International Journal of Human-Computer Studies 122 (2019) 168–173

Contents lists available at ScienceDirect

International Journal of Human-Computer Studies

journal homepage: www.elsevier.com/locate/ijhcs

Anthropomorphization of artificial agents leads to fair and strategic, but not altruistic behavior

Roy de Kleijn*,[a], Lisa van Es[a], George Kachergis[b], Bernhard Hommel[a]

[a] Cognitive Psychology Unit, Leiden University, The Netherlands
[b] Department of Psychology, Stanford University

ARTICLE INFO

Keywords:
Strategic behavior
Fairness preference
Human-robot interaction
Anthropomorphism

ABSTRACT

With robots playing an increasing role in our daily lives, our emotional responses to them have become an active subject of study. The process of anthropomorphization, ascribing human affordances to non-human objects, is thought to play a large role in human-robot interaction. However, earlier studies have relied largely on experimenter's manipulation of anthropomorphism, and the use of virtual robots. The aim of this study was to investigate peopleâs fairness preference and strategic and altruistic behavior toward different opponents (a human, a semi-humanoid and a spider-like robot, and a laptop) in two economic games. Anthropomorphization questionnaires and mood measures were also administered. Our findings suggest that fairness preference and strategic behavior are not predicted by the opponent's physical appearance, but instead predicted by individual differences in the tendency to anthropomorphize others. Altruistic behavior, on the other hand, is affected by the opponent's physical appearance.

1. Introduction

In a society where humans and robots increasingly interact with each other, it will become important for robots to be accepted by humans as members of a shared society. Humans have been shaped by both natural selection and cultural tradition to cooperate and interact with other humans and animals. In fact, some theories even state that human intelligence is *the evolutionary result* of complex social interactions, requiring the ability to predict the behavior of others (Dautenhahn, 1998; Dunbar, 1998). In contrast, robots are a relatively novel addition to our social environment and, as such, we are not shaped by evolution to accurately predict their behavior. Without such predictive ability, we believe it is important to investigate what would be critical design features to build up trust and collaboration between humans and robots. Increasing our insight into such features is likely to ease our transition to a more automated society.

The physical design of modern robots is rather heterogeneous. At the moment, consumer robots take the form of vacuum cleaners (e.g., Roomba) or self-driving vehicles, but as we expect robots to perform more and more everyday human action, we could expect robots to increasingly look like us. Many research robots that are used to study everyday action (e.g., Willow Garage's PR2) are equipped with two arms and a binocular camera system arranged similarly to the anatomy of humans (albeit with more wheels and fewer legs). This physical similarity to humans has sparked the interest of human-robot interaction researchers, and some have suggested the existence of a non-linear relationship between physical similarity to human appearance and likeability of robots, nicknamed the *uncanny valley* (Mori, 2012; Pollick, 2010). The uncanny valley theory states that there is a positive relationship between the likeability of a robot and its human-likeness. However, at very high levels of human-likeness there is a sharp decrease in likeability, which is coined the uncanny valley.

1.1. Anthropomorphization

Another consequence of a robot's physical similarity to humans is an increase in potential *anthropomorphization:* the tendency to attribute human characteristics to non-human agents or even objects, such as animals or computers (Bartneck et al., 2009). It has been proposed (e.g. Epley et al., 2007) that the extent to which we anthropomorphize an agent is dependent on its physical similarity due to the inaccessibility of the phenomenological experience of others. While we are unable to imagine what it would be like to be, let's say, a bat (Nagel, 1974), it is easier for us to imagine what it would be like to be another *person*, and humanoid robots would potentially fall somewhere in between.

There has been an increasing interest in studying anthropomorphism, both on a psychological and a neuroscientific level. The perhaps most comprehensive psychological framework for

* Corresponding author.
E-mail address: kleijnrde@fsw.leidenuniv.nl (R. de Kleijn).

https://doi.org/10.1016/j.ijhcs.2018.09.008
Received 4 February 2017; Received in revised form 11 September 2018; Accepted 20 September 2018
Available online 21 September 2018
1071-5819/

anthropomorphism, described by Epley et al. (2007), predicts that people tend to anthropomorphize agents when they are motivated to be effective social agents, when they lack a connection to other humans, and when anthropocentric knowledge is accessible and applicable.

Other studies have instead argued for the importance of goal-directed, meaningful action. In as early as 1944, Heider and Simmel (1944) showed that people tend to attribute human states such as intent to elementary geometric shapes moving in a seemingly meaningful way. Similarly, neuroscientific studies on anthropomorphism have shown that the human mirror neuron system responds to observed actions performed even by industrial robots (Gazzola et al., 2007), especially when the action seems to be goal-directed. When anthropomorphizing, the superior temporal sulcus, which is also involved in dispositional attribution to people, and amygdala, which is involved in social categorization, seem to play an important role (Harris and Fiske, 2008). And indeed, amygdala-damaged patients, as well as patients with autism seem to exhibit impaired anthropomorphization (Castelli et al., 2002; Heberlein and Adolphs, 2004).

From a psychological viewpoint, it is interesting to investigate the social consequences of such anthropomorphization. Robots that look like humans, or for another reason are attributed with human-like characteristics, might be expected to elicit a higher empathic response than non-humanoid robots. However, as the uncanny valley may predict, this anthropomorphization could also lead to negative emotional responses, depending on the similarity to humans.

1.2. Altruistic and strategic behavior toward robots

Riek et al. (2009) have investigated the influence of anthropomorphization on empathic behavior. Subjects were presented with a film clip featuring one of five protagonists, ranging in physical appearance from a Roomba to a human. Film clips were either neutral or emotionally evocative in which the protagonist was being treated cruelly. After the film clip, subjects were asked which one of the four robots they would save in the event of an earthquake. More human-like protagonists induced higher empathy in subjects, feeling more sorry for them and reporting taking higher risks to save them.

Another paradigm in the study of empathy uses economic games to measure altruistic and strategic behavior. Underlying this design is the idea that altruistic behavior is necessarily preceded by empathic concern for others, known as the *empathy-altruism hypothesis* (Batson, 1991; Cialdini et al., 1997). In such economic games, some amount of money is given to a human participant, the proposer, who is subsequently asked to offer a stake of this amount to another player, the receiver. The dictator and ultimatum games are among the most widely-used economic games in the social sciences (Andersen et al., 2011; Engel, 2011; Güth et al., 1982). The premise of these games is that the amount of money given away by the proposer is an indicator of altruistic or strategic behavior, and reflects a preference for fairness.

In the ultimatum game, the proportion of the stake offered by the proposer is thought to reflect both an altruistic “taste for fairness” as well as the strategic anticipation that small offers may be turned down (Oosterbeek et al., 2004). Earlier research has shown that the amount proposed is dependent on the information given to both proposer and receiver, i.e. the proposer is more likely to make a fair offer if the proposer knows that the receiver is aware of the amounts to be divided (Pillutla and Murnighan, 1995). Also, the proposer is thought to reflect on the mental state of the responder (Campbell-Meiklejohn and Frith, 2012). In other words, the amount offered to the receiver is a function of the perceived capability of the receiver to know and reason with the proposed amount; it is hypothesized that “smart” receivers will be offered a larger stake due to (1) being perceived as able to reason with the proposed amount, and (2) the expectation that smart receivers will keep track of reciprocity, rejecting low offers to punish the proposer.

In the dictator game, the amount given away is considered a “more pure” measure of altruism (Eckel and Grossman, 1996; Fehr and Schmidt, 2006), as the receiver does not have the option of turning down an offer, removing the fear of rejection. Although the dictator game can be considered to measure a more pure form of altruism, factors such as experimental demand characteristics and social norms play a role as well (Bardsley, 2008).

Torta et al. (2013) investigated rejection rates in an ultimatum game in which human participants played as a receiver against a (virtual) proposer that was either human, a humanoid robot, or a computer. In their study, participants rejected offers made by a computer more often than offers made by a human or humanoid robot, although this effect was only marginally significant. However, these findings seem to contradict earlier studies, which have more consistently shown that when offers are made by a computer rather than a human player rejection rates are much lower (Moretti and di Pellegrino, 2010; Sanfey et al., 2003). Sanfey et al. (2003) showed that this is reflected in neural activity, and found weaker activation of the anterior insula when unfair offers were randomly generated by a computer instead of a human opponent in an fMRI study investigating ultimatum game rejection behavior.

In a similar paradigm, van Dijk (2013) also had participants play as the receiver in an ultimatum game. In addition, participants completed anthropomorphism questionnaires in which they rated how much they anthropomorphized their (virtual) opponent, which could either be a human, a robot, or a computer. While this study did not find an effect of opponent type on rejection behavior, a correlation was found between anthropomorphization and rejection behavior, where offers being made by proposers who were anthropomorphized more were less likely to be rejected.

This finding suggests that it is not the opponent *type*, which is often manipulated by experimenters to manipulate different levels of anthropomorphism, but individual *differences in anthropomorphization* that determine rejection behavior.

1.3. The current study

So far, several human-robot interaction studies looking at ultimatum game behavior have focused on rejection rate behavior (Moretti and di Pellegrino, 2010; Sanfey et al., 2003; Torta et al., 2013). However, relatively few studies have investigated altruism using proposer behavior and individual differences in the tendency to anthropomorphize (van Dijk, 2013).

In the current study, we investigated the role of both physical human-robot similarity and the individual degree of anthropomorphization on altruistic and strategic behavior. To assess altruistic and strategic behavior, we used the dictator and ultimatum games. In this study, human participants were proposers, and we used different types of robots as well as a human confederate as receivers. The manipulation of the type of opponent was thought to tap into the physical similarity between proposer and receiver, with the other human being the most similar and the laptop being the least similar opponent.

Due to the criticism toward the use of virtual robots or avatars (Bainbridge et al., 2011; Li, 2015), we used physical, co-present robots. Importantly, we also assessed the individual degree of anthropomorphization. Using this design, we could more carefully investigate the effect of anthropomorphization on altruistic and strategic behavior. We were particularly interested in comparing the impact of physical similarity and individual anthropomorphization on altruistic and strategic behavior. From the viewpoint of physical similarity, the latter should merely reflect the former, which should be the main factor accounting for the degree of altruistic behavior—which in turn should be most pronounced for the human opponent and least pronounced for the laptop. From an anthropomorphization point of view, however, it might be mainly the individual tendency to perceive an opponent as human-like that determines the proposer’s altruistic and strategic behavior.