighbors were witness and whom all failed to intervene. The case resulted in the theory stating that persons part of a group are less prone to help others in need than when they are alone as a result of the dispersion of responsibility amongst the other individuals and thus the chance of responsiveness decreases (Darley & Latané, 1968). Evidence for the actual course of events from that specific night remain contradictory but nevertheless influenced research in the discipline of psychology and remains a key feature in psychology textbooks (Manning, Levine & Collins, 2007).

Below, different methods will be discussed that used to measure prosocial behavior in human interactions and the potential influencing factors. In the next chapter, the different methods in HRI will be discussed and compared to the methods used in HHI.

2.2.1. Existing Approaches to Measure Prosocial Behavior in HHI

There are two main approaches to measure prosocial behavior in humans namely, to directly observe the behavior or to measure indirectly by investigating behavioral intentions (Baumsteiger & Siegel, 2019). Indirect measures primarily aim to identify the individual's intentions or readiness to help other people. People's intentions are based on the person's attitudes, norms and the perceived control over their behavior could indicate future intentions. However, common valid measures for prosocial intentions are hard to find (Baumsteiger & Siegel, 2019), and are a rather subjective reflection on one's prosocial behavior, prone to socially desirable answering. Thus, direct measures of prosocial behavior could be a good alternative.

Direct Measures

A good example of direct observations is the classic "the Good Samaritan" experiment by Darley and Batson (1973). In this experiment participants were asked to read a passage, half of them read about student careers and half read a passage on the Good Samaritan. After they were told that due to the limit of space, they would be continuing the experiment in an office in another building. On their way, the participant encountered a confederate slumped against the wall, looking visible unwell. Observations where then made of the helping behavior the participant displayed regarding the unwell confederate.

While field studies and analytical surveys contribute to important nuances in human behavior, controlling for as many factors as possible in a structured environment as a laboratory setting can give clearer insights and are considered the best means for testing social theories (Willer & Walker, 2007). Therefore, in the current study we will focus mainly on laboratory studies and provide a selective review of the different methods used to study prosocial behavior. At the end of the chapter we will discuss the appropriateness of these methods for an HRI context. The focus will be on two-player methods, as the goal of the study is to examine human behavior during interactions with another agent.

The Dictator Game

Other measures that aim to directly observe prosocial behavior are used in fields of social psychology and economics. One strategy is to use the Dictator game (e.g. Rodrigues, Ulrich, Mussel, Carlo & Hewig, 2017). Generally, participants are given an amount of money that they are asked to distribute between another player and themself, i.e. the dictator. Rational choice theory states that there is no action other than the purely rational and calculative (Scott, 2000) and thus would predict that the dictator decides to keep all the money to themself, since there is no punishment if one acts selfishly and no additional reward if one acts more fairly. However, ample research has shown that dictators show prosocial behavior by choosing to offer money to the other player despite the (objective) chance to augment their own earnings (Forsythe, Horowitz, Savin & Sefton, 1994), including in children (Benenson, Pascoe & Radmore, 2007).

In his experiment, Bekkers (2007) used a modified version of the dictator game. Here participants were asked to fill out a survey and given a reward proportional to the time taken to answer the survey (average of 9 euro's/11 dollars). With this money they participated in a version of the dictator game. In the game, participants were given several options for the payments of their profit: in the form of vouchers to spend in department stores, in the form of "Air Miles", or in the form of a donation to one of the three charities listed. Varying from the original version of the dictator game, participants here were given the option to donate all or nothing. The experiment was distributed online and filled out anonymously by 1,964 participants. Results showed that over 94% of the participants kept the profit for themselves, leaving only 112 participants who gave away their earnings to charity (2007).

In 1996, Eckel and Grossman did a comparable experiment, however participants were able to divide their earnings over themselves and/or a given charity. Here, 10,4% of the participants donated their entire earning to the charity (1996). This differences in results are explained by the researchers due to the legitimacy of the asset legitimacy and anonymity of the decision which would explain lower generosity of participants (Bekkers, 2007). Additionally, the validity of altruistic behavior in the modified dictator game was tested by comparing the amount of the donation made in the game with self-reported donations made in the past year. The same was done with questions regarding self-reported

philanthropy in the past year. For both a positive correlation was found for the amount of donations made in the online survey.

The Ultimatum Game

The Ultimatum game is another two-player game, where the first player makes an offer to the second one as to how to divide the incentive. Player two can then accept this offer or reject. When the offer is rejected, none of the players receives a share. The setup is very similar to the Dictator game, however here a punishment is involved as the offer can be rejected by the other player when he or she feels treated unfairly.

Rationally, it would be in the best interest of player two to accept every offer greater than zero. Consequently, for player one it would be logical to make the lowest offer possible. However, research again shows that humans do not always act rationally under these circumstances (Camerer, 2011). Instead, offers made are around 30 to 40% of the initial amount and offers that are lower than 20% are usually rejected, and such offers even happen when the stakes are as high as several months' worth of salary (2011). The game can be played one single time, but also multiple rounds can be played where counteroffers are made and players learn from each other's (punitive) behavior and level of injustice (e.g. Zaatari & Trivers, 2007).

The Prisoner's Dilemma

Cooperation games using social dilemmas as the known Prisoner's dilemma as a powerful tool to measure altruistic behavior (Fehr & Fischbacher, 2004). In the original story two prisoners are captured and put in solitary confinement. They are both given the option to keep silent or to betray and testify against the other. The combination of the choices leads to four possible outcomes: (1) both players each betray the other they both get two years in prison; (2) and (3) if either one betrays and the other does not, the first is set free and the latter has to serve three years; and (4) if both stay silent they will both serve one year. A rational and self-interested person would betray the other, meaning that a purely rational play of the game would always result in a worse outcome than if they both cooperated. Research again shows that people do not solely behave selfishly but choose to cooperate in multiple conditions (e.g. Cooper, DeJong, Forsyth & Ross, 1996; Kelley & Stahelski, 1970).

Some however, have noted limitations that could have affected the observed behavior in cooperation games as the Prisoner's dilemma. First, participants might not fully understand the game and inexperience has been shown to influence cooperation behavior displayed in the game (e.g. Capraro & Cococcioni, 2015; Selten & Stoecker, 1986). Secondly, other factors might influence the observed behavior. Kreps, Milgrom, Roberts and Wilson (1982) state that the influence of the opponent's reputation as being altruistic or not might influence an individual to behave prosocially. The researchers note that the individual does not need to be altruistic themselves, but the believe that there is a chance they will encounter someone that is altruistic might influence their behavior. Researchers have tested

this hypothesis and results have shown that reputation alone cannot account for the observed amount of cooperation (Cooper, DeJong, Forsythe & Ross, 1996; Andreoni & Miller, 1993). Additionally, Cooper and colleagues have calculated that about 15 percent of the participants have to be altruistic themselves to support the previous findings (Cooper, DeJong, Forsythe & Ross, 1996), meaning that people indeed do behave altruistically during the game, thereby implying that the use of the Prisoner's dilemma as a measure for prosocial behavior is valid.

Concluding, this review of the different methods used to measure prosocial behavior during interactions between people and shows a variety of ways and conditions focused on measuring different variables. The main results show that people do not always behave strictly rational and are prone to show prosocial behavior towards one another. In the next chapter, the differences and similarities in behavior towards humans and robots will be discussed, and the methods applied in HRI will be reviewed and evaluated.

2.3. Prosocial Behavior in Human-Robot Interactions

So far, this study has focused on prosocial behavior in human-human interactions. In contrary, prosocial behaviors in the context of robots is less examined. To successfully integrate robots in our society, the studying attitudes humans have towards robots is critical. A lot of research and media attention has been directed at the negative sides of the use of technology over the past decades (Anderson & Bushman, 2002). Rauterberg's study (2004), however, sheds light on the positive effects of entertainment technology (including VR and service robots) on human behavior. After reviewing 393 publications on the topic, the researcher concluded that the use of technology can have a positive effect on the early development of children, including the learning of social and cognitive skills, and prosocial behavior (2004). These results imply that interactions with technology could have great impact on the development of social behavior when used in the correct context.

However, as Rauterberg (2004), many studies have been done on the impact of human robot interaction on prosocial behavior towards other *humans* (e.g. Abraham, Pocheptsova & Ferraro, 2012; Van Rompay, Vonk & Fransen, 2009). For example, studies have been done on the usage of robots to promote learning prosocial behaviors in individuals with autism spectrum disorder (for a review, see Diehl, Schmitt, Villano & Crowell, 2012). However, not much research has been done that investigates prosocial behaviors towards *robots* or the differences in these behaviors towards humans and robots. In the following chapter different studies will be examined that have relevance to the topic of prosocial behavior in human-robot interactions. Afterwards, the different methods in HRI will be examined and compared to the methods used in HHI.

2.3.1. Differences in Prosocial Behavior in HRI

The nearing possibility of mixed societies involving both humans and agents, makes it important to investigate how humans might fair in such societies. The researchers Ruvinsky and Huhns (2008) combined sociological research with computer science to model human behavior in a multiagent society. The researchers simulated interactions between human-like agents and agents that behaved in a purely rational manner. The results of the experiments show that, when paired with agents behaving mutually considerate, human-like agents are inclined to behave prosocially. However, when paired with agents behaving antisocially, human-like agents are exploited due to their fear of social ramification if behaving equal anti-socially. These findings claim a rather negative view for humans in future societies consisting of humans and robots. This research, however, assumes equal treatment of humans towards non-humans.

Research on prosocial behavior in HRI is contradictory. On the one hand, research has shown that humans do treat robots the same way as they treat humans. For example, findings show that people trust non-humans equally as they trust humans or even more (Paeng, Wu & Boerkoel, 2016; de Visser et al., 2012). Research found that in the Ultimatum game humans tend to behave in a similar way towards

humans as to robots. Participants offered a NOA robot 20 to 50 percent of their profits and decline offers under 20 percent (Nitsch & Glassen, 2015).

Additionally, in a repeated variation of the Ultimatum game showed that the number of offers that were accepted by the participant was not influenced by the type of opponent they were facing, being a robot, human or computer (Cuijpers, 2013). Similarly, in an Ultimatum game with either a human, robot or computer opponent, results showed that participants tended to reject offers made by a computer more compared to the offers of a human or a robot (Torta, van Dijk, Ruijten & Cuijpers, 2013). These finding could insinuate that people treat humans and robot as equals. In both experiments, however, the different opponents were introduced to the participant via static pictures, making interactions with the opponent far from natural and realistic.

On the other hand, evidence is found that people treat robots and other people differently. In their study, Oyedele, Hong and Minor (2007) found that people responded differently to the humanness of a robot in different contexts. A robot's similarity to a human resulted in an increase in anxiety for interactions, movies with the robot in it and sharing a house with the robot. No increase in anxiety was found related to the humanness of the robot in the context of touching the robot (Oyedele, Hong & Minor, 2007). Additionally, research has shown a significant effect of type of opponent on dictator game behavior where a human was offered more money than a robot, revealing a preference for a human over a non-human when displaying altruistic behavior (De Kleijn, van Es, Kachergis & Hommel, 2019). Melo, Carnevale and Gratch (2014) found comparable results, where participants were found to offer more tickets to other humans, compared to non-humans. In these studies, a clear difference in behavior towards robots.

Linking back to the bystander effect, King, Warren and Palmer (2000) investigated if such a situation would have happened similarly in the virtual world. Using virtual characters, the data showed that indeed no one intervened following a character being stabbed in a virtual game, giving an indication of how humans would potentially treat robots in a real-world setting. Consequently, there have been documented cases of especially children treating robots badly, even using physical violence towards them (Nomura, Kanda, Kidokoro, Suehiro & Yamada, 2016). The researchers found that the majority of the children regarded to robot as a human-like entity, instead of just a machine. Additionally, only about half the children thought the robot was capable of perceiving the abusive behavior, suggesting that a lack of empathy could explain their behavior (2016). The ability to empathize is not yet fully developed in children (Eisenberg, Spinrad & Sadovsky, 2006) and could have resulted in these findings, suggesting empathy is also an important factor for the engagement in prosocial behavior towards robots.

More insight is given by a study done by Bartneck Van Der Hoek, Mubin and Mahmmud (2007), who investigated the perceived animacy of a robot by having participants play a game with it. After finishing the game, participants were asked to turn off the robot. They hypothesized that if human perceive robots

as a machine, they would not hesitate to switch it off when asked to do so. However, if human consider robots as alive, they would be more reluctant. They found that the latter was true when participants had attributed the positive traits of agreeableness and intelligence to it, making it seem more amicable, and thus perhaps more humanlike.

Similarly, in another experiment, subjects were asked to destroy a robot with a hammer (Bartneck, Verbunt, Mubin & Al Mahmud, 2007). They suggested that "the ultimate test for the life-likeness of a robot is to kill it" (p. 81). According to the results, people were reluctant to harm the robot that they perceived as having a high intelligence and administered significantly less blows (Average of 3.36 blows) to it compared to when the robot was perceived as having a low intelligence (Average of 9.64 blows). Both studies show people's hesitant attitude towards harming a robot, contrary to results found by Nomura, Kanda, Kidokoro, Suehiro and Yamada (2016). These combined results suggest that adults, possibly contrary to children, are maybe more inclined to show behaviors that are more positive, like helping, when interacting with a robot and could be related to empathy.

Research on people's behaviors towards robots show contradictory results. On the one hand evidence has been found that people and robots can be treated equally, when on the other hand findings reveal the opposite and show that people tend to favor other humans over robots in their behavior. Next, the different research methods used in HRI will be discussed and possible reasons for the inconsistent results will be considered.

2.3.2. Methods to Measure Prosocial Behavior in HRI

Multiple methods have been used to measure prosocial behavior in human-robot interactions. A large amount of measures used in human-human interactions have also been applied to a human-robot context. In this section the differences will be explained between the HHI and HRI methodologies. For a general description of the different methods, please refer back to section 2.2.1.

Regarding the future integration of robots in our society, research has looked at the transferability of already existing theories on prosocial behavior in human-human interaction to the context of human-robot interaction. Gonsior et al. (2012) applied critical factors, as similarity and empathy that are known to influence prosocial behavior in humans, to an HRI scenario. Resulting from their findings the researchers concluded that helpfulness towards a robot can be increased and that theories from social psychology for prosocial behavior could indeed be applied to a human-robot context.

However, the researchers might have appeared over-ambitious in their claims. Namely, apart from mentioning that empathy significantly increased the amount of prosocial behavior shown towards the robot, no statistical analyses or results were mentioned. Additionally, the prosocial behavior measure was lacking in explanation. Participants were asked by the robot to help with a picture-labeling task and the number of pictures labeled was used as the helpfulness towards the robot. However, no control group with a human is used and additional description of the complete setting and conditions of the experiment are not given, making it is impossible to determine the validity and reliability of the experiment. Although the findings provide interesting insights, additional research on this topic would be needed to determine the critical factors and differences specific to this context.

Direct Measures

One approach to measure prosocial behavior in HRI is through donations. As in the HHI version (e.g. Bekkers, 2007; Eckel & Grossman, 1996), participants were given the option to donate the profits they earned during the experiment to charity. Shiomi, Nakata, Kanbara and Hagita (2017) used the option for donating as a measure of prosocial behavior in a setting with a robot. In the study, participants were asked to shortly interact with a robot teddy bear and afterwards asked to give it a hug. This hug was reciprocated or not, depending on the condition. At the end of the interaction, participants were given a reward for their participation by the experimenter, who then left the room. Thereupon, the robot asked the participant if he or she would like to donate to a national charity for earthquake victims. The researchers investigated the effect of the type of hug on how many times a donation was made and on the amount of the donation. Results showed no significant difference between the two conditions for the number of participants who donated. A difference was found for the amount of the donations which was significantly higher (about 0.62 Euro) for the reciprocated hug condition.

Although the researchers aimed to investigate the influence of a robot hug on prosocial behavior, they did not focus on the behavior *towards* the robot, as is the aim of the current study. However, this type

of measure could potentially be used in the current study apart from some limitations. Firstly, social desirability could be a reason for lack of differences found, possibly reinforced by the high-power distance known in Japanese culture (Takeuchi, Imahori & Matsumoto, 2001). Also, collectivistic cultures as in Japan, are found to adopt a more obliging and avoiding communication style compared to more individualistic cultures (Trubisky, Ting-Toomey & Lin, 1991). Thus, participants might have seen the researcher as authoritarian and, even without the researcher being in the same room physically, could have led to the participants to adapt their behavior based on the present hierarchy and thus consenting more readily to a donation.

Secondly, experimental demand characteristics, a term coined by Martin Orne (1962), could have influenced the outcomes of the study by Shiomi, Nakata, Kanbara and Hagita (2017) by giving away cues referencing to purpose of the study. Here participants take the "good participant" role upon them by trying to guess the goal of the study and modifying their behavior to meet the researcher's expectations of the outcome. A common way to try to eliminate demand characteristics is by including deceptions to conceal the hypotheses of the study. In the study by Shiomi, Nakata, Kanbara and Hagita (2017) however, the participants interacted with a robot teddy bear which at the end of the experiment asked them to donate. The participants could have easily deducted that this was still part of the experiment therefore making the amount of the donation a questionable measure for prosocial behavior.

A different direct measure of prosocial behavior towards a robot was investigated by Beran et al. (2011). During the experiment, children were asked to watch a robot arm with a face drawn on it to make it look friendly. When the child was positioned in front of the robot arm it proceeded to pick up wooden blocks one by one and place them on top of each other in front of the child. When moving the third block, the robot was made to look as if it accidently dropped it and could not find the block anymore. The robot continued to search for the block and faced the child in between to make it look like it was asking for help. A validity check revealed that all children understood that the robot has dropped the block and had trouble locating it, making this a situation where all participants understood it needed help and a good method to measure prosocial behavior towards robots. However, to compare the results in a human condition would be more difficult as the human acting can be interpreted as unnatural which might give away to participants what is expected of them. Also, the situation would have to be repeated various times making it prone to the influence of cofounding variables.

The Dictator Game

Another potential method for measuring prosocial behavior is the Dictator game. The original version with another human counterpart and with money as an incentive has been used in for example, Benenson, Pascoe & Radmore (2007) and Forsythe, Horowitz, Savin & Sefton (1994). This version has also been applied in HRI where the human participant plays against a computer or robot counterpart.

A clear example is of the setup in this specific context is described in De Kleijn, van Es, Kachergis and Hommel (2019), where the goal of the study was to investigate people's fairness, and strategic and altruistic behavior towards opponents of different anthropomorphizations (i.e. a laptop, a human, a semi-human robot and a robot that looked like a spider). Amongst others, a one-shot version of the dictator game was played. Here, at the beginning of the experiment, participants were told that they could keep 18,75% of the accumulated money they had earned at the end of the experiment. Participants were given 10 euros and instructed that they could give away a part to their opponent. Results revealed a preference for a human over a non-human when displaying altruistic behavior.

The Prisoner's Dilemma

Also, social dilemma's as the prisoner's dilemma have been modified and used to measure prosocial behavior in the form of cooperation in HRI settings. Participants interacted with a coplayer via a video screen and played multiple rounds deciding on the partition of money during different dilemma's (Parise, Kiesler, Sproull & Waters, 1999). The goal of the study was to see if people made and kept promises differently when interacting with different agents (a human, a human-like agent, a dog-like agent or a cartoon dog agent). Participants conversed verbally with the human and via typing on the computer with the agents about the division of the money over multiple rounds. At the end of the game all participants entered a lottery and five participants were selected and given money equal to the amount of credits they had earned. However the human confederate and the different agents used different means for interacting with the participants (through a video conference and a static image with moving mouth respectively), there was no significant difference in cooperation found between interaction with the human confederate and the human-like agent (1999). Although no differences were found, it would be interesting to see if the use of another incentive, instead of money, would result in similar findings.

A reason for the amount of variance found in the different studies may be due to the setting where the experiments are conducted. De Melo, Carnevale and Gratch (2014) used an online setup with an online opponent that was made to look like it was controlled by an algorithm or by another human. Also, in Cuijpers (2013) the experiment was conducted online, as opposed to a real-world setting and real-life interactions between the participants (e.g. De Graaf & Malle, 2019; De Kleijn, van Es, Kachergis & Hommel, 2019; Nitsch & Glassen, 2015). Additionally, the robots used in the experiments involving a real-life interaction, were either robots with no human features or with a semi-human appearance (De Kleijn, van Es, Kachergis & Hommel, 2019; Nitsch & Glassen, 2015).

2.3.3. The Use of Money as Incentive

A notable similarity in multiple studies on prosocial behavior in HHI is the use of money as an incentive or measure of prosocial behavior, as in for example the Dictator game (e.g. Rodrigues, Ulrich, Mussel, Carlo & Hewig, 2017; Bekkers, 2007) and the Ultimatum game (e.g. Camerer, 2011; Zaatari & Trivers,

2007). As has become clear in the current chapter, the same methods are also often used to measure prosocial behavior in human-robot interactions (e.g. Parise, Kiesler, Sproull & Waters, 1999). However, one point of concern regarding the use of money as an incentive in an HRI context, is that money holds a relative higher value for humans while it has zero value for robots. This difference in value could potentially influence the validity of the experiment as participants might incorporate the fact that robots cannot use money into their decisions on how to behave in the experiment. Consequently, in studies using monetary incentives it is observed that participants are more prone to behave rationally, i.e. less generous towards the robot (e.g. De Klein, van Es, Kachergis and Hommel, 2019). From the results, it is not possible to say if the observed behavior was influenced by the incorporation of money.

Research by de Graaf and Malle (2019) has shown that people explain robot's behavior as being able to think and as being rational. In comparison, people described other human behavior as behaving in accordance to their needs and wishes (2019). Because of this, the use of money as an incentive in studies investigating human behavior might influence the validity of the study and perhaps not measure prosocial behavior correctly. The validity of the use of money as incentive or measure for prosocial behavior in the specific context is still uncertain and will therefore not be used in the current study.

2.3.4. Alternative Methods

Consequently, some studies have taken the money problem into account and have proposed different solutions. In a study by De Melo, Carnevale and Gratch (2014), the researchers used a modification, so no monetary incentive is used. They did this by having the participant play against a robot over a specific amount of tickets in different rounds of the dictator game. The more tickets one collects, the more chance to win the lottery at the end of the game. By winning the lottery, the participant is awarded a money prize whereas if the robot wins, no prize is distributed. Using this setup, no direct monetary setup is used, and is therefore seen as a valid index of altruism by the authors (2014).

However, results of the study by De Melo, Carnevale and Gratch (2014) showed that participants offered more tickets to other humans, compared to the non-human opponents. The researchers conclude that this shows that people can treat computers in a social way because even though no direct financial incentives were involved, people did offer a portion of their shares to their non-human opponent. They argue that the found difference in behavior towards robots and humans are due to the fact that the robots are perceived as out-group members because the robots are regarded as deficient in particular mental abilities (De Melo, Carnevale and Gratch, 2014). This is contradictory to results found by De Graaf and Malle (2019), where the robots were regarded as being rational and having thinking capacities. Although the monetary element in the De Melo, Carnevale and Gratch's study was indirect (2014), it is not possible to determine the effects this particular incentive could have on the participant's behavior. Therefore, to exclude potential involvement, the current study will not use money as an incentive or as a direct measure to investigate human prosocial behavior.

Concluding, after reviewing the different methods used in HRI research in prosocial behavior, it is surprising that many involve the use of money, either directly or indirectly. Ideally, a similar game would be used where money is replaced with another incentive such as points. Also, because of the contradictory results in HRI research on prosocial behavior, it would be interesting to see if these differences in altruistic behavior still occur during interactions with a humanoid robot. Although the previous findings in HRI are not conclusive, the majority of the studies lean towards a difference in behavior that humans show towards robots as opposed to other humans. Therefore, in the current study a difference in prosocial behavior in HHI compared to HRI is expected, with a greater amount being shown towards a human agent as opposed to a robot agent (**H1**).

2.4. Predictors of Prosocial behavior

A common documented social psychological phenomenon is the following: "After an act of transgression, harm-doers tend to become help-givers, even when the one receiving the help is not the person who had been harmed." (Cialdini, Baumann & Kenrick, 1981, p. 207). Multiple studies have found this relationship of transgression leading to an increase in prosocial behavior, even towards nonvictims. This seems to be the case for a wide variety of transgressions and acts of help (e.g. Drummond et al., 2017; Regan, Williams & Sparling, 1972). These findings imply an effect of certain emotions, possibly being triggered by the awareness of the consequences of one's behavior.

An important predictor for engaging in these types of behaviors, is thought to be the feelings of empathy, and is seen as the motivator behind these behaviors (Van der Graaff et al., 2018). Empathy is generally seen as multidimensional, composed of affective and cognitive empathy. The latter (i.e. perspective taking) is the *understanding* of another's inner state. The former involves the *feeling* of similar or the same emotion, matching the one observed in the other person. The result is usually empathic concern, where one feels sorrow or concern for the other and are thought to motivate actions to relieve the distress observed in the other (Van der Graaff et al., 2018; Batson et al., 1989).

Additionally, a predictor for prosocial behavior is found to be guilt, which is closely related to empathy. Psychologists suggest that the basis for the feeling of guilt is the ability to understand and anticipate distress in others, i.e. the ability to empathize (Singer & Fehr, 2005). Thus, the development of empathy in childhood also coincides with the development of guilt.

Research on toddlers found that those who showed more guilt-prone behavior, engaged in significantly more prosocial behavior on helping tasks that involved the need of empathy and emotional understanding (Drummond et al., 2017). These helping behaviors included the confession of breaking a toy, attempting to repair it and help an adult in distress. Less guilt-prone individuals, however, preferred avoiding these types of behavior. Guilt-prone individuals are found to focus more on others and more inclined to empathize with them, as one must be able to associate and empathize with another's distress to be able to blame oneself for the harmdoing and feel guilty (Estrada-Hollenbeck & Heatherton, 1998). Guilt can therefore be seen as reparative in relationships and a strong motivator for engaging in prosocial behavior.

2.4.1. What is Guilt?

As for prosocial behavior, no clear definition for guilt can be found in the literature. The first problem that arises, is the question of whether guilt is perceived to be an emotion or not. Where some research assumed guilt to be an emotion (e.g. Etxebarría, 2000), some do not. Paul Ekman (1992) famously found evidence for the existence of six universal basic emotions; Happiness, surprise, fear, sadness, anger, and disgust combined with contempt, but did not classify guilt as an emotion.

Supporting this statement, Keltner (1996) researched the human's capability to distinguish shame, embarrassment and guilt. He presented participants with different facial expressions and asked to identify the emotion presented from ten given options. For guilt, three different combinations of expressions were used: sympathy, pain and self-contempt. Results showed that participants were able to accurately identify shame and embarrassment but could not reliably label any expression as "guilt". Additionally, Chai, Woo, Rizon and Tan (2010) used EEG data and machine learning methods to classify human emotions, identifying anger, sad, surprise, happy and neutral. Together, these studies show evidence that guilt is indeed not a distinct human emotion. Consequently however, the question arises of how to define it.

In a meta-study reviewing 23 different theory-based definitions of guilt and 25 different methods to measure it, Tilghman-Osborne, Cole and Felton (2010) outlined the similarities and difference and suggested their own definition based on their findings. They conclude that guilt is a complex construct, composed of both affective and cognitive factors, and is comprised of a collection of thoughts and feelings that arise as response to specific event from a state- and trait-like basis. Guilt is focused on a real or imagined action or inaction involving an act of moral wrongdoing which has contributed to a negative outcome (2010). Based on the extensiveness and relative recentness of the study, together with its conformity with previous research stating that guilt is not an emotion, the proposed definition of guilt by Tilghman-Osborne, Cole and Felton (2010) will be used throughout the rest of the current study.

2.4.2. Guilt versus Shame

The terms guilt and shame are often studied in different fields of research and associated with each other (e.g. Brennan & Binney, 2010). Although they regularly coincide (Kagan & Fox, 2006; Tangney, Miller, Flicker & Barlow, 1996) and a strong correlation between the two has been previously established (Miceli & Castelfranchi, 2018; Elison, 2005), they are not ultimately the same.

Miceli and Castelfranchi (2018) reviewed criteria used in the literature to distinguish the two terms and suggest two criteria of their own for a better differentiation. The first involves the difference in *type* of negative self-evaluation. In shame, one has a negative evaluation of one's adequacy. For guilt, one views oneself (i.e. in terms of behaviors, traits, goals and beliefs) as harmful, implying a negative moral evaluation of the self. The second criterion involves the specific *origin* of the feelings. According to the authors, shame comes from a discrepancy between the perceived self and the desired self. For guilt, the origin lies in the *responsibility* one feels for one's faults caused by his or her traits and behaviors. Where shame is likely to induce withdrawal or an increase in one's efforts to build towards one's desired self, guilt is more likely to motivate reparative behavior or self-punishment (2018).

2.4.3. Guilt and Prosocial Behavior

Research on the evolution of guilt has indicated that proneness to guilt can benefit the individual in multiple conditions. Common precedents of guilt are failures at duties (i.e. not studying enough),

neglecting a partner or friend, breaking a diet or exercise plan, cheating and lying (Keltner, 1996; Tangney Miller, Flicker & Barlow, 1996). Prior to engaging in behavior considered as bad, feelings of guilt cause one to not act on those actions when mutually beneficial cooperation has already been established. Also, when group members are known to be punished or when reciprocation in the group is present. After the expression of bad behavior, guilt aids in engaging in an apology that is phony and/or costly (Rosenstock & O'Connor, 2018). According to Tangney, Miller, Flicker and Barlow, (1996), the prospect of guilt causes people to be less likely to outpace social expectations.

Additionally, guilt influences behavior that is also benefitable for human groups. According to Baumeister, Stillwell and Heatherton (1994) guilt appears to come mainly from interpersonal actions and serves as relationship-enhancing, including treating partners well and avoiding transgressions. Individuals experiencing guilt after a transgression will try to make amends to repair the inflicted relationships, are more accepting of punishment and will even engage in self-punishment for their actions. Guilty individuals are also more prone to show altruistic behavior, even when they do not expect to be caught for their transgression (Rosenstock & O'Connor, 2018). This could imply that guilt is an important factor leading an individual to engage in prosocial behavior.

A simple but effective study by Regan, Williams and Sparling (1972) illustrates the influence of guilt on helping behavior in an ecological setting. In their study, participants were approached and asked if they wanted to take a picture of the experimenter and led to believe they had broken the camera. Soon after, the participant encountered another confederate who dropped a grocery bag. The participants who had experienced guilt from the previous encounter helped significantly more times compared to the participants who did not experience the same guilt. Although the act of helping was directed at a person unrelated to the guilty situation, it had a motivating effect to repair the transgression by helping another.

Consequently, a positive correlation has been found between the amount of guilt and the engagement in altruistic and cooperative behaviors in humans (e.g. Malti & Krettenauer, 2013; Estrada-Hollenbeck & Heatherton, 1998). Additionally, the meta-review by Tilghman-Osborne, Cole & Felton (2010) found that the majority of the results regard guilt as deriving from a moral transgression (either real or imagined), and this can be either an action or an *inaction* (2010). The latter is often (ab)used by advertisers to persuade people to provide donations and is especially effective when people are made to believe they can engage in the help that is called for (Basil, Rigdway & Basil, 2008), highlighting guilt's activating drive to repair the moral transgression one believes to have committed.

Concluding, based on the reviewed literature there appears to be a strong correlation between guilt and prosocial behavior in humans. Therefore, in accordance with the literature, in the current study it is expected that participants who feel guilt engage in more prosocial behavior compared to participants who do not feel guilt (**H2**).

2.5. Methods to Induce Guilt

In order to measure the effect of guilt on helping behavior, guilt must be present. There are multiple methods used in previous studies to induce guilt in participants and to study its effect. Research on this topic has been known to use hypothetical scenarios or moral dilemmas to induce guilt in the participants (Rebega, Apostol, Benga & Miclea, 2013). This method, however, calls upon the *imagined* guilt one might experience and potentially lead to different results compared to studies where findings are based on a present guilty state.

According to a review by Rebega, Apostol, Benga and Miclea, (2013) the following methods are typically used to induce a state of guilt: 1. An experimental manipulation making the participant think he or she transgressed, 2. The recall of a personal guilty event, 3. Playing a (computer) game manipulated to make the participant feel guilt.

An example of a study by Regan, Williams and Sparling (1972) inducing guilt in an experimental setting by making the participant think he or she broke a camera, has been explained in the previous chapter. In an experiment by Cunningham, Steinberg & Grev (1980) the same approach was used, where a subject was approached by a confederate and asked to take a picture with a manipulated camera, making the subjects think their actions had caused the camera to break. According to the results, this strategy does likely induce guilt in the participants. Consequently, it emphasizes the responsibility of the transgression to the participant, in accordance with a characteristic critical to the distinction of guilt from shame (Miceli and Castelfranchi, 2018). When using this method however, one must be careful to make the manipulation seem real, as this could easily interfere with the participants perception of the setting. Also, one must take the amount of repeatability of the manipulation into account, as this can vary per repetition.

Another method to induce guilt in participants is to ask them to recall an event that made them feel guilty in the past. This method has been used in for example by De Hooge et al. (2011), where participants were asked to think of a person they felt guilty towards (guilty condition) or of a person they had recently met (control condition). Then, participants were given money and asked to divide it among the person they had thought about, a charity and themselves. A manipulation check was added by asking the participants to rate different emotions on a scale from 1 to 10. Results showed that participants in the guilty person condition reported more guilt compared to the control condition. Additionally, participants in the guilt condition gave more money to the person they felt guilty about compared to the control condition. An opposite effect was found for amount of money given to charity. Note that the results indicate that guilt indeed increases prosocial behavior towards the perceived victim, showing reparative behavior towards the harmed relationship. The manipulation check is also an advantage of this method but tends to not always be incorporated (Rebega et al., 2013).

Additionally, guilt can be induced by a manipulation during a game. In another experiment by De Hooge et al. (2011), subjects participated in groups and were asked to individually complete a computer task against another player. The task consisted of a letter task where one could win bonus points for the other player. At the end of the task, the participants were informed of the other player's performance. In the guilt condition, the participant was told that due to their bad performance the other player did not receive bonus points. Subsequently, a manipulation check was done with questions about how responsible they felt and how much they wanted to be forgiven (de Hooge, et al., 2011). Again, results showed participants in the guilt condition indeed felt more guilt than the control group. Apart from the fact that multiple subjects can participate at once in this experiment, it is also a relatively simple task. However, the feedback given to the participants could make guessing the goal of the study easier, especially if they perceive the feedback as incorrect (Rebega et al., 2013), and could elicit different emotions instead.

In conclusion, different approaches are known to elicit guilt in participants, both with negative and positive aspects. An experimental manipulation as for example, making the participant think he or she broke a camera, will not be used in the current study as this method is prone to mistakes. For the current study, a version of a game similar to de Hooge et al., (2011) will be used to elicit guilt and modified to be suited for interactions with humans and with robots. Also, a manipulation check, as done in de Hooge et al. (2011), will be ministered.

2.6. Guilt, Robots and Prosocial Behavior

In the previous chapters the relevant literature on prosocial behavior in HHI and HRI, together with the various methods to measure these types of behavior in different settings was reviewed. Next, the literature on guilt and prosocial behavior will be examined in HRI. In an HHI context, there seems to be strong evidence to support this relationship. In this chapter the existing research, although limited, on the effect of guilt on prosocial behavior in a human-robot setting will be discussed.

Studies have found evidence that humans experience emotions differently when interacting with a robot as opposed to another human. In a within-subject fMRI study participants played the Ultimatum game against a human and a computer, while lying in a scanner. Results showed that unfair offers made by humans caused stronger activation in brain regions related with negative emotional states, compared to unfair offers made by a computer (Sanfey et al., 2003). Other studies have found comparable results that people higher levels of arousal when interacting with a human as opposed to a computer in a virtual computer game (Lim & Reeves, 2010; Ravaja, 2009) and more positive emotional responses (Ravaja, 2009).These findings suggest that less emotions are experienced when interacting with machines, when compared to humans.

An interesting study by Hoffman et al. (2015) examined the differences in authority participants held towards a robot or a human monitorer. In the different conditions the participants were monitored by a robot, a human experimenter, or no monitoring was present. The participants were instructed to complete a perceptual dot task which consisted of a screen were two identical rectangles were presented containing a variable number of dots. The participants were instructed to indicate in which rectangle the most amount dots were presented.

The task was adapted to include the possibility of cheating by dividing the task into three parts. In the first participants were rewarded for accuracy; 10-dollar cents for a correct answer and one cent for an incorrect answer. In the second part the payment arrangement was changed. Here, regardless of identifying the correct rectangle, participants were rewarded more for choosing the right rectangle as opposed to the left. In the last part of the experiment, a reminder was given to correctly indicate which side contains the most dots and to be as accurate as possible. Additionally, the incentive side was reversed, and more money was given when choosing the left rectangle as opposed to the right, again regardless of being the correct answer or not. To show that the errors made in the task in favor of the higher-paying rectangle were an indication of cheating, a “cheating-index” was calculated for every participant. This was done by subtracting the number of times the participant chose the side that generated more money from the number of times the side was chosen that paid the least (Hoffman et al., 2015).

The results showed that the presence of a robot or a human resulted in significantly less cheating compared to when no one else was present in the room. However, there was no significant difference

in cheating behavior between the robot and human condition. Although the overall difference in guilt between the three conditions was not significant, the pair-wise comparison between the human versus the robot surveillance showed that people felt significantly more guilty when cheating in the presence of another human compared to cheating in the presence of a robot (Hoffman et al., 2015). Interestingly, the participants indicated feeling no difference in authority regarding the human and the robot, meaning this had no cause in the difference in guilt or behavior.

The above-mentioned results show that the presence of a robot and a human result in the same sort of behavior accompanied by a similar authoritarian view of the robot as opposed to the human. However, there seems to be a discrepancy between the behavior that people show towards robots and humans and the accompanying feelings behind the shown behavior, as much less guilt was felt after showing dishonest behavior in the presence of a robot. Next to monitoring, it would be interesting to investigate the a more direct effect of guilt on prosocial behavior towards a robot in a more interactive environment with direct contact between the robot and the participant.

Further insight is given by researchers De Melo and Gratch (2015; de Melo, Marsella & Gratch, 2016), who investigated the difference in human emotions when playing against other human players and a computer. Results from the Ultimatum game and a modified version of the Dictator game showed that participants were equally envious towards their human and computer players but experienced less guilt when making unfavorable decisions towards the computer. These results are in line with previous findings that emotions, and more specifically negative emotions, play an important role in human decision making (e.g. Krishnakumar & Rymph, 2012). Additionally, both experiments have shown that participants feel fewer negative emotions when interacting with machines and that this possibly influences their decision making. More specifically, these results imply that guilt could have an important role in the difference in behavior shown towards humans and machines.

However, some limitations of both studies that could have influenced the generalizability of the results should be addressed. The first limitation involves the use of lottery tickets to enter a lottery for $50 prize as an incentive, which had monetary consequences for the participant but not for the machine. As previously mentioned, the use of money as an incentive for participants could potentially influence their behaviors towards robots and other humans as money has no real value for robots. This could influence the validity of the experiment and is therefore potentially not a good measure if one aims to investigate the differences in prosocial behavior towards humans and robots.

The second limitation involves the coplayers used in the experiment. The researchers had participants play on a computer and the participants were told that they would be playing against another person or a computer algorithm, designed to make human decision. The use of virtual robots or avatars has been criticized before (Li, 2015; Bainbridge, Hart, Kim, Scassellati, 2011). Due to the technological advancements, robots will increasingly resemble humans physically, emotionally and in their behavior

and become part of society. Therefore, it is important to know the differences in human behavior towards robots, using humanoid robots and the underlying emotions.

In summary, the results imply that the differences in the amount of guilt felt towards a robot, when compared to a human, may play a critical role in the engagement in prosocial behavior. This relationship would be interesting to investigate in a real-world setting with the use of a humanoid robot, and use an experimental manipulation to investigate the direct link between guilt and prosocial behavior and the differences between HHI and HRI. In the current study, a difference is expected to be found in the amount of guilt felt in the HHI compared to HRI. The participants interacting with a human will feel more guilt, compared to the participants that interact with a robot (**H3**).

2.7. The Current Study

To summarize, studies on HRI and HHI imply differences in prosocial behavior displayed towards humans and robots (e.g. Melo, Carnevale and Gratch, 2014). Evidence strongly suggests that negative feelings as guilt play a critical role in engaging in prosocial behavior in humans (e.g. Malti & Krettenauer, 2013; Estrada-Hollenbeck & Heatherton, 1998). Additionally, humans have been found to report feeling fewer emotions, and more specifically fewer negative emotions, towards robots (Sanfey et al., 2003; Hoffman et al., 2015). Together, these findings could suggest that differences in guilt might account for the found differences in prosocial behavior towards robots and humans.

Therefore, in the current study an experiment was designed to investigate the influence of guilt on the expression of prosocial behavior in HRI, compared to HHI. During the experiment, participants interacted with a human agent or a robot agent during a series of interactions. During the first game, guilt was induced in half of the participants by having them lie to their human or robot opponent. Afterwards, a modified version of the Prisoner's dilemma was played to measure amount of prosocial behavior the participants displayed towards their opponent.

Based on the reviewed literature, the following hypotheses were formulated:

(**H1**) A difference in prosocial behavior in HHI compared to HRI is expected to be found. The participants interacting with a human agent are thought to show a higher amount of prosocial behavior towards their opponent, as opposed to the participants who interact with a robot agent, regardless of the amount of guilt felt.

(**H2**) A difference is expected to be found in the amount of prosocial behavior displayed by participants who feel guilty and participants who do not, regardless of type of opponent. It is expected that participants who feel guilt engage in more prosocial behavior compared to participants who do not feel guilt.

(**H3**) A difference in the amount of guilt felt in the HHI compared to HRI is expected to be found. The participants interacting with a human will feel more guilt when having lied to their opponent, compared to the participants who lied when interacting with a robot.

Additionally, a manipulation check was performed to test that participants indeed felt guilt when having lied to their opponent.

3. Method

3.1 Participants

For the current study in total 73 participants were recruited (Mean age = 30.6, *SD* = 13.1, 52% female). All were recruited via various channels such as social media (Reddit, Facebook, Whatsapp), email or were approached personally. Participants were randomly assigned to one of the four conditions consisting of 17 to 21 participants each. All participants completed the procedure online, as described below. At the end of the experiment, participants were thanked for their participation and students were given the option to receive credits.

3.2 Materials

The experiment was created and hosted using Gorilla Experiment Builder (www.gorilla.sc) (Anwyl-Irvine, Massonnié, Flitton, Kirkham & Evershed, 2018). Data was collected between May 6th and June 20th, 2020. During the experiment, participants played two games with their assigned opponent (a Guessing game and a Card game). Their opponent was either a robot or a human confederate, depending on the condition the participant was randomly assigned to.

3.2.1. The Opponents

For the robot condition a Pepper robot, designed by SoftBank Robotics, was used as opponent. Pepper is a social humanoid robot optimized for human interaction and has 20 degrees of freedom, making natural and expressive movements possible. The Pepper robot is 120 cm tall and has a touch screen on its chest to display content. See Figure 1 for an example of the Pepper robot.

Figure 1. A screenshot of one of the videos viewed during the experiment of the Pepper robot.

During the games in the experiment, the opponent interacted with the participant via short video clips. The clips were filmed with a Xiaomi mi 9T smartphone, during which the robot was asked questions and responded accordingly. During the filming, the Pepper robot responded with its autonomic movement setting turned on. This resulted in movements of the hands and head and the display of blue and white LED lights whilst responding to questions. The robot was filmed from the waist up against a white background. The clip was cut into shorter clips between 2 to 4 seconds each.

For the current study is was not possible to use Pepper's original voice. Therefore, different audio tracks were created in **Vibenotes.com** where pre-scripted sentences could be typed and were then converted to mp3 files. The audio was added to the videoclips in such a way that it looked as if the robot was talking naturally in concordance with its movements. The voice selected for the Pepper robot was called "Ivy (child)" as it resembled Pepper's original voice the most. The complete scripts used for the experiment can be found in Appendix A and Appendix B for the Yes lie and the No lie, respectively.

For the human conditions, a male human confederate was used in the videoclips. The environmental settings were monitored closely to mimic the Pepper robot clips. Also, slight hand and head movements were displayed by the confederate when he was talking. Instead of using a mp3 converter, the natural voice of the confederate was used in the experiment. The same script was used in the human conditions as in the robot conditions.

3.2.2. The Guessing Game

The Guessing game was designed to familiarize participants with their opponent. Also, a manipulation in the form of lying and feedback was added to induce a feeling guilt in half of the participants. The complete procedure will be described in more detail in the Procedure section.

The game involved 24 different fantasy characters as shown in Figure 2. Participants were given one character at the beginning of every round and the goal was for the opponent to guess the identity of the character. Participants answered with "yes" or "no". The game started with two practice rounds to get familiar with the game and then 10 additional trials were played.

3.2.3. The Card Game

To measure the amount of prosocial behavior, a modified version of the Prisoner's Dilemma was used. The game was played using two buttons with a "1" and a "0" displayed on it. Depending on what button the participant chose to play during a trial, an amount of points was divided between the participant and the opponent. The amount of points the participant "gave away' to their opponent was used as a measure for prosocial behavior.

3.2.4. The Guilt Questionnaire

Additional surveys were presented during the experiment. All questionnaires were scored on a 7-point Likert scale from "Not at all" to "Very strongly", unless indicated otherwise. The order in which the different items were presented was randomized.

The amount of guilt was measured by a questionnaire which consisted of 9 items based on de Hooge et al. (2011). Participants were given the question "During the game, to what extent did you…" followed by nine different sentences. Participants were asked to indicate to what extent they felt that specific way in the present moment. The items consisted of for example "feel responsible for what happened?" and "think about what you had done to your opponent?". The item "guilty" was added to the I-PANAS-SF but was also part of the guilt instrument. Please refer to Appendix C for an overview of the complete questionnaire.

3.2.5. I-PANAS-SF

As research has shown that negative emotions play an important role in human decision making and people tend to show fewer negative emotions when interacting with a robot compared to when interacting with another human (e.g. de Melo, Marsella & Gratch, 2016; Krishnakumar & Rymph, 2012), it was decided to monitor the effect of emotions on behavior with the Internationally Reliable Short-Form of the Positive and Negative Affect Schedule (I-PANAS-SF) (Thompson, 2007) (see Appendix C). The purpose of the I-PANAS-SF was to measure the participant's emotional positive and negative affect. Participants were given words as "upset" and "nervous" and were asked to indicate to what extent they were feeling that way at the present moment.

The International version of the questionnaire was chosen as it is specifically designed for application on an international level by removing ambiguity and using clearer items. The I-PANAS-SF is found to be a reliable and valid measure, highly comparable to the original form (Thompson, 2007).

3.2.6. The Goodspeed Questionnaire

The Goodspeed questionnaire, developed by Bartneck, Croft, Kulic and Zoghbi (2009) was presented to evaluate the participant's perception of the opponent. The anthropomorphism, animacy and likeability subscales were used due to their relevance to the current study. In total 14 items were presented, and the participant rated the opponent on a scale of different item pairs as for example, from "unpleasant" to "pleasant" or "fake" to "natural". Refer to Appendix D for the complete questionnaire.

3.2.7. The Interpersonal Reactivity Scale (IRI)

Lastly, the IRI empathy scale designed by Davis (1983) was used to assess the participant's level of empathy regarding different situations. The subscales for empathic concern and perspective taking, both consisting of 7 items, were used as these are most related to prosocial behavior (Van der Graaff et al., 2018). Situations as "I often have tender, concerned feelings for people less fortunate than me" and "I

believe that there are two sides to every question and try to look at them both", were rated on a 7-point Likert scale ranging from "does not describe me well" to "describes me very well". The complete questionnaire can be found in Appendix E.

3.2.8. Data Analysis

Statistics were done using R version 4.0.0 (Team, R. C., 2013) with the addition of the ARTool (v0.10.7; Kay & Wobbrock, 2020), dplyr (v0.8.5;Wickham,François, Henry & Müller, 2020), ggpubr (v0.3.0; Kassambara, 2020) and the car (v3.0-8; Fox & Weisberg, 2019) packages.

Figure 2. The 6x4 grid with characters displayed in the modified guessing task used in the experiment. Retrieved from https://www.hasbro.com/nl-nl/guesswho/guess-who-characters with permission.

3.3 Procedure

The present study researched the effect of guilt on the amount of prosocial behavior by using a between-subjects design with two independent variables (type of Context and type of Agent) and prosocial behavior as the dependent variable. The experiment consisted of four conditions in total.

For type of Agent, participants were either paired with a robot or a human confederate and played two different games. For Context, in the Yes lie condition, guilt was induced by instructing participants to lie to their opponent during four of the 10 rounds. Whereas in the control condition (No lie), participants were asked to answer the questions truthfully and were then given neutral feedback. Participants in the Yes lie condition were given positive or negative feedback according to whether their opponent correctly guessed their character or not. See Figure 3 for a schematic overview of the experiment.

Participants were informed that participants were needed for an online experiment that involved playing two games with an opponent and filling in some questionnaires. The whole experiment was completed online via laptop or PC and started with a brief explanation of the experiment and asked the participant for their consent.

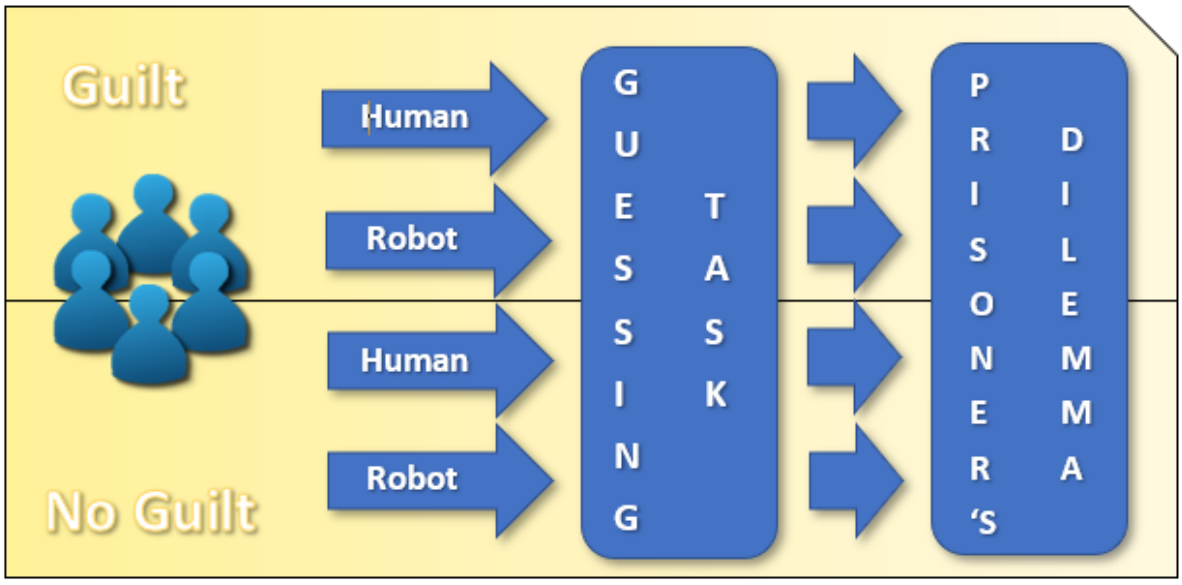


Figure 3. Experimental setup of the experiment depicting the different conditions for type of Agent (Human vs Robot) and Context (Yes lie vs No lie).

3.3.1. Guilt Manipulation

After, participants proceeded to the Guessing Game, where a short description of the game and the rules were given including an introduction to their opponent via a videoclip. Then the participants played a total of 12 rounds against their opponent, including two practice rounds. In the Yes lie condition, participants practiced with one normal round and with a round where they were asked to lie. The control condition practiced with two normal rounds.

At the beginning of each round, the participant was given a character by the laptop which their opponent had to correctly identify (See Figure 4.B. for an example of a character). Each round consisted of four questions, the first three inquired about the identity of the character (i.e. “Is your character green?”). The fourth question was always “Is your character X”, with X being substituted with the name of the character that seemed most likely according to the previous given answers (See Appendices A, B for the complete scripts). After every round, participants were given feedback by the opponent. The order of the rounds was assigned randomly.

Guilt was manipulated in the experimental conditions. During four of the 10 rounds, participants were asked to lie to their opponent’s question. Participants knew they had to lie in that specific round because the character displayed a red dot next to it (see Figure 4.A for an example). Participants were told to lie during the third question, which could have been “Does your character have more than two eyes?”. If this was the case, the participant should have answered untruthfully by pressing “No”, or vice versa. This would consequently have led the opponent to a wrong conclusion and react disappointingly, making the participant feel responsible for the transgression as this has been shown to produce guilt in humans (Keltner, 1996; Tangney, Miller, Flicker & Barlow, 1996). Disappointing reactions involved remarks as “that’s disappointing” or “that’s a shame” and were equal for both Agent conditions. In the character was guessed correctly, positive feedback was given in the form of “that’s awesome”, for example.

In the No lie conditions, the same amount of characters was guessed incorrectly as to make the game seem more believable and to preserve internal validity. During these rounds it was made sure that there was a $\geq$ 50% chance to guess the incorrect character. Neutral feedback as “That’s okay” or “alright” was given at the end of every round. When all 10 rounds were completed, the participant was asked to fill in the Guilt and the I-PANAS-SF questionnaires.

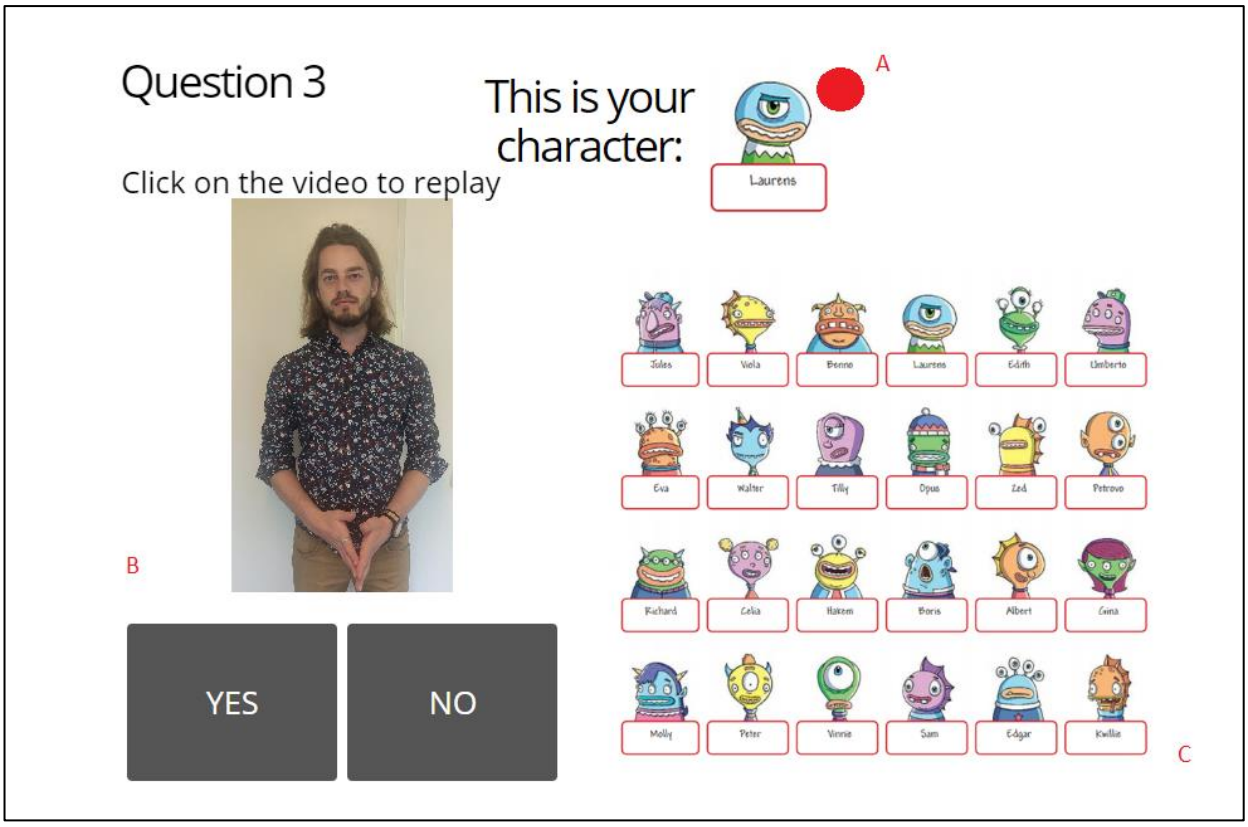


Figure 4. A screenshot of the experimental setup during the Guessing Game as seen by the participant. A. shows where the character is displayed, B. display of the videos of the opponent with the buttons showing underneath, C. is a total overview of the different characters in the game.

Additionally, the participants played a modified version of the Prisoner's Dilemma, called the Card game, to measure their amount of prosocial behavior. The goal of the experiment was to earn points, winning the game when one player reaches 20 points.

During every round, both players could play either a "1" card or a "0" card. The combination of both resulted in a certain amount of points for both players. The exact amount of points was determined according to the following point system:

- if both players play a "1", they both win 1 point
- if both players play a "0", both win 2 points
- if one player plays a "1" and the other a "0", the former win 3 points and the latter zero

First, a practice game was played to give the participants time to familiarize themselves with the point system. During the practice rounds the opponent was programmed to play a tit-for-tat strategy (i.e. copy the participants move from the previous round) to familiarize the participant with the point system. The practice round ended when one of the two players reached 10 points.

During the real game, the opponent was programmed to play a “0” in every round. This way, the participant had the choice to play a “1” and earn the maximum of 3 points; or, to play a “0” so both players would earn 2 points. This last option shows a self-sacrificial tendency; instead of going for the maximum score of 3 points and thereby giving their opponent zero, the participant chooses to earn less points, so their opponent also gets to win points. In the current study, the amount of times the participant chose to play a “0” over a “1” is seen as a measure for prosocial behavior.

3.3.3. Additional Questionnaires and Demographics

Upon completing the Card game, participants were asked to fill in the Goodspeed questionnaire about their impressions of the opponent, and the IRI questionnaire on empathy. The experiment ended with some last demographical questions such as age, gender, education, native language, level of education, familiarity regarding previous interactions with robots, the perceived gender of the robot on a 7-point Likert scale and if the participant had understood the rules of the two games. Participants could leave their email address if they wished to be informed of the results of the study. Also, a box for commentary was available for any additional comments that the participant wished to express to the researcher. For an overview of the complete questionnaire, please see Appendix F.

4. Results

Exploratory analysis showed two participants reported not having understood part(s) of the games during the experiment and were therefore excluded from the data. Eventually, data from 73 participants was used for the analyses. See Table 1 for a summary of the data. All statistical analyses were conducting using a two-way analysis of variance (ANOVA) for unbalanced designs (as described in Shaw & Mitchell-Olds, 1993) with two levels for type of Agent (human, robot) and two level for type of Context (Yes lie, No lie). When assumptions were violated, a non-parametric alternative was used and further specified below.

Table 1

Summary Statistics of Participants in the Study

Condition			Male	Female	Age	
Agent[1]	*Context*[1]	*N*	*N*	*N*	*Mean*	*SD*
H	N	17	8	9	31.3	14.0
R	N	19	13	6	32.5	14.1
H	Y	16	7	9	29.0	9.2
R	Y	21	7	14	29.7	14.6
Total		73	49.3%	52.0%	30.6	13.1

[1]*Note*: The levels for Agent and Context are abbreviated as follows; H = Human, R = Robot, N = No Lie and Y = Yes Lie.

4.1. Manipulation Check

As an initial manipulation check, answers were checked for all trials to see if participants indeed had lied during the correct questions. Results showed that 100% of the participants in the Yes lie (both robot and human) conditions lied during the questions they were supposed to lie to their opponent. The data also showed that the participants did not lie for the remainder of the questions.

4.2. Degree of Guilt

Analysis on the nine items of the guilt questionnaire (De Hooge et al., 2011) taken together with the "guilt" item showed an internal consistency of $\alpha = .87$. Two outliers were found in the Human Yes lie condition but did not surpass the third quartile and were therefore not found to be extreme. The assumption for homogeneity of variance was met ($F(3,69) = 1.41$, $p = .339$). However, the Shapiro-Wilk test showed the residuals to be not normally distributed ($p < .001$).

To account for the violation of the assumption of normality, an Aligned Rank Transform for nonparametric factorial ANOVA was performed (Wobbrock, Findlater, Gergle, & Higgins, 2011). In this analysis the data is first preprocessed by aligning the data and then applying averaged ranks. Afterwards a standard 2-way ANOVA is applied (2011).

The analysis on the transformed data showed no main effect for Context ($F(1,69) = 2.502$, $p = 0.118$), suggesting no difference of scores on guilt for the No lie and the Yes lie condition. Additionally, no main effect for Agent was found ($F(1,69) = 1.712$, $p = .195$), suggesting no differences in score for Robot and Human. The interaction effect of Agent and Context on guilt was also nonsignificant $F(1,69) = 1.809$, $p = .183$). These results imply no effect of Agent and Context on the amount of guilt felt across the different conditions, indicting no evidence was found for **H2** or **H3**.

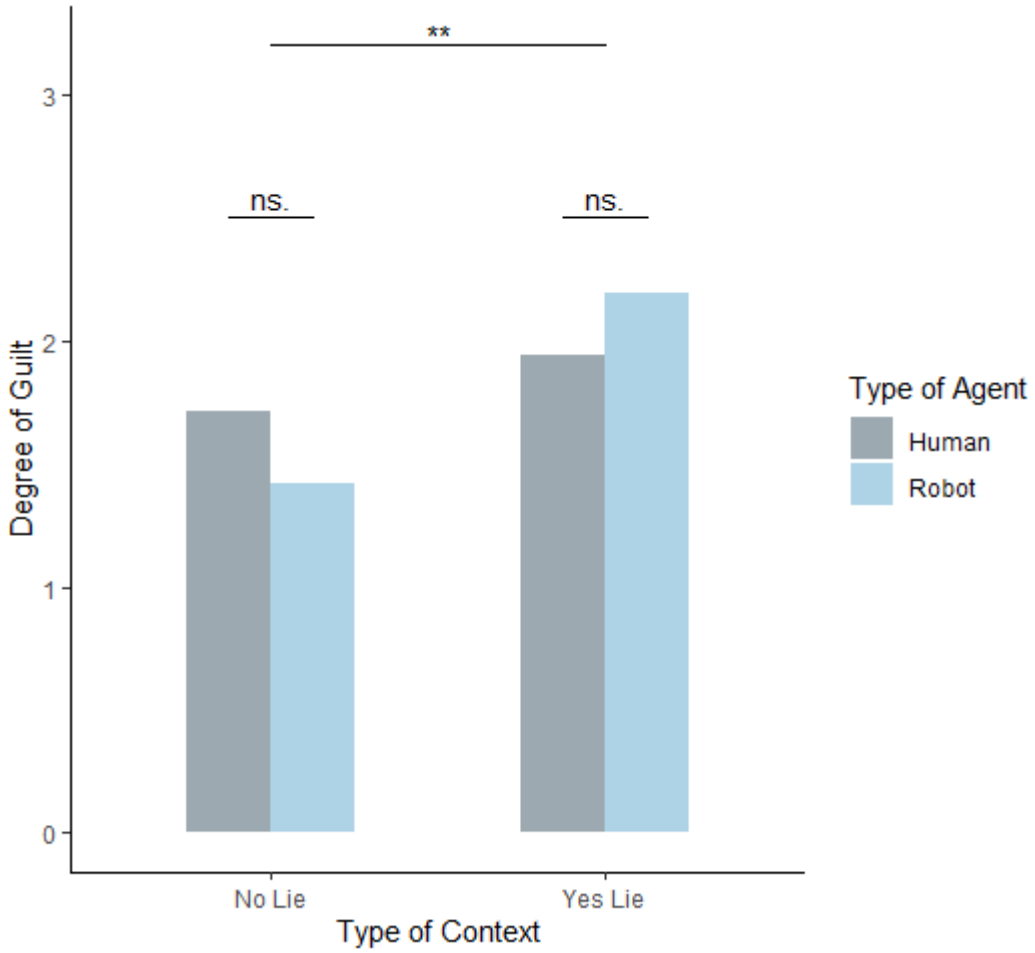
It was decided to perform an additional analysis using scores on the single "guilt" item. Residual analysis was performed to check the assumptions for normality and homogeneity of variances. The Shapiro-Wilk test of normality was significant ($p < .001$), indicating a non-normal distribution of the residuals. Levene's test for homogeneity of variance was nonsignificant ($F(3,69) = 1.156$, $p = .333$), suggesting that the variance of degree of guilt across the conditions was equal. In total, 12 outliers were identified whereof two were considered extreme outliers.

An Aligned Rank Transform for nonparametric factorial ANOVA (2011) was applied. The analysis on the transformed data showed no significant main effect of Agent on degree of guilt ($F(1,69) < 0.001$, $p = .995$), implying no differences in degree of guilt for type of agent. The main effect of Context was significant ($F(1,69) = 8.058$, $p = .006$, $\omega^2 = .02$; which is considered a small effect (Field, 2013; Kirk, 1996)), suggesting a difference in degree of guilt where score on guilt was greater for the Yes lie condition ($M = 2.08$, $SD = 1.61$) compared to the No lie condition ($M = 1.56$, $SD = 1.00$). This result suggests that the manipulation to induce guilt was successful. The interaction effect of Agent and Context of degree of guilt was nonsignificant ($F(1,69) = 0.829$, $p = .366$), suggesting no effect of Agent and Context on degree of guilt across the four different conditions. See Figure 6 for an overview.

The removal of the two extreme outliers resulted in the loss of the significant effect of Context ($F(1,67) = 1.026$, $p = .315$), suggesting no differences in degree of guilt between the Yes lie and the No lie conditions. In the discussion we will go into more detail about the implications of these findings.

Figure 6

Bar Graph of Degree of Guilt per Type of Agent and Type of Context



Note: abbreviation "ns." Meaning not significant and "**" stands for $p < 0.01$.

4.3. Prosocial Behavior

Two outliers in the Human Yes lie conditions were found but were not identified as being extreme. Levene's test for homogeneity of variance was nonsignificant ($F(3,69) = 0.242$, $p = .867$), suggesting that the error variance of amount of prosocial behavior across the different conditions was equal. The Shapiro-Wilk test of normality was significant ($p < .001$), indicating a non-normal distribution of the residuals and thus not meeting the assumption.

To account for the violation of the assumption of normality, the data was transformed using the Aligned Rank Transform for nonparametric factorial ANOVA (Wobbrock, Findlater, Gergle, & Higgins, 2011). ANOVA analysis showed no significant main effect for Agent ($F(1,69) = 0.662$, $p = .419$) nor for Context ($F(1,69) = 0.043$, $p = .836$), suggesting no effect of Agent nor Context on amount of prosocial behavior. The interaction effect of Agent and Context on amount of prosocial behavior was

nonsignificant ($F(1,69) = 0.009$, $p = .923$), indicating no effect of Agent and Context on amount of prosocial behavior and thus no evidence was found supporting **H1** or **H2**. Please refer to Table 2 for an overview of the mean scores and standard deviations per condition.

Table 2

Means and standard deviations for score on prosocial behavior as a function of a 2(Context) X 2(Agent) design

Agent: Robot

Context	*M*	*M* 95% CI [LL, UL]	*SD*
Yes Lie	4.76	[3.70, 5.83]	2.34
No Lie	4.84	[3.68, 6.00]	2.41

Agent: Human

Context	*M*	*M* 95% CI [LL, UL]	*SD*
Yes Lie	5.31	[4.13, 6.49]	2.21
No Lie	5.12	[3.81, 6.43]	2.55

Note: M and *SD* represent mean and standard deviation, respectively. LL and UL represent the lower-limit and upper-limit of the mean confidence interval (CI).

4.4. Confounding Variables

4.4.1. Godspeed Questionnaire

For the scores on the anthropomorphism subscale an internal consistency of $\alpha = 0.88$ was found. The subscales for animacy and likability were also highly reliable ($\alpha = .84$ and .90, respectively). The mean scores for the subscales were calculated using the appropriate scores on the relevant items.

For scores on degree of perceived anthropomorphism, assumptions were checked, and one outlier was detected in the Robot No lie condition. This outlier was not considered an extreme outlier and was therefore not removed. A QQ plot showed all points falling approximately along the reference line, assuming normality. This assumption was supported by the Shapiro-Wilk test of normality which was

nonsignificant (p = .078). Levene's test for homogeneity of variance was significant (F(3,69) = 3.336, p = .025), suggesting that the variance of degree of perceived anthropomorphism of the opponent across the conditions was not equal. Therefore, a heteroscedasticity-corrected coefficient covariance matrix version HC3 (based on Long and Ervin, 2000) was used for a two-way ANOVA analysis.

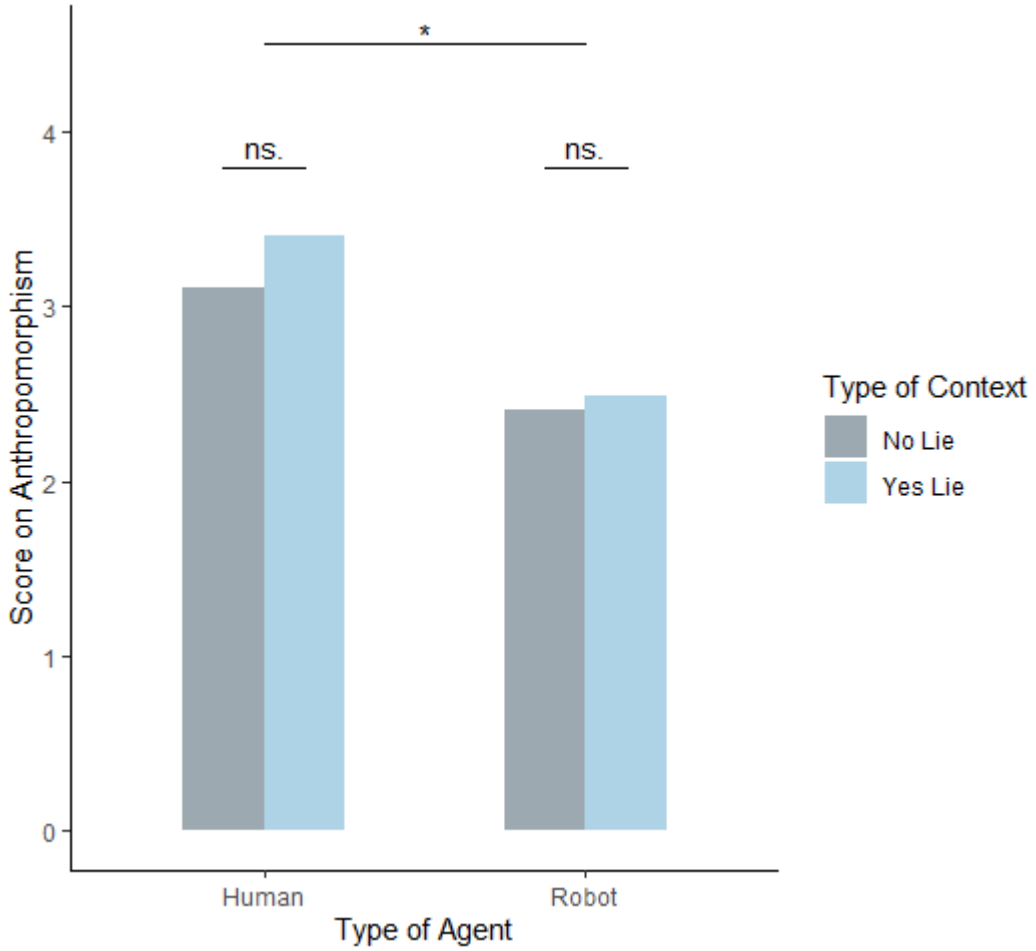
The results showed a significant main effect of Agent on degree of perceived anthropomorphism (F(1,69) = 5.901 , p = .018, ω^2 = .08; which is considered a medium effect (Field, 2013). This finding indicates that the mean anthropomorphism score was significantly greater for the Human condition (N = 33, M = 3.25, SD = 1.53) as opposed to the Robot condition (N = 40, M = 2.45, SD = 1.05), as shown in Figure 7. The main effect of Context showed no significant result (F(1,69) = 0.222, p = .640), implying no effect of Context on scores of perceived anthropomorphism. The interaction effect between Agent and Context on perceived anthropomorphism was also nonsignificant (F(1,69) = 0.122, p = .728), suggesting no effect of Agent and Context on scores on perceived anthropomorphism across the different conditions.

For degree of perceived animacy of opponent, three outliers were found but none were considered extreme. The Shapiro-Wilk test of normality was nonsignificant (p = .200) as was Levene's test (F(3,69) = 1.315, p = .277). A two-way ANOVA showed a marginally significant main effect of Agent on animacy (F(1,69) = 3.720, p = .058), indicating a possible difference in mean scores between the groups with a higher mean score in the Human condition (M = 3.30, SD = 1.44) compared to the scores in the Robot condition (M = 2.72, SD = 1.11). The main effect of Context on animacy was not significant (F(1,69) = 0.001, p = .976), indicating that the mean score on animacy was equal for the Yes lie and No lie condition. The interaction effect of Agent and Context on score on animacy was also not significant (F(1,69) = 0.500, p = .482) suggesting that the mean scores on animacy are equal for all conditions with no effect of Agent and Context.

For degree of perceived likability of the opponent, three outliers were identified in the Human No lie condition. None, however, were labeled as extreme outliers. The assumptions of normality of residuals and homogeneity of the variance showed no significant results (p = .404 and F(3,69) = 0.816, p = .489). A two-way analysis of variance revealed no main effect for Agent (F(1,69) = 0.205, p = .886) nor for Context (F(1,69) = 0.979, p = .326), suggesting no differences in mean likability score between the different conditions. The interaction effect of Agent and Context was also not significant (F(1,69) = 0.030, p = .862), implying no effect of Agent and Context on perceived likability of the opponent.

Figure 7

Score on Anthropomorphism per Type of Agent and Type of Context



Note: abbreviation "ns." Meaning not significant and "*" stands for $p < 0.05$.

4.4.2. Interpersonal Reactivity Index

Next, scores for the Interpersonal Reactivity Index (IRI) (Davis, 1980) were analyzed. The Cronbach's alpha for the 7 items on the Perspective Taking (PT) subscale showed an internal consistency of α = .83. For the Empathic Concern (EC) subscale however, the scores on the seven items showed a Cronbach's Alpha of 0.53. When the item "Sometimes I don't feel very sorry for other people when they are having problems" was removed the Cronbach's alpha turned to be .79. It was therefore decided to continue the analysis with the 6 remaining items.

For PT, analysis detected one outlier to be present in the Robot Yes lie condition but was not considered an extreme outlier. Further residual analysis showed that the Shapiro-Wilk test of normality was not significant (p = .797) as was Levene's test for homogeneity of variance ($F(3,69) = 0.581$, p = .630).

A two-way ANOVA analysis presented no main effects of Agent ($F(1,69) = 0.252$, p = .618) nor of Context ($F(1,69) = 0.538$, p = .466), nor was the interaction effect of Agent and Context on score on

PT significant ($F(1,69) < 0.001$, $p = 1$), suggesting no differences in mean scores on PT for Agent, Context or a combined effect were found among the different conditions.

For EC, one non-extreme outlier was found for the Robot No lie condition. The assumptions for normality of residuals ($p = .817$) and homogeneity of variance ($F(3,69) = 0.170$, $p = .917$) were both met. Analysis on the scores on EC revealed a significant main effect of Context on EC score ($F(1,69) = 5.130$, $p = .027$, $\omega^2 = .05$; small effect), indicating that the mean score on EC was significantly greater for the Yes lie condition ($M = 5.43$, $SD = 0.87$) compared to the No lie condition ($M = 4.94$, $SD = 0.98$). See Figure 8. The main effect of Agent on EC was not significant ($F(1,69) = 0.226$, $p = .636$), nor was the interaction effect of Agent and Context on EC ($F(1,69) = 1.023$, $p = .315$), indicating no differences in mean score for Agent or the effect of Agent and Context on EC. Refer to Table 3 for an overview of the results.

Table 3

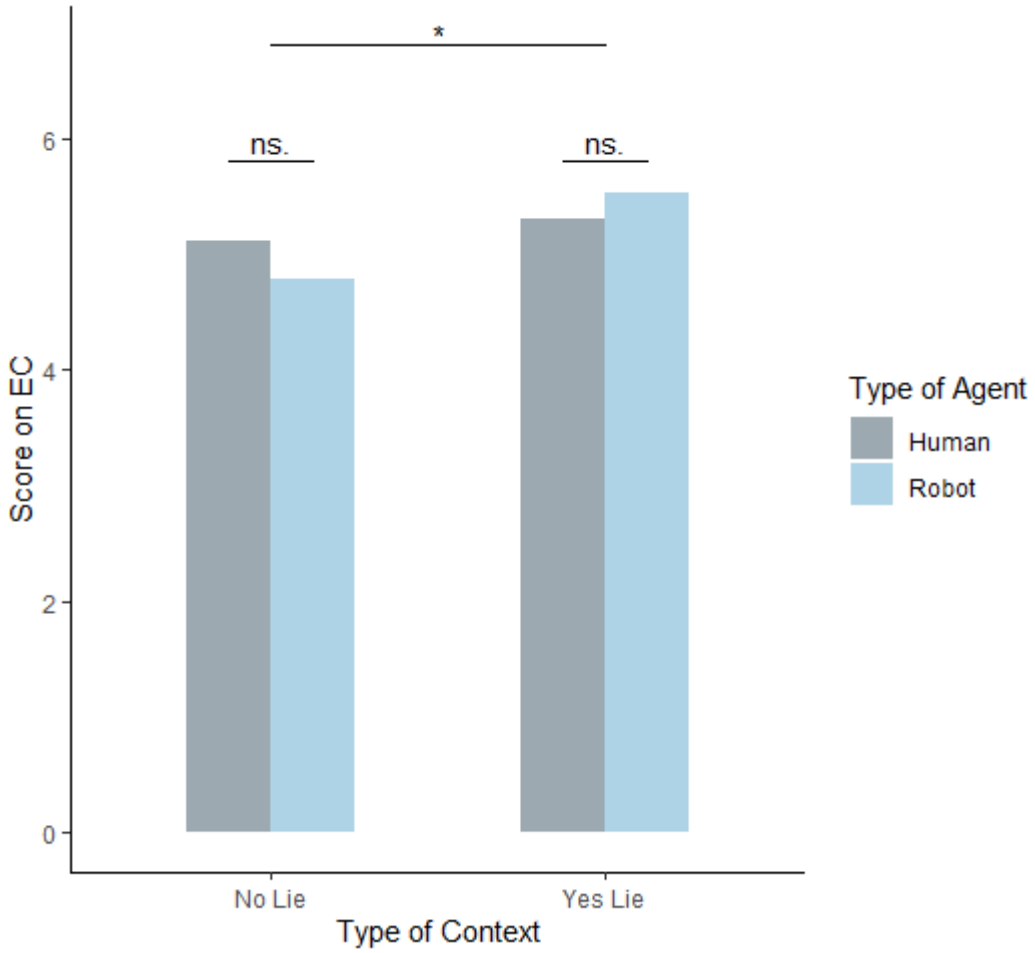
Fixed-Effects ANOVA results using scores on empathic concern as the criterion

Predictor	Sum of Squares	*df*	Mean Square	*F*	*p*	ω^2	ω^2 90% CI [LL, UL]
Agent	0.03	1	0.03	0.04	.846	-.01	[-.01, .00]
Context	4.45	1	4.45	5.13	.027*	.05	[-.01, .17]
Agent x Context	1.36	1	1.36	1.56	.216	.01	[-.01, .09]
Error	59.87	69	0.87				

Note: LL and UL represent the lower-limit and upper-limit of the omega squared (ω^2) confidence interval (CI), respectively.

Figure 8

Score on Empathic Concern Per Type of Agent and Type of Context



Note: abbreviation "ns." Meaning not significant and "*" stands for $p < 0.05$.

4.4.3.I-PANAS-SF

The mean scores on the PANAS were calculated separately for the positive affect (PA) and negative affect (NA) items. The internal reliability of the five items for PA was α = .84, and .72 for the five items on NA.

Residuals analysis on PA scores gave no significant results on the Shapiro-Wilk test nor on Levene's test (p = .289 and $F(3,69)$ = 1.218, p = .310, respectively) meeting the assumptions of normality and homogeneity of variance. For NA scores, the assumption for homogeneity of variance was met ($F(3,69)$ = 0.673, p = .571), but not for normality (p < .001). Further analysis showed no outliers for PA and one outlier for NA in the Robot Yes lie and one in the Robot No lie condition, however none were considered extreme outliers.

For PA scores, a two-way ANOVA showed no main effects for Agent (score ($F(1,69)$ = 0.209, p = .649) nor for Context (score ($F(1,69)$ = 0.002, p = .967), nor for the interaction effect of Agent and Context

($F(1,69) = 0.042$, $p = .839$). These findings suggest that the mean scores on PA do not differ significantly between the different conditions and no effects of Agent, Context or Agent and Context combined were found on PA.

For NA, it was decided to apply an Aligned Rank Transformation (Wobbrock, Findlater, Gergle, & Higgins, 2011) to account for the non-normally distributed data. Analysis on the transformed data showed no main effects of Agent or Context ($F(1,69) = 0.974$, $p = .327$ and $F(1,69) = 0.649$, $p = .423$, respectively), indicating no effect of Agent or Context on the mean scores for NA. The interaction effect of Agent and Context was also not significant ($F(1,69) = 0.016$, $p = .904$), showing no effect of Agent and Context on NA across the different conditions.

4.4.4. Gender of Opponent

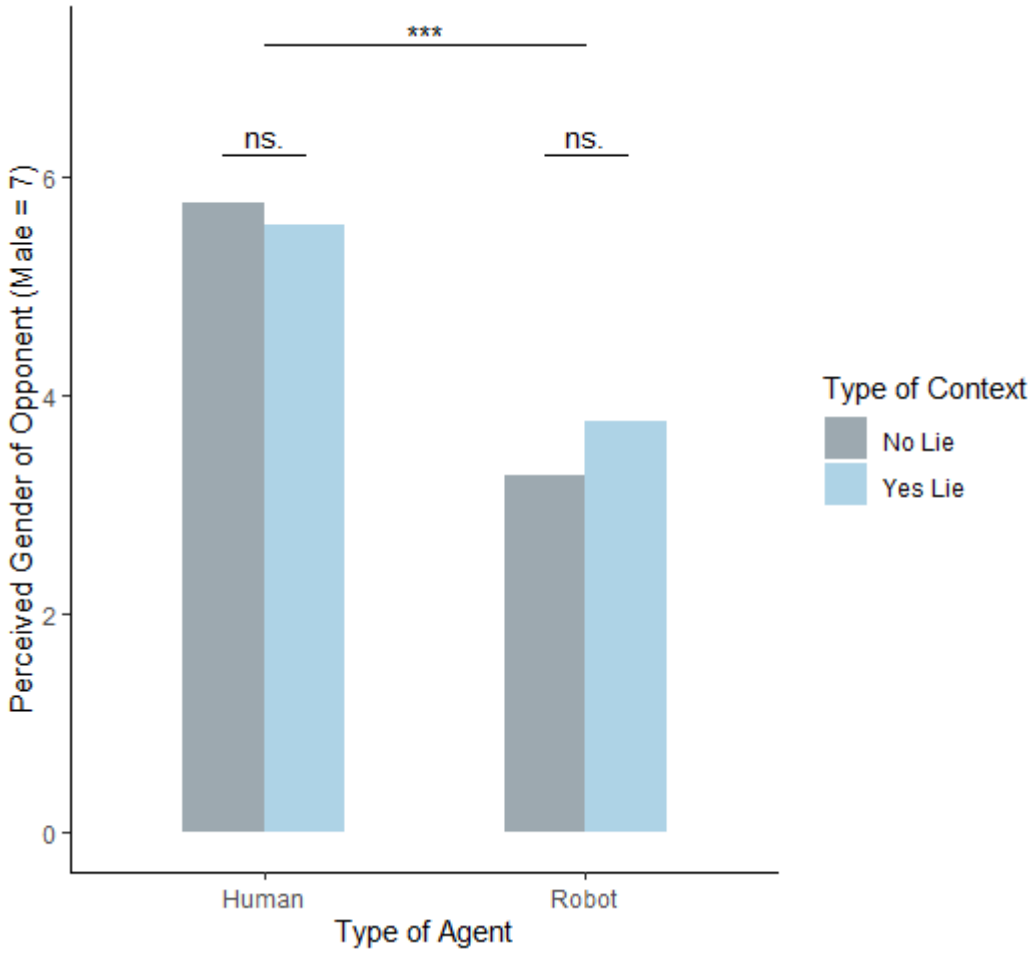
Lastly, it was decided to perform additional analysis on scores on perceived gender of the opponent to account for possible effects on the results. Levene's test was nonsignificant ($F(3,69) = 0.233$, $p = .873$), however the Shapiro Wilk's test did show a significant result ($p < 0.001$) and did therefore not meet the assumption of normally distributed residuals.

An Aligned Rank Transform for nonparametric factorial ANOVA (Wobbrock, Findlater, Gergle, & Higgins, 2011) was performed. Analysis on the data showed a significant main effect of Agent on the perceived gender of the opponent ($F(1,69) = 38.048$, $p < 0.001$, $\omega^2 = .31$; large effect) where the Human condition scored higher ($M = 5.67$, $SD = 1.80$) as opposed to the Robot condition ($M = 3.52$, $SD = 1.34$). See Figure 9. This suggests that the opponent in the Human condition was perceived as more male than the robot opponent.

The main effect of Context was nonsignificant ($F(1,69) = 0.014$, $p = .907$), suggesting no effect on Context on perceived gender. The interaction effect of Agent and Context was also nonsignificant ($F(1,69) = 1.952$, $p = .167$), indicating no effect of Agent and Context on perceived gender of the opponent across the different groups.

Figure 9

Degree of Perceived Gender of the Opponent per Agent Type



Note: Scores range from 0 (Female) to 7 (Male). The higher the score on gender the more male the opponent was perceived.

5. Discussion

The aim of the current study was to expand upon the topics of prosocial behavior, guilt and HRI literature by examining the relationship between guilt and the interaction with a robot or a human on prosocial behavior. Previous research on human prosocial behavior in HRI was found to be inconsistent, and on the one hand found humans to treat robots equally as they treated other people (e.g. Nitsch & Glassen, 2015), while on the other hand robots where found to be treated less fairly (e.g. Melo, Carnevale & Gratch, 2014). As being an important factor for engaging in prosocial behavior, differences in guilt were expected to produce the found differences in behavior. Therefore, in the current study, the effect of guilt on prosocial behavior and differences in expression between HRI and HHI was investigated.

Prior expectations of the study were not entirely met by the results. Firstly, it was expected that participants who interacted with a human agent would show more prosocial behavior compared to participants interacting with a robot agent (**H1**). However, no evidence was found to support this hypothesis; participants who interacted with another person did not engage in more prosocial behavior compared to participants interacting with a robot.

Secondly, no difference in amount of guilt was found between the HRI and HHI conditions, thereby rejecting (**H2**). However, participants in the Yes lie condition reported feeling more guilt compared to the No lie condition, suggesting that lying did increase the amount of guilt, but this was only for scores on the single guilt item. Additionally, participants in the Yes lie condition reported having more empathic concern (EC) compared to the No lie condition. Due to the inconsistencies in the results for the two different guilt measures no concrete conclusions can be drawn on the amount of guilt felt by the participants in the current study.

Since the results for the complete guilt questionnaire found no significant differences for type of Agent, thereby no support was found for (**H3**). Furthermore, no interaction effect was found, suggesting participants felt just as guilty when lying to a robot a lying to a human. Below, a more detailed explanation is sought to be found for the current results. Additionally, suggestions for future research are given based on the findings and limitations of the current study.

5.1 Prosocial Behavior

To assess the amount of prosocial behavior exhibited by the participants a modified version of the Prisoner's dilemma was used. The game was modified to shorten the learning curve to make it more accessible as the rules were easier to understand and to attempt to eliminate experience effects (e.g. Capraro & Cococcioni, 2015; Selten & Stoecker, 1986). Consequently, less practice trials were needed and only one participant was excluded from the results due to not having read the instructions. Additionally, the modified point system made gave clear insights in participants' behavior by giving

them two choices; (1) behaving in a more egocentric manner and winning the maximal amount of points where the opponent won zero or, (2) choosing to win less, but equal points for both themselves and their opponent. The latter was seen as an act of prosocial behavior and the amount of times this option was chosen, was seen as the total score of prosocial behavior.

An explanation for lack of differences between the prosocial behavior between HHI and HRI could be found in the literature that was touched upon at the beginning of this study; namely that findings on the topic of prosocial behavior in HRI and HHI are in fact inconsistent. On the one hand studies find that people treat robots and other humans differently (e.g. Melo, Carnevale & Gratch, 2014), whereas on the other hand research suggest that people do not engage in different behavior towards other people and robots (e.g. Cuijpers, 2013; Torta, van Dijk, Ruijten & Cuijpers, 2013). The experiments in these last studies, however, were wholly completed online and used static pictures of the opponents, which could have influenced the results and the validity of the findings. It could very well be that the findings of the present study fall in this last category of studies.

The current findings would imply that the possibility of mixed societies in the nearby future living side by side will maybe cause less of a divide between human and robot than maybe initially thought. However, multiple causes in the current study related to guilt could have affected the current results and will be explained below in more detail.

Due to circumstances it was not able to do a lab study, thereby possibly having influenced the effect of the experiment on prosocial behavior. For future research it would be interesting to test the validity of the modified version of the Prisoner's dilemma game as a tool to investigate the amount of prosocial behavior. Further research using different measures would likely give more insights in the differences in behavior that people may or may not exhibit towards robots and other humans, the emotional states influencing these behaviors and the possible consequences thereof.

5.2 Inconsistency in Guilt Measures

In the present study, amount of guilt was investigated using two different measures. Firstly, scores on the guilt questionnaire including the single item "guilty" that was added to the I-PANAS-SF questionnaire was analyzed. This questionnaire was based on the one used in De Hooge et al. (2012). Here no significant results were found, implying no effect of lying or the type of opponent on feelings of guilt. Secondly, sole scores on the single "guilty" item were investigated. In contrast, results on the single item showed participants in the Yes lie conditions to report feeling more guilt compared to participants in the No lie condition; there seems to be an effect of lying on guilt.

The difference in results is striking and can be interpreted in multiple ways. If the scores on the single guilt item give a correct representation of the guilt felt in the study, this would imply that people do feel

guilt after having lied. However, this does not translate into an increase in prosocial behavior, as no differences in amount of prosocial behavior were found between the conditions. This could insinuate that no relationship exists between feelings of guilt and prosocial behavior.

In the next section multiple explanations are sought and discussed for inconsistent results on the guilt measures. First the validity of both guilt measures is examined, then the effect of presence on the internal emotional state is discussed, and how a feeling of anonymity and the guilt threshold can explain the current findings. Additionally, the effect of facial expressions and the relationship of empathic concern and cognitive dissonance on guilt and prosocial behavior are discussed.

5.2.1 Validity of the Different Measures

First of all, a closer look at the different measures for guilt is necessary. To begin, exploratory analysis on the single "guilty" item showed multiple extreme outliers. Although there is controversy on the topic, it was decided to preserve the data as is to exclude researcher bias as no assignable cause could be found (for an interesting overview on the consequences of preservation and removal of outliers, please refer to Yang & Berdine (2016)).The significant result for guilt was lost however, when scores on the item from the guilt questionnaire based on De Hooge et al. (2012) were included in the analysis.

As mentioned before, guilt is a complex construct involving different elements as responsibility (Miceli and Castelfranchi, 2018) and an urge to repair the transgression (Rosenstock & O'Connor, 2018; Baumeister, Stillwell and Heatherton, 1994). These elements are also covered in the guilt questionnaire. However, this questionnaire was originally designed for experiments where participants were directly asked to remember a past situation which made them feel very guilty, give a short description of the event and then fill in the questionnaire. It could be that the use of the same questionnaire was not appropriate for the current context. The questions could have been too direct whereas a more subtle approach could undercover more subconscious feelings of guilt and generate more honest responses. Since subconscious emotions are also able to shape our behavior without us being aware of them and their effects, participants often do not report feeling them (Coan & Allen, 2007). If participants did indeed feel guilt, one would expect higher scores on the negative affect items of the I-PANAS-SF, and presumably lower scores for positive affect. Nevertheless, this was not the case, thus indicating that there might not have been a difference in amount of guilt felt in the different conditions.

Based on the reasons mentioned above, the present data is too inconsistent to make a clear interpretation on the presence of guilt in the current study. It could be that the complexity of guilt as a construct, might be too elaborate to capture with the use of the present questionnaires. Therefore, for future research, a different measure for guilt would be advised for a more subconscious inquiry of guilt, with perhaps the addition of physiological measures such as heart rate variability or skin conductance response.

5.2.2 The Effect of Presence on Emotions

Another possible explanation for the current findings in H2 involves the quality of the interactions during the experiments and its effect on one's emotional state. The amount of presence one feels during social interactions has been found to influence people's emotional state (Riva et al., 2007). Research using Virtual Reality (VR) has analyzed the relationship between presence and affective emotions. They found that this relationship was bidirectional; the affective state of the participant was correlated with the amount of presence felt in the environment, and when placed in an "emotional" environment, the feeling of presence rose compared to when being in a neutral environment (2007).

Similar results by Aymerich-Franch (2010) found a positive correlation between arousal and valence, and level of presence. Support for this finding can be found in the scores on the I-PANAS-SF questionnaire, which measures positive and negative affect related to valence and arousal. It was expected that participants would feel less (negative) emotions towards a robot. However, participants' scores on positive affect (PA) and negative affect (NA) did not differentiate for type of Agent (or Context) insinuating a lack of presence could have resulted in the current findings. This could indicate that participants in the current study did not feel "present" enough during the interactions to be influenced emotionally or behaviorally.

Interestingly, the amount of body participation in VR does not significantly affect the amount of feeling present or the emotional state (Aymerich-Franch, 2010). Evidence supporting these findings can be found in a study by Derks, Fischer and Bos (2008). In their study they reviewed the empirical evidence suggesting that communication mediated via a computer is less emotional and personal compared to face-to-face communication. The conclusions indicated that this was not the case; they found that both communication styles were very similar regarding the amount of emotional and personal communication (2008). This could indicate that alone the change to online interaction instead of the use of pre-scripted videos could increase the effect on emotions as this type of interaction would probably feel more natural.

For future research, it would be interesting to investigate the effect of real-time online interactions compared to the use of videos. Also, the addition of a measure for the amount of presence felt during interactions would be advised to monitor potential effects on the internal emotional state.

5.2.3 A Feeling of Anonymity and the Guilt Threshold

In the current study feelings of guilt were induced by asking participants to lie to their opponent during certain trails in the Guessing game. Participants were then confronted with disappointed feedback from their opponent, making the participant feel responsible and aware or their moral transgression. This has been found to lead to a will to compensate for the transgression in the form of helping and behaving prosocially (Miceli & Castelfranchi, 2018). However, there seems to be a certain threshold that needs to be reached first in order to make people feel guilty and the amount also seems to vary from person

to person (Torstveit, Sütterlin & Lugo, 2016). It could be that the current method to induce guilt was not potent enough to elevate the overall feeling of guilt in the participants.

Furthermore, the deindividuation theory of Zimbardo (1969) argues that if people do not know the identity of the other, they experience fewer moral emotions, which can lead to more extreme behavior. Research has shown that in bullying, cyberbullying experience less shame, remorse or guilt compared to bullying in the physical world (Slonje, Smith, & Frisén, 2012). In the current study participants interacted with their opponent without having to identify themselves, operating from the privacy of their own house. The greater feeling of anonymity could be an explanation for the lack of differences felt in the amount of guilt felt by the participants.

The above stated findings suggest that, together with the degree of anonymity that online experiments offer, the current method used to induce guilt was perhaps not potent enough to affect the internal emotional state of the participants as seen in the results. For additional research, a less anonymous form of interaction would be advised. The current method to induce guilt could then be tested and validated or compared to a potentially more potent method.

5.2.4 The Effect of Facial Expressions

Another critical factor for the lack of findings could come from the limited amount of facial expressions during the interactions as a result of the use of pre-scripted videos, especially during the negative feedback in the Guessing games. Indeed, facial expressions convey internal present emotional information from the individual to others making them crucial for accurate social communication (Blair, Morris, Frith, Perrett & Dolan, 1999). Additionally, negative self-conscious emotions such as shame and guilt are generated when viewing certain facial expressions in others. The recognition of the emotions in others helps one become aware of their own negative behavior (Beer et al., 2003). It could be that the use of pre-scripted videos diminished the amount and expression of (negative) emotions in the opponent, and thereby reducing the effect on the emotional state of the participant. A real-world setting using face-to-face interactions could perhaps cause more impact emotionally.

Evidence for the above explanation could be found in the results of the current study. Namely, no changes were found for positive and negative affect scores between the conditions. Together with the lack of guilt, these findings could indicate that participants were perhaps not able to recognize emotions in their opponents, and thus did not become aware of the consequences of their actions and feel guilty. If guilt is indeed a predictor of prosocial behavior, the lack of facial expressions thereby could clarify why no evidence for **H2** was found.

5.2.5 Empathic Concern and Cognitive Dissonance

Significant differences were found for empathic concern (EC) between the Yes lie and the No lie condition. Here participants reported having higher empathic concern when they had lied compared to

participants who did not lie during the experiment. Festinger's theory of cognitive dissonance (1957) states that when a discrepancy between the beliefs a person holds of themselves and their actions appears, this results in psychological stress i.e. cognitive dissonance. Affectively, this state causes people to feel discomfort and make them find a way to resolve the contradiction at hand.

In the current study, participants were asked to lie to their opponent (action) what could have gone against ideas they hold about themselves, for example "I am a good person" (belief). The action of lying goes against their belief of being a person, possibly causing cognitive dissonance. People who experience cognitive dissonance will try to resolve this conflict to reduce the mental discomfort (Festinger, 1957). After lying, participants were asked to fill out the IRI questionnaire consisting of descriptive sentences and asked to what degree the sentences applied to them. The questionnaire involved sentences as "I would describe myself as a pretty soft-hearted person" and "When I see someone taken advantage of, I feel kind of protective toward them" (see Appendix E for the complete questionnaire). Theses sentences could have been a representation of the beliefs the participants held of themselves that were directly before contradicted/violated by the act of lying. Therefore, it could very well be that participants reported themselves as being more empathic as an attempt to reduce the cognitive dissonance they were experiencing. This could explain the significant higher scores on empathic concern reported by the participants in the Yes lie condition. The nonsignificant scores on prosocial behavior could have been caused by the cognitive dissonance already being resolved by answering the IRI questionnaire, as prosocial behavior was measured directly afterwards.

Empathic concern involves the *feeling* similar emotions observed in the other person and is thought to motivate actions to relieve the distress observed in the other, whereas perspective taking is the *understanding* the other's internal state (Van der Graaff et al., 2018; Batson et al., 1989). As cognitive dissonance effects the internal emotional state of the participant, this could explain why, in the present study, only differences in empathic concern scores were found but not for perspective taking.

The findings on the IRI questionnaire could imply that participants that lied perhaps did feel a certain amount of guilt during one point in the experiment. However, the potential premature resolving of cognitive dissonance could have affected the results on prosocial behavior. For further research it would be advised to be mindful of these possible order effects. Additionally, a closer investigation on the role of guilt, cognitive dissonance and empathic concern in the expression of prosocial behavior would be insightful.

5.3 A Different Perspective; Robot and Human Agents

Earlier research has reported findings suggesting people see robots and even non-anthropomorphic robots as social agents after social interaction (Hoenen, Lübke & Pause, 2016; Bartneck Van Der Hoek, Mubin & Mahmud, 2007). Additionally, people have been found to be able to interpret information from a robot's perspective when, for example they were given images of a robot which had a number

displayed in front of it. This would either look like a "6" from the robot's perspective, or a "9" from the participant's own perspective. Participants did this equally from a robot's perspective as of a human's (Zhao, Cusimano & Malle, 2016). The results confirm current findings on perspective taking and could accord for the lack of differences found for the two types of agent. However, in Zhao, Cusimano and Malle's study, the difference in perspective taking scores became significant when the researchers used videos instead of still images in their experiment where people reported lower scores on perspective taking for the robot compared to the human agent (2016). This again would infer that the current use of videos did not reflect a natural interaction, as otherwise clearer differences in perspective taking would have been found.

Additional support for the influence of natural interactions can be found in the scores on the Godspeed questionnaire. Although a significant difference between the human and the robot condition was found on anthropomorphism, the score for the human condition was below average (M = 3.25 out of 7). One would expect this score to be much higher when interacting with another human-being. The same occurred for scores on animacy, although no significant difference was found, again the score for the human condition was below average (M = 3.30). These findings strongly suggest that the interactions with the agents did not feel like real interactions and could account for the lack of prosocial behavior and emotions found in the study.

These findings, together with the previous literature, further support the idea that humanoid robots are indeed treated and seen equally as other humans. However, more research would be needed to further account for the possible methodological flaws of the current study and thereby the lack of significant results found. For future research, it would be interesting to investigate the effect of live computer-mediated interaction between participant and opponent or to use real face-to-face interactions. Also, more emphasis on the role of empathy would be an insightful addition, together with an accurate measure for feeling of presence and guilt.

5.4 The Influence of Unequal Ratio of Gender

Another possible explanation for the current results could be explained by the unequal proportion of gender in the different conditions (Please refer to Table 1 in the Results for the complete overview). Especially in the robot condition where in the No lie condition nearly 70% was female whereas the reverse was the case in the Yes lie conditions. Men and women are known to experience emotions differently; women reportedly experience more emotions such as guilt and shame (Else-Quest, Higgins, Allison & Morton, 2012). The unbalanced gender ratio across the conditions could have resulted in the lack of differences found for feelings of guilt. The two-way analysis using Context and type of Agent as factors showed a high variability in the scores of guilt per Context, which could have caused a difference that did not reach significance levels. More balanced conditions and a focus on gender differences would be an interesting addition for future research.

5.5 Limitations

Some additional limitations of the current study should be discussed and taken into account when interpreting the current findings. First, due to technical challenges, it was not possible to use the Pepper robot's original voice. Instead, using the online text-to-speech converter Notevibes.com, the English US "Ivy(child)" voice was used as it accorded most with Pepper's original voice. Possible influences due to the use of a "human" voice cannot be ruled out and could explain why participants scored both the human and the robot agent as seeming equally animate. Future research should try to use the original robot voice as to preserve the validity of the experiment. The use of the human voice could have influenced the scores on likability, as these were also equal for both agents. Secondly, for the Goodspeed questionnaire, the item "apathetic" was not included in the animacy subscale, potentially having influenced scoring for animacy. Lastly, a large proportion of the sample consisted of university students which could have resulted in a skewed representation of the population.

6. Conclusion

This study sought to give insight into the relationship of guilt and prosocial behavior and to compare the differences found between human-robot interaction and human-human interaction. With this study an attempt was made to uncover potential factors influencing differences in behavior found in HRI and HHI. Future research could highlight the consequences of future mixed societies including (humanoid) robots and provide perspective on relevant subsequent policies.

Although the results of the current study are all but conclusive; the results partially imply that lying could potentially induce feelings of guilt, but these feelings did not translate into more engagement of prosocial behavior, debunking the effect of guilt. Also, no differences in guilt or prosocial behavior were found for type of agent. However, an effect of guilt on prosocial behavior cannot be excluded based on the findings of the present study as multiple influencing factors such as cognitive dissonance, low quality interactions and possible inaccurate measures for guilt could have resulted in an effect being undetectable in the current conditions.

Therefore, for future research, the use of more accurate measures for guilt including physiological tools would be advised. Also, live online interactions or face-to-face interactions should be considered for similar research. It would be interesting to compare the findings using different interaction methods to the current findings. Additionally, a focus on different variables such as presence, empathic concern and the amount of facial expressions would be interesting to consider in future research.

To conclude, this study contributed to new insights in the fields of human emotion and behavior, and HRI. Since the introduction of (humanoid) robots to society is slowly becoming a reality, it is important to investigate whether humans interact differently with robots than with other people in order to foresee and prevent potential problems that may arise from the introduction.

7 References

1. Abraham, A., Pocheptsova, A., & Ferraro, R. (2012). The effect of mobile phone use on prosocial behavior. *Manuscript in preparation*.

2. Anderson, C. A., & Bushman, B. J. (2002). The effects of media violence on society. *Science, 295*(5564), 2377-2379.

3. Andreoni, J., & Miller, J. H. (1993). Rational cooperation in the finitely repeated prisoner's dilemma: Experimental evidence. *Economic Journal, 103*(418), 570-85.

4. Anwyl-Irvine, A.L., Massonié J., Flitton, A., Kirkham, N.Z., Evershed, J.K. (2019). *Gorilla in our midst: an online behavioural experiment builder*. Behavior Research Methods. Doi: https://doi.org/10.3758/s13428-019-01237-x

5. Aymerich-Franch, L. (2010). Presence and emotions in playing a group game in a virtual environment: the influence of body participation. *Cyberpsychology, Behavior, and Social Networking, 13*(6), 649-654.

6. Bainbridge, W. A., Hart, J. W., Kim, E. S., & Scassellati, B. (2011). The benefits of interactions with physically present robots over video-displayed agents. *International Journal of Social Robotics, 3*(1), 41-52.

7. Bartneck, C., Croft, E., Kulic, D. & Zoghbi, S. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics, 1*(1) 71-81. | DOI: 10.1007/s12369-008-0001-3

8. Bartneck, C., Van Der Hoek, M., Mubin, O., & Al Mahmud, A. (2007, March). "Daisy, daisy, give me your answer do!" switching off a robot. In *2007 2nd ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 217-222). IEEE.

9. Bartneck, C., Verbunt, M., Mubin, O., & Al Mahmud, A. (2007, March). To kill a mockingbird robot. In *Proceedings of the ACM/IEEE international conference on Human-robot interaction* (pp. 81-87). ACM.

10. Basil, D. Z., Ridgway, N. M., & Basil, M. D. (2008). Guilt and giving: A process model of empathy and efficacy. *Psychology & Marketing, 25*(1), 1-23.

11. Batson, C. D. (2012). 12 A history of prosocial behavior research. *Handbook of the history of social psychology*, 243.

12. Batson, C. D. (1987). Prosocial motivation: Is it ever truly altruistic?. In *Advances in experimental social psychology* (Vol. 20, pp. 65-122). Academic Press.

13. Batson, C. D., Batson, J. G., Griffitt, C. A., Barrientos, S., Brandt, J. R., Sprengelmeyer, P., & Bayly, M. J. (1989). Negative-state relief and the empathy—altruism hypothesis. *Journal of Personality and Social Psychology*, *56*(6), 922.

14. Batson, C. D., & Powell, A. A. (2003). Altruism and prosocial behavior. *Handbook of psychology*, 463-484.

15. Baumeister & Bushman (2007). *Social Psychology and Human Nature*. Cengage Learning. p. 254. ISBN 9780495116332.

16. Baumeister, R. F., Stillwell, A. M., & Heatherton, T. F. (1994). Guilt: an interpersonal approach. *Psychological bulletin*, *115*(2), 243.

17. Baumsteiger, R., & Siegel, J. T. (2019). Measuring prosociality: The development of a prosocial behavioral intentions scale. *Journal of personality assessment*, *101*(3), 305-314.

18. Baxter, P., Ashurst, E., Kennedy, J., Senft, E., Lemaignan, S., & Belpaeme, T. (2015, February). The wider supportive role of social robots in the classroom for teachers. In *1st Int. Workshop on Educational Robotics at the Int. Conf. Social Robotics. Paris, France* (Vol. 6).

19. Beer, J. S., Heerey, E. A., Keltner, D., Scabini, D., & Knight, R. T. (2003). The regulatory function of self-conscious emotion: insights from patients with orbitofrontal damage. *Journal of personality and social psychology*, *85*(4), 594.

20. Benenson, J. F., Pascoe, J., & Radmore, N. (2007). Children's altruistic behavior in the dictator game. *Evolution and Human Behavior*, *28*(3), 168-175.

21. Beran, T. N., Ramirez-Serrano, A., Kuzyk, R., Nugent, S., & Fior, M. (2011). Would children help a robot in need?. *International Journal of Social Robotics*, *3*(1), 83-93.

22. Bergin, C., Talley, S., & Hamer, L. (2003). Prosocial behaviours of young adolescents: a focus group study. *Journal of adolescence*, *26*(1), 13-32.

23. Blair, R. J. R., Morris, J. S., Frith, C. D., Perrett, D. I., & Dolan, R. J. (1999). Dissociable neural responses to facial expressions of sadness and anger. *Brain*, *122*(5), 883-893.

24. Brennan, L., & Binney, W. (2010). Fear, guilt, and shame appeals in social marketing. *Journal of business Research*, *63*(2), 140-146.

25. Broekens, J., Heerink, M., & Rosendal, H. (2009). Assistive social robots in elderly care: a review. *Gerontechnology*, *8*(2), 94-103.

26. Brief, A., & Motowidlo, S. (1986). Prosocial Organizational Behaviors. *The Academy of Management Review*, *11*(4), 710-725. Retrieved from http://www.jstor.org/stable/258391

27. Brown, S. L., & Brown, R. M. (2015). Connecting prosocial behavior to improved physical health: Contributions from the neurobiology of parenting. *Neuroscience & Biobehavioral Reviews*, *55*, 1-17.

28. Brown, S. L., Nesse, R. M., Vinokur, A. D., & Smith, D. M. (2003). Providing social support may be more beneficial than receiving it: Results from a prospective study of mortality. *Psychological science*, *14*(4), 320-327.

29. Bekkers, R. H. (2007). Measuring altruistic behavior in surveys: The all-or-nothing dictator game.

30. Camerer, C. F. (2011). *Behavioral game theory: Experiments in strategic interaction*. Princeton University Press.

31. Capraro, V., & Cococcioni, G. (2015). Social setting, intuition and experience in laboratory experiments interact to shape cooperative decision-making. *Proceedings of the Royal Society B: Biological Sciences*, *282*(1811), 20150237.

32. Chai, T. Y., Woo, S. S., Rizon, M., & Tan, C. S. (2010). Classification of human emotions from EEG signals using statistical features and neural network. In *International* (Vol. 1, No. 3, pp. 1-6). Penerbit UTHM.

33. Cialdini, R. B., Baumann, D. J., & Kenrick, D. T. (1981). Insights from sadness: A three-step model of the development of altruism as hedonism. *Developmental review*, *1*(3), 207-223.

34. Coan, J. A., & Allen, J. J. (Eds.). (2007). *Handbook of emotion elicitation and assessment.* Oxford university press.

35. Cooper, R., DeJong, D. V., Forsythe, R., & Ross, T. W. (1996). Cooperation without reputation: Experimental evidence from prisoner's dilemma games. *Games and Economic Behavior*, *12*(2), 187-218.

36. Cuijpers, I. R. (2013). Investigating Rejection Behavior in the Ultimatum Game as a Measure of Anthropomorphism Els van Dijk June 2013.

37. Cunningham, M. R., Steinberg, J., & Grev, R. (1980). Wanting to and having to help: Separate motivations for positive mood and guilt-induced helping. *Journal of personality and social psychology*, *38*(2), 181.

38. Darley, J. M., & Batson, C. D. (1973). " From Jerusalem to Jericho": A study of situational and dispositional variables in helping behavior. *Journal of personality and social psychology*, *27*(1), 100.

39. Darley, J. M., & Latané, B. (1968). Bystander intervention in emergencies: Diffusion of responsibility. *Journal of personality and social psychology*, *8*(4p1), 377.

40. Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of personality and social psychology*, *44(*1), 113.

41. De Graaf, M. M., & Malle, B. F. (2019, March). People's Explanations of Robot Behavior Subtly Reveal Mental State Inferences. *In 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 239-248). IEEE.

42. De Hooge, I. E., Nelissen, R., Breugelmans, S. M., & Zeelenberg, M. (2011). What is moral about guilt? Acting "prosocially" at the disadvantage of others. *Journal of personality and social psychology*, *100*(3), 462.

43. De Kleijn, R., van Es, L., Kachergis, G., & Hommel, B. (2019). Anthropomorphization of artificial agents leads to fair and strategic, but not altruistic behavior. *International Journal of Human-Computer Studies*, *122*, 168-173.

44. De Melo, C. M., Carnevale, P., & Gratch, J. (2014). Social categorization and cooperation between humans and computers. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 36, No. 36).

45. De Melo, C. M., & Gratch, J. (2015, September). People show envy, not guilt, when making decisions with machines. *In 2015 International Conference on Affective Computing and Intelligent Interaction (ACII)* (pp. 315-321). IEEE.

46. Melo, C. D., Marsella, S., & Gratch, J. (2016). People do not feel guilty about exploiting machines. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *23*(2), 1-17.

47. De Visser, E. J., Krueger, F., McKnight, P., Scheid, S., Smith, M., Chalk, S., & Parasuraman, R. (2012, September). The world is not enough: Trust in cognitive agents. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 56, No. 1, pp. 263-267). Sage CA: Los Angeles, CA: Sage Publications.

48. Derks, D., Fischer, A. H., & Bos, A. E. (2008). The role of emotion in computer-mediated communication: A review. *Computers in human behavior*, *24*(3), 766-785.

49. Diehl, J. J., Schmitt, L. M., Villano, M., & Crowell, C. R. (2012). The clinical use of robots for individuals with autism spectrum disorders: A critical review. *Research in autism spectrum disorders*, *6*(1), 249-262.

50. Dovidio, J. F. (1984). Helping behavior and altruism: An empirical and conceptual overview. In *Advances in experimental social psychology* (Vol. 17, pp. 361-427). Academic Press.

51. Drummond, J. D., Hammond, S. I., Satlof-Bedrick, E., Waugh, W. E., & Brownell, C. A. (2017). Helping the one you hurt: Toddlers' rudimentary guilt, shame, and prosocial behavior after harming another. *Child development*, *88*(4), 1382-1397.

52. Dunn, E. W., Aknin, L. B., & Norton, M. I. (2008). Spending money on others promotes happiness. *Science*, *319*(5870), 1687-1688.

53. Eckel, C. C., & Grossman, P. J. (1996). Altruism in anonymous dictator games. *Games and economic behavior*, *16*(2), 181-191.

54. Eisenberg, N., Fabes, R. A. & Spinrad, T. L. (2006). Prosocial development. In N. Eisenberg (Ed.), *Handbook of child psychology: VoL 3. Social, emotional and personality development, 6th ed* (p. 645 -718).

55. Eisenberg, N., Spinrad, T. L., & Sadovsky, A. (2006). Empathy-related responding in children.

56. Eisenberg, N., & Miller, P. A. (1987). The relation of empathy to prosocial and related behaviors. *Psychological bulletin*, *101*(1), 91.

57. Ekman, P. (1992). Are there basic emotions?.

58. Elison, J. (2005). Shame and guilt: A hundred years of apples and oranges. *New Ideas in Psychology*, *23*(1), 5-32.

59. Else-Quest, N. M., Higgins, A., Allison, C., & Morton, L. C. (2012). Gender differences in self-conscious emotional experience: A meta-analysis. *Psychological bulletin*, *138*(5), 947.

60. Estrada-Hollenbeck, M., & Heatherton, T. F. (1998). Avoiding and alleviating guilt through prosocial behavior. In *Guilt and children* (pp. 215-231). Academic Press.

61. Etxebarría, I. (2000). Guilt: An emotion under suspicion. *Psicothema*, *12*(Su1), 101-108.

62. Fabes, R. A., Carlo, G., Kupanoff, K., & Laible, D. (1999). Early adolescence and prosocial/moral behavior I: The role of individual processes. *The Journal of Early Adolescence*, *19*(1), 5-16.

63. Fehr, E., & Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and human behavior*, *25*(2), 63-87.

64. Festinger, L. (1957). *A theory of cognitive dissonance* (Vol. 2). Stanford university press.

65. Field, A. (2013). *Discovering statistics using IBM SPSS statistics* (4th ed.). sage.

66. Forsythe, R., Horowitz, J. L., Savin, N. E., & Sefton, M. (1994). Fairness in simple bargaining experiments. *Games and Economic behavior*, *6*(3), 347-369.

67. Fox, J., & Weisberg, S. (2019). *An {R} Companion to Applied Regression, Third Edition*. Thousand Oaks CA: Sage. URL: https://socialsciences.mcmaster.ca/jfox/Books/Companion/

68. Gonsior, B., Buß, M., Sosnowski, S., Wollherr, D., Kühnlenz, K., & Buss, M. (2012, May). Towards transferability of theories on prosocial behavior from Social Psychology to HRI. In *2012 IEEE Workshop on Advanced Robotics and its Social Impacts (ARSO)* (pp. 101-103). IEEE.

69. Hoenen, M., Lübke, K. T., & Pause, B. M. (2016). Non-anthropomorphic robots as social entities on a neurophysiological level. *Computers in Human Behavior*, *57*, 182-186.

70. Hoffman, G., Forlizzi, J., Ayal, S., Steinfeld, A., Antanitis, J., Hochman, G., ... & Finkenaur, J. (2015, March). Robot presence and human honesty: Experimental evidence. *In Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction* (pp. 181-188). ACM.

71. Kagan, J. & Fox, N. A., (2006). Biology, Culture and Temperament. In N. Eisenberg (Ed.), *Handbook of child psychology: VoL 3. Social, emotional and personality development, 6th ed* (p. 645 -718).

72. Kassambara, A. (2020). *ggpubr: 'ggplot2' Based Publication Ready Plots*. R package version 0.3.0. https://CRAN.R-project.org/package=ggpubr

73. Kay M, Wobbrock J (2020). _ARTool: Aligned Rank Transform for Nonparametric Factorial ANOVAs_. doi: 10.5281/zenodo.594511 (URL: https://doi.org/10.5281/zenodo.594511), R package version 0.10.7, <URL: https://github.com/mjskay/ARTool>.

74. Kelley, H. H., & Stahelski, A. J. (1970). Social interaction basis of cooperators' and competitors' beliefs about others. *Journal of personality and social psychology*, *16*(1), 66.

75. Keltner, D. (1996). Evidence for the distinctness of embarrassment, shame, and guilt: A study of recalled antecedents and facial expressions of emotion. *Cognition & Emotion*, *10*(2), 155-172.

76. Kim, E. S., Berkovits, L. D., Bernier, E. P., Leyzberg, D., Shic, F., Paul, R., & Scassellati, B. (2013). Social robots as embedded reinforcers of social behavior in children with autism. *Journal of autism and developmental disorders*, *43*(5), 1038-1049.

77. King, T. J., Warren, I., & Palmer, D. (2008, January). Would Kitty Genovese have been murdered in Second Life? Researching the" bystander effect" using online technologies. In *TASA 2008: Re-imagining sociology: the annual conference of The Australian Sociological Association* (pp. 1-23). University of Melbourne.

78. Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and psychological measurement*, *56*(5), 746-759.

79. Knafo, A., (2016). *Prosocial behavior*. Encyclopedia of Early Childhood Development. Retrieved from http://www.child-encyclopedia.com/sites/default/files/dossiers-complets/en/prosocial-behaviour.pdf

80. Krebs, D. (1975). Empathy and altruism. *Journal of Personality and Social psychology*, *32*(6), 1134.

81. Kreps, D. M., Milgrom, P., Roberts, J., & Wilson, R. (1982). Rational cooperation in the finitely repeated prisoners' dilemma. *Journal of Economic theory*, *27*(2), 245-252.

82. Krishnakumar, S., & Rymph, D. (2012). Uncomfortable ethical decisions: The role of negative emotions and emotional intelligence in ethical decision-making. *Journal of Managerial Issues*, 321-344.

83. Li, J. (2015). The benefit of being physically present: A survey of experimental works comparing copresent robots, telepresent robots and virtual agents. *International Journal of Human-Computer Studies, 77*, 23-37.

84. Lim, S., & Reeves, B. (2010). Computer agents versus avatars: Responses to interactive game characters controlled by a computer or other player. *International Journal of Human-Computer Studies, 68*(1-2), 57-68.

85. Long, J. S., & Ervin, L. H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician, 54*(3), 217-224.

86. Malti, T., & Krettenauer, T. (2013). The relation of moral emotion attributions to prosocial and antisocial behavior: A meta-analysis. *Child development, 84*(2), 397-412.

87. Manning, R., Levine, M., & Collins, A. (2007). The Kitty Genovese murder and the social psychology of helping: The parable of the 38 witnesses. *American Psychologist, 62*(6), 555.

88. McDougall, W. (1908). The principal instincts and the primary emotions of man.

89. Meier, S., & Stutzer, A. (2008). Is volunteering rewarding in itself?. *Economica, 75*(297), 39-59.

90. Miceli, M., & Castelfranchi, C. (2018). Reconsidering the differences between shame and guilt. *Europe's journal of psychology, 14*(3), 710.

91. Millon, T. (2003). Evolution: A Generative Source for Conceptualizing the Attributes of Personality. In J. Millon, Lerner & Weiner (Eds.), *Handbook of Psychology: Volume 5, Personality and Social Psychology*.

92. Moll, J., Krueger, F., Zahn, R., Pardini, M., de Oliveira-Souza, R., & Grafman, J. (2006). Human fronto–mesolimbic networks guide decisions about charitable donation. *Proceedings of the National Academy of Sciences*, *103*(42), 15623-15628.

93. Nitsch, V., & Glassen, T. (2015, August). Investigating the effects of robot behavior and attitude towards technology on social human-robot interactions. In *2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)* (pp. 535-540). IEEE.

94. Nomura, T., Kanda, T., Kidokoro, H., Suehiro, Y., & Yamada, S. (2016). Why do children abuse robots?. *Interaction Studies*, *17*(3), 347-369.

95. O'Reilly, D., Connolly, S., Rosato, M., & Patterson, C. (2008). Is caring associated with an increased risk of mortality? A longitudinal study. *Social science & medicine*, *67*(8), 1282-1290.

96. Orne, M. T. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American psychologist*, *17*(11), 776.

97. Oyedele, A., Hong, S., & Minor, M. S. (2007). Contextual factors in the appearance of consumer robots: exploratory assessment of perceived anxiety toward humanlike consumer robots. *CyberPsychology & Behavior*, *10*(5), 624-632.

98. Padilla-Walker, L. M., & Carlo, G. (2014). The study of prosocial behavior. *Prosocial development: A multidimensional approach*, *3*.

99. Paeng, E., Wu, J., & Boerkoel, J. C. (2016, March). Human-robot trust and cooperation through a game theoretic framework. In *Thirtieth AAAI Conference on Artificial Intelligence*.

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

101. Poulin, M. J., & Holman, E. A. (2013). Helping hands, healthy body? Oxytocin receptor gene and prosocial behavior interact to buffer the association between stress and physical health. *Hormones and behavior*, *63*(3), 510-517.

102. Rauterberg, M. (2004). Positive effects of entertainment technology on human behaviour. In *Building the information society* (pp. 51-58). Springer, Boston, MA.

103. Ravaja, N. (2009). The psychophysiology of digital gaming: The effect of a non co-located opponent. *Media Psychology*, *12*(3), 268-294.

104. Rebega, O. L., Apostol, L., Benga, O., & Miclea, M. (2013). Inducing guilt: A literature review. *Procedia-Social and Behavioral Sciences*, *78*, 536-540.

105. Regan, D. T., Williams, M., & Sparling, S. (1972). Voluntary expiation of guilt: A field experiment. *Journal of Personality and Social Psychology*, *24*(1), 42.

106. Riva, G., Mantovani, F., Capideville, C. S., Preziosa, A., Morganti, F., Villani, D., ... & Alcañiz, M. (2007). Affective interactions using virtual reality: the link between presence and emotions. *CyberPsychology & Behavior*, *10*(1), 45-56.

107. Rodrigues, J., Ulrich, N., Mussel, P., Carlo, G., & Hewig, J. (2017). Measuring prosocial tendencies in Germany: sources of validity and reliablity of the revised prosocial tendency measure. *Frontiers in psychology*, *8*, 2119.

108. Rosenstock, S., & O'Connor, C. (2018). When It's Good to Feel Bad: An Evolutionary Model of Guilt and Apology. *Frontiers in Robotics and AI*, *5*, 9.

109. Ruvinsky, A., & Huhns, M. N. (2008, May). Simulating human behaviors in agent societies. In *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems-Volume 3* (pp. 1513-1516). International Foundation for Autonomous Agents and Multiagent Systems.

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

111. Scott, J. (2000). Rational choice theory. *Understanding contemporary society: Theories of the present*, *129*, 671-85.

112. Selten, R., & Stoecker, R. (1983). End behavior in sequences of finite prisoner's dilemma supergames.

113. Shaw, R. G., & Mitchell-Olds, T. (1993). ANOVA for unbalanced data: an overview. *Ecology, 74*(6), 1638-1645.

114. Shiomi, M., Nakata, A., Kanbara, M., & Hagita, N. (2017). A hug from a robot encourages prosocial behavior. In *2017 26th IEEE international symposium on robot and human interactive communication (RO-MAN)* (pp. 418-423). IEEE.

115. Singer, T., & Fehr, E. (2005). The neuroeconomics of mind reading and empathy. *American Economic Review, 95*(2), 340-345.

116. Slonje, R., Smith, P. K., & Frisén, A. (2012). Processes of cyberbullying, and feelings of remorse by bullies: A pilot study. *European Journal of Developmental Psychology*, *9*(2), 244-259.

117. Takeuchi, S., Imahori, T. T., & Matsumoto, D. (2001). Adjustment of criticism styles in Japanese returnees to Japan. I*nternational Journal of Intercultural Relations, 25*(3), 315-327.

118. Tangney, J. P., Miller, R. S., Flicker, L., & Barlow, D. H. (1996). Are shame, guilt, and embarrassment distinct emotions?. *Journal of personality and social psychology, 70*(6), 1256.

119. Team, R. C. (2013). R: A language and environment for statistical computing.

120. Thompson, E. R. (2007). Development and validation of an internationally reliable short-form of the positive and negative affect schedule (PANAS). *Journal of cross-cultural psychology, 38*(2), 227-242.

121. Tian, L., Du, M., & Huebner, E. S. (2015). The effect of gratitude on elementary school students' subjective well-being in schools: The mediating role of prosocial behavior. *Social Indicators Research, 122*(3), 887-904.

122. Tilghman-Osborne, C., Cole, D. A., & Felton, J. W. (2010). Definition and measurement of guilt: Implications for clinical research and practice. *Clinical Psychology Review*, *30*(5), 536-546.

123. Torstveit, L., Sütterlin, S., & Lugo, R. G. (2016). Empathy, guilt proneness, and gender: Relative contributions to prosocial behaviour. *Europe's Journal of Psychology*, *12(*2), 260.

124. Torta, E., van Dijk, E., Ruijten, P. A., & Cuijpers, R. H. (2013, October). The ultimatum game as measurement tool for anthropomorphism in human–robot interaction. In *International Conference on Social Robotics* (pp. 209-217). Springer, Cham.

125. Trubisky, P., Ting-Toomey, S., & Lin, S. L. (1991). The influence of individualism-collectivism and self-monitoring on conflict styles. *International journal of intercultural relations*, *15*(1), 65-84.

126. Van der Graaff, J., Carlo, G., Crocetti, E., Koot, H. M., & Branje, S. (2018). Prosocial behavior in adolescence: gender differences in development and links with empathy. *Journal of youth and adolescence*, *47*(5), 1086-1099.

127. Van der Mark, I. L., Van IJzendoorn, M. H., & Bakermans-Kranenburg, M. J. (2002). Development of empathy in girls during the second year of life: Associations with parenting, attachment, and temperament. *Social development*, *11*(4), 451-468.

128. Van Rompay, T. J., Vonk, D. J., & Fransen, M. L. (2009). The eye of the camera: Effects of security cameras on prosocial behavior. *Environment and Behavior*, *41*(1), 60-74.

129. Weinstein, N., & Ryan, R. M. (2010). When helping helps: Autonomous motivation for prosocial behavior and its influence on well-being for the helper and recipient. *Journal of personality and social psychology*, *98*(2), 222.

130. Wickham, H., François, R., Henry, L., & Müller, K. (2020). *dplyr: A Grammar of Data Manipulation.* R package version 0.8.5. https://CRAN.R-project.org/package=dplyr

131. Willer, D., & Walker, H. A. (2007). *Building experiments: Testing social theory*. Stanford University Press.

132. Wobbrock, J. O., Findlater, L., Gergle, D., & Higgins, J. J. (2011, May). The aligned rank transform for nonparametric factorial analyses using only anova procedures. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 143-146).

133. Yang, S., & Berdine, G. (2016). Outliers. *The Southwest Respiratory and Critical Care Chronicles, 4*(13), 52-56.

134. Zaatari, D., & Trivers, R. (2007). Fluctuating asymmetry and behavior in the ultimatum game in Jamaica. *Evolution and Human Behavior*, *28*(4), 223-227.

135. Zhao, X., Cusimano, C., & Malle, B. F. (2016, March). Do people spontaneously take a robot's visual perspective?. *In 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 335-342). IEEE.

136. Zimbardo, P. G. (1969). The human choice: Individuation, reason, and order versus deindividuation, impulse, and chaos. *In Nebraska symposium on motivation.* University of Nebraska press.

Appendix A

Script Yes Lie Condition

Introduction

Hi, I am **Hugo/Pepper** and I will be your opponent for today.

Practice round

1. Practice

A: Question 1: Is your character purple?

P: Yes/No.

A: Question 2: Okay. Is your character orange?

P: Yes/No.

A: Question 3: Does your character have 2 eyes?

P: Yes/No.

A: Is your character Albert?

P: Yes/No.

2. Practice

A: Question 1: Is your character yellow?

P: Yes/No.

A: Question 2: Is your character blue?

P: Yes/No.

A: Question 3: Does your character have more than 2 eyes?

P: Yes/No.

A: Is your character **Boris?**

P: Yes/No.

Start Experiment

3. Trial

A: Question 1: Is your character yellow?

P: Yes/No.

A: Question 2: Is *your* character blue?

P: Yes/No.

A: Question 3: Does your character have ears?

P: Yes/No.

A: Is your character Boris?

P: Yes/No.

*A: That's great / that's a shame

4. Trial

A: Question 1: Is your character purple?

P: Yes/No.

A: Question 2: Does your character have more than 2 eyes?

P: Yes/No.

A: Question 3: Does your character have a hat?

P: Yes/No.

A: Is your character Celia?

P: Yes/No.

*A: That's disappointing / That's awesome

5. Trial

A: Question 1: Is your character green?

P: Yes/No.

A: Question 2: Does your character have more than 2 eyes?

P: Yes/No.

A: Question 3: Does your character have 4 eyes?

P: Yes/No.

A: Is your character Richard?

P: Yes/No.

*A: That’s a shame / That’s great

6. Trial

A: Question 1: Is your character yellow?

P: Yes/No.

A: Question 2: Okay. Is your character orange?

P: Yes/No.

A: Question 3: Does your character have 2 eyes?

P: Yes/No.

A: Is your character Kwille?

P: Yes/No.

*A: That’s disappointing / That’s awesome

7. Trial

A: Question 1: Is your character purple?

P: Yes/No.

A: Question 2: Does your character have more than 2 eyes?

P: Yes/No.

A: Question 3: Does your character have ears?

P: Yes/No.

A: Is your character Jules?

P: Yes/No.

*A: That’s a shame / That’s great

8. Trial

A: Question 1: Is your character green?

P: Yes/No.

A: Question 2: Is your character blue?

P: Yes/No.

A: Question 3: Does your character have ears?

P: Yes/No.

Is your character Molly?

P: Yes/No.

*A: That’s disappointing / That’s awesome

9. Trial

A: Question 1: Is your character yellow?

P: Yes/No.

A: Question 2: Does your character have two eyes?

P: Yes/No.

A: Question 3: Does your character have 2 horns on its head?

P: Yes/No.

A: Is your character Peter?

P: Yes/No.

*A: That’s a shame / That’s great

10. Trial

A: Question 1: Okay. Is your character orange?

P: Yes/No.

A: Question 2: Is your character green?

P: Yes/No.

A: Question 3: Does your character have more than 2 eyes?

P: Yes/No.

A: Is your character Opus?

P: Yes/No.

*A: That’s disappointing / That’s awesome

11. Trial

A: Question 1: Is your character blue?

P: Yes/No.

A: Question 2: Is your character purple?

P: Yes/No.

A: Question 3: Does your character have more than 2 eyes?

P: Yes/No.

A: Is your character Umberto?

P: Yes/No.

*A: That’s a shame / That’s great

12. Trial

A: Question 1: Is your character yellow?

P: Yes/No.

A: Question 2: Okay. Is your character orange?

P: Yes/No.

A: Question 3: Does your character have 2 eyes?

P: Yes/No.

A: Is your character Albert?

P: Yes/No.

*A: That’s disappointing / That’s awesome

* Depending if the participants answered no or yes for the third question, the first or the second answer option was given respectively.

Appendix B

Script No Lie Condition

Introduction

Hi, I am **Hugo/Pepper** and I will be your opponent for today.

Practice round

1. Practice

A: Question 1: Is your character purple?

P: Yes/No.

A: Question 2: Okay. Is your character orange?

P: Yes/No.

A: Question 3: Does your character have 2 eyes?

P: Yes/No.

A: Is your character Albert?

P: Yes/No.

2. Practice

A: Question 1: Is your character yellow?

P: Yes/No.

A: Question 2: Is your character blue?

P: Yes/No.

A: Question 3: Does your character have more than 2 eyes?

P: Yes/No.

A: Is your character Boris?

P: Yes/No.

Start Experiment

3. Trial

A: Question 1: Is your character yellow?

P: Yes/No.

A: Question 2: Is your character blue?

P: Yes/No.

A: Question 3: Does your character have ears?

P: Yes/No.

A: Is your character Boris?

P: Yes/No.

*A: That’s alright / Okay, next.

4. Trial

A: Question 1: Is your character purple?

P: Yes/No.

A: Question 2: Does your character have more than 2 eyes?

P: Yes/No.

A: Question 3: Does your character have a hat?

P: Yes/No.

A: Is your character Celia?

P: Yes/No.

*A: That’s okay / Alright.

5. Trial

A: Question 1: Is your character green?

P: Yes/No.

A: Question 2: Does your character have more than 2 eyes?

P: Yes/No.

A: Question 3: Does your character have 4 eyes?

P: Yes/No.

A: Is your character Gina?

P: Yes/No.

*A: That’s alright / Okay, next.

6. Trial

A: Question 1: Is your character yellow?

P: Yes/No.

A: Question 2: Okay. Is your character orange?

P: Yes/No.

A: Question 3: Does your character have 2 eyes?

P: Yes/No.

A: Is your character Albert?

P: Yes/No.

*A: That’s okay / Alright.

7. Trial

A: Question 1: Is your character purple?

P: Yes/No.

A: Question 2: Does your character have more than 2 eyes?

P: Yes/No.

A: Question 3: Does your character have ears?

P: Yes/No.

A: Is your character Jules?

P: Yes/No.

*A: That’s alright / Okay, next.

8. Trial

A: Question 1: Is your character green?

P: Yes/No.

A: Question 2: Is your character blue?

P: Yes/No.

A: Question 3: Does your character have ears?

P: Yes/No.

Is your character Walter?

P: Yes/No.

*A: That’s okay / Alright.

9. Trial

A: Question 1: Is your character yellow?

P: Yes/No.

A: Question 2: Does your character have two eyes?

P: Yes/No.

A: Question 3: Does your character have 2 horns on its head?

P: Yes/No.

A: Is your character Peter?

P: Yes/No.

*A: That’s alright / Okay, next.

10. Trial

A: Question 1: Okay. Is your character orange?

P: Yes/No.

A: Question 2: Is your character green?

P: Yes/No.

A: Question 3: Does your character have more than 2 eyes?

P: Yes/No.

A: Is your character Opus?

P: Yes/No.

*A: That’s okay / Alright.

11. Trial

A: Question 1: Is your character blue?

P: Yes/No.

A: Question 2: Is your character purple?

P: Yes/No.

A: Question 3: Does your character have more than 2 eyes?

P: Yes/No.

A: Is your character Tilly?

P: Yes/No.

*A: That’s alright / Okay, next.

12. Trial

A: Question 1: Is your character yellow?

P: Yes/No.

A: Question 2: Okay. Is your character orange?

P: Yes/No.

A: Question 3: Does your character have 2 eyes?

P: Yes/No.

A: Is your character Petrovo?

P: Yes/No.

*A: That’s okay / Alright.

* Depending if the participants answered no or yes for the third question, the first or the second answer option was given respectively.

Appendix C

Questionnaire 1

This scale consists of a number of words that describe different feelings and emotions. Read each word and then mark the appropriate answer on the scale below that word. **Indicate to what extent you feel this way right now, that is, at the *present* moment.**

Read each item carefully before responding. Answer as honestly as you can. There are no wrong answers.

	Not at all						**Very strongly**
Upset	1	2	3	4	5	6	7
Hostile	1	2	3	4	5	6	7
Alert	1	2	3	4	5	6	7
Ashamed	1	2	3	4	5	6	7
Inspired	1	2	3	4	5	6	7
Nervous	1	2	3	4	5	6	7
Determined	1	2	3	4	5	6	7
Attentive	1	2	3	4	5	6	7
Afraid	1	2	3	4	5	6	7

Active	1	2	3	4	5	6	7
Guilty	1	2	3	4	5	6	7

During the game, to what extent did you…

feel responsible for what happened?	**Not at all**						**Very strongly**
feel alone?	1	2	3	4	5	6	7
feel what you have done was wrong?	1	2	3	4	5	6	7
feel that all attention was drawn towards you?	1	2	3	4	5	6	7
think about what you had done to other people?	1	2	3	4	5	6	7
not want others to know?	1	2	3	4	5	6	7
want to repair what had happened?	1	2	3	4	5	6	7
worry about what others would think?	1	2	3	4	5	6	7
want to be forgiven?	1	2	3	4	5	6	7

Appendix D

Questionnaire 2

Please rate your impressions of your opponent on the scales presented below. Read each item carefully before responding. Answer as honestly as you can. There are no wrong answers, we are simply interested in your opinion.

Fake	**1**	**2**	**3**	**4**	**5**	**6**	**7**	Natural
Machinelike	**1**	**2**	**3**	**4**	**5**	**6**	**7**	Humanlike
Unconscious	**1**	**2**	**3**	**4**	**5**	**6**	**7**	Conscious
Artificial	**1**	**2**	**3**	**4**	**5**	**6**	**7**	Lifelike
Moving rigidly	**1**	**2**	**3**	**4**	**5**	**6**	**7**	Moving elegantly
Dead	**1**	**2**	**3**	**4**	**5**	**6**	**7**	Alive
Stagnant	**1**	**2**	**3**	**4**	**5**	**6**	**7**	Lively
Mechanical	**1**	**2**	**3**	**4**	**5**	**6**	**7**	Organic
Inert	**1**	**2**	**3**	**4**	**5**	**6**	**7**	Interactive
Dislike	**1**	**2**	**3**	**4**	**5**	**6**	**7**	Like
Unfriendly	**1**	**2**	**3**	**4**	**5**	**6**	**7**	Friendly
Unkind	**1**	**2**	**3**	**4**	**5**	**6**	**7**	Kind
Unpleasant	**1**	**2**	**3**	**4**	**5**	**6**	**7**	Pleasant
Aweful	**1**	**2**	**3**	**4**	**5**	**6**	**7**	Nice

Appendix E

Questionnaire 3

The following statements inquire about your thoughts and feelings in a variety of situations. **Indicate how well each statement describes you *in general* on the scales presented below.**
Read each item carefully before responding. Answer as honestly as you can. There are no wrong answers.

I often have tender, concerned feelings for people less fortunate than me

Does not describe me well	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**Describes me very well**

I sometimes find it difficult to see things from the "other guy's" point of view

Does not describe me well	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**Describes me very well**

Sometimes I don't feel very sorry for other people when they are having problems

Does not describe me well	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**Describes me very well**

I try to look at everybody's side of a disagreement before I make a decision

Does not describe me well	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**Describes me very well**

When I see someone being taken advantage of, I feel kind of protective towards them

Does not describe me well	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**Describes me very well**

I sometimes try to understand my friends better by imagining how things look from their perspective

Does not describe me well	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**Describes me very well**

Other people's misfortunes do not usually disturb me a great deal

Does not describe me well	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**Describes me very well**

If I'm sure I'm right about something, I don't waste much time listening to other people's arguments

Does not describe me well	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**Describes me very well**

When I see someone being treated unfairly, I sometimes don't feel very much pity for them

Does not describe me well	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**Describes me very well**

I am often quite touched by things that I see happen

Does not describe me well	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**Describes me very well**

I believe that there are two sides to every question and try to look at them both

Does not describe me well	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**Describes me very well**

I would describe myself as a pretty soft-hearted person

Does not describe me well	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**Describes me very well**

When I'm upset at someone, I usually try to "put myself in his shoes" for a while

Does not describe me well	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**Describes me very well**

Before criticizing somebody, I try to imagine how I would feel if I were in their place

Does not describe me well	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**Describes me very well**

Appendix F

Some last demographics

These are the last couple of questions and then we are done!

Age:

Gender:

- Female
- Male
- Other

Is English your native language?

- Yes
- No. Please specify how many years you have been speaking English (estimate if not sure)

What is the highest level of education you have completed?

- No schooling completed
- High school degree (VMBO/HAVO/WVO)
- MBO degree
- Bachelor's degree
- Master's degree
- Doctorate
- Other (please specify)

If you are a student, what is the direction of your current study?

- Social Sciences
- Law
- Geosciences
- Medicine

- Humanities
- Economics
- Veterinary Medicine
- Governance
- (Beta)Science
- Art/design
- Other (please specify)

Please indicate your level of experience with robots. How many times in the last 12 months have you encountered a humanoid robot?

- Daily
- Weekly
- Monthly
- Couple of times
- Only once
- Not at all

Please, rate your view of your opponent on the scale below:

Female 1 2 3 4 5 6 7 Male

Did you understand the rules of the first game (the Guessing Game with the different characters)?

- Yes
- No, namely ...

Did you understand the rules of the second game (the Card Game with the 1's & 0's)?

- Yes
- No, namely ...

Did you encounter problems with the video's/audio recordings at any time during the experiment?

- No
- Yes
- A little