UNIVERSITY OF
LINCOLN

Determining the effect of human cognitive biases in social robots for human-robot interactions

Mriganka Biswas

A thesis submitted in partial fulfilment of the requirements of the University of Lincoln for the degree of Doctor of Philosophy

Doctor of Philosophy

2016

I would like to dedicate this thesis to my loving parents

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration.

This research has been documented, in part, within the following publications:

Journal articles:

1. Biswas, M. and Murray, J. (2017) The effects of cognitive biases and imperfectness in long-term robot-human interactions: case studies using five cognitive biases on three robots. *Cognitive Systems Research Journal*, 43(C) 266-290. Available from doi:10.1016/j.cogsys.2016.07.007 [accessed 14 May 2017].

2. Biswas, M. and Murray, J. (2016) Can cognitive biases in robots make more 'likeable' human-robot interactions than the robots without such biases: case studies using five biases on humanoid robot. *International Journal of Artificial Life Research* (*IJALR*), 6(1) 1-29. Available from doi:10.4018/IJALR.2016010101 [accessed 14 May 2017].

Conference papers:

1. Biswas, M. and Murray, J. (2014) Effect of cognitive biases on human-robot interaction: a case study of a robot's misattribution. In: *23rd IEEE international symposium on robot and human interactive communication, IEEE RO-MAN 2014,* Edinburgh, UK, 25-29 August.
2. Biswas, M. and Murray, J. (2015) Robotic Companionship: how Forgetfulness Affects Long-Term Human-Robot Interaction. In: *8th international conference on intelligent robotics and applications*, Portsmouth, UK, 14 August.
3. Biswas, M. and Murray, J. (2015) Towards an imperfect robot for long-term companionship: case studies using cognitive biases. In: *IEEE/RSJ international conference on intelligent robots and systems*, Hamburg, Germany, 28 September – 2 October.

4. Biswas, M. and Murray, J. (2016), Robots that refuse to admit losing – a case study in game playing using self-serving bias in the humanoid robot MARC. In: *9th international conference on intelligent robotics and applications*, Tokyo, Japan, 16 August.
5. Biswas, M. and Murray, J. (2016) The influences of 'self-serving' bias in robot-human companionship: case studies using the humanoid robot MARC. Workshop on behaviour adaptation, interaction and learning for assistive robotics. In: *RO-MAN 2016*, New York, US, 14 May 2017.
6. Biswas, M. and Murray, J. (2016) The effects of cognitive biases in long-term human-robot interactions: case studies using three cognitive biases on MARC the humanoid robot. In: *The eighth international conference on social robotics*, Kansas City, USA, 1-3 November.

Posters:

1. Biswas, M. and Murray, J. (2013) Building a long term human-robot relationship: how emotional interaction plays a key role in attachment. In: *Fourth EUCogIII members conference*, 23-24 October, Falmer/Brighton.
2. Biswas, M. and Murray, J. (2013) Developing long-term human-robot interaction. *HFR2013*, 25-26 September, Rome.
3. Biswas, M. and Murray, J. (2014) A model for long-term human-robot interaction and relationships in a companion robot. In: *Sixth EUCogIII members conference*, 17-18 October, Genoa.

Acknowledgements

I would like to express my special thanks to my supervisor Dr John Murray. You have been a tremendous mentor for me. I would like to thank you for encouraging and guiding my research and for allowing me to grow as a research scientist. Your advice on both research as well as on my career have been invaluable.

I would like to express my gratitude to my family for their support and encouragement over the years. Words cannot express how grateful I am to my mother, Namita Biswas and father, Sasanka Sekhar Biswas for all of the sacrifices they’ve made on my behalf. Your prayer for me was what sustained me this far. I would also like to thank to my beloved wife, Payel. Thank you for supporting me for everything.

Abstract

The research presented in this thesis describes a model for aiding human-robot interactions based on the principle of showing behaviours which are created based on 'human' cognitive biases by a robot in human-robot interactions. The aim of this work is to study how cognitive biases can affect human-robot interactions in the long term.

Currently, most human-robot interactions are based on a set of well-ordered and structured rules, which repeat regardless of the person or social situation. This trend tends to provide an unrealistic interaction, which can make difficult for humans to relate ‘naturally’ with the social robot after a number of relations. The main focus of these interactions is that the social robot shows a very structured set of behaviours and, as such, acts unnaturally and mechanical in terms of social interactions. On the other hand, fallible behaviours (e.g. forgetfulness, inability to understand other’ emotions, bragging, blaming others) are common behaviours in humans and can be seen in regular social interactions. Some of these fallible behaviours are caused by the various cognitive biases. Researchers studied and developed various humanlike skills (e.g. personality, emotions expressions, traits) in social robots to make their behaviours more humanlike, and as a result, social robots can perform various humanlike actions, such as walking, talking, gazing or emotional expression. But common human behaviours such as forgetfulness, inability to understand other emotions, bragging or blaming are not present in the current social robots; such behaviours which exist and influence people have not been explored in social robots.

The study presented in this thesis developed five cognitive biases in three different robots in four separate experiments to understand the influences of such cognitive biases in human–robot interactions. The results show that participants initially liked to interact with the robot with cognitive biased behaviours more than the robot without such behaviours. In my first two experiments, the robots (e.g., ERWIN, MyKeepon) interacted with the participants using a single bias (i.e., misattribution and empathy gap) cognitive biases accordingly, and participants enjoyed the interactions using such bias effects: for example, forgetfulness, source confusions, always showing exaggerated happiness or sadness and so on in the robots. In my later experiments, participants interacted with the robot (e.g., MARC) three times, with a time interval between two interactions, and results show that the likeness the interactions where the robot shows biased behaviours decreases less than the interactions where the robot did not show any biased behaviours.

In the current thesis, I describe the investigations of these traits of forgetfulness, the inability to understand others’ emotions, and bragging and blaming behaviours, which are influenced by cognitive biases, and I also analyse people’s responses to robots displaying such biased behaviours in human–robot interactions.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Outline

This thesis investigates social interactions between humans and robots and how these interactions might be improved. The approaches we consider include the introduction of cognitive biases in a robot, and we investigate the effect of those biases on human–robot interaction. The thesis paper presents a model which can be used to develop the social interactions between humans and robots.

In social robotics studies, scientists follow various approaches to make robots' behaviours sociable and understandable to humans, so that humans can easily interact with the robots: for example, robots mimicking human behaviours, expressing various emotions, developing trait-based behaviours and different moods, etcetera. In the current study, I investigate different cognitive biases of humans and develop few biases in the robot to study their effects in the human–robot interactions. The founding assumption of this research is that the effects of cognitive bias exist in humans alongside their emotions, moods and characteristics, and such biases can have an important impact in human interactions. In the existing social robotics research, the role of anthropomorphic emotions, moods and character traits in robots are widely studied, but the role of cognitive biases has not been investigated in detail and thus calls for further study. The research presented in this thesis investigates the effects of such biases in human–robot interactions.

Section 2.1 discusses the motivations and significance of this study. The section offers a brief literature review of the topic and a hypothesis. Subsequently, I state the hypothesis and research objectives. Section 1.3 states the questions I seek to answer in the study. Section 1.4 offers an overview of the remaining parts of the thesis.

1.2 Motivation

This thesis is based on social robot–human interactions. A social robot can interact with others based on social skills (e.g. social rules and norms) which humans follow in daily social exchanges (Breazeal, 2003). A number of studies have used different social robots which exhibit such social behaviours in human–robot interactions. For example, Sony's Artificial Intelligence Robot (AIBO) dog is a social robot which interacts in different ways so that the user can understand; for example, it jumps and barks to seek attention and follows the user around (Sony, n.d.). The AIBO dog has been used in many studies, one of which was to increase diversity in an undergraduate computer science program (Gini, Pearce & Sutherland, 2006) where students had to develop different AIBO-related projects and program the robot to play (e.g. develop a game on AIBO for children with attention deficit/hyperactivity disorder [ADHD]). Another was to develop game activities for AIBO which elderly people could enjoy. Another popular social robot, Paro the robot seal, can learn to behave according to user preferences and respond to its name, as if it is alive, moving its head and legs and making sounds (PARO, n.d.). Paro has mainly been used for therapeutic purposes and in care homes, where it was intended to provide company for the elderly (Shibata & Wada, 2010). Paro interacts differently than AIBO, as it makes noises of a seal, closes it eyes when someone touches its fur, and moves its head and tail (PARO, 2017). These two robots demonstrate such understandable behaviours so that the human interaction is possible. Many other social robots also interact in ways that people can understand; for example, the Keepon (Kozima, Michalowski & Nakagawa, 2009) robot dances and makes meaningful noises; the Kismet (Breazeal, 2003) and KASPAR (Dautenhahn et al., 2009) robots can expresses emotions; and many other robots manifest social cues, skills and interactions based on various anthropocentric behaviours.

Researchers worldwide continue to study the different approaches by which social robots can interact with people. Different social skills and anthropomorphic behaviours are developed in robots so that their behaviours can be recognised and easily understood by people. Social robots are being used in different studies where various human-like behaviours are developed and investigated in robots for human–robot interactions, for example, mimicking human actions (Scheeff, Pinto, Rahardia, Snibbe & Tow, 2002), accompanying musicians (Otsuka, Nakadai, Takahashi, Ogata & Okuno, 2009), walking and jumping in humanoid form (ASIMO, n.d.), dancing to music (Michalowski, Simmons & Kozima, 2009), and doing various other things. Such skills and behaviours in social robots make them superior to similar technologies; however, scientists argue that the actions of such robots are limited and may not provide sufficient reason for people to create and maintain social relationships for a long period of time (Baxter et al., 2012). Baxter, in this case, has pointed out that social robots have memory issues, as occasionally the robots were unable to pick up conversations from a previous interaction. In a later chapter (5.1) of this thesis, we will see that the forgetfulness of a robot can play an important role in human–robot interactions. However, in human–human interactions, people meeting others are able to form different kinds of relationships, but not always in human-robot interactions. This inability raises the question: 'What happens in human–human interactions that is absent in human–robot interactions, thus preventing a social relationship forming between the human and the robot?'

To answer this question, I investigate how humans interact and form relationships. Russell and Shoare (1994) concluded that human behaviour is derived from the thoughts, feelings and actions of individuals, and it develops from 10 variables found within each human being: individual's will, motivation, psychological dimensions, learning styles, cognitive development, beliefs, physiological needs, social and emotional development, language development and spiritual beliefs (Russell & Shoare, 1994). These variables can differentially influence individuals, so an individual will act uniquely in different situations. This individuality does not mean, however, that people behave randomly, but that they behave in certain ways to particular situations. For example, one individual's act of aggressiveness may be wholly different than another's.

A further example clarifies the above point: Some may feel happy watching videos of cats on the Internet, but others may think that the Internet is too full of cat videos. Such differences in reactions are, perhaps, as scientists suggest, based on individuals' idiosyncratic thoughts, social norms and cultures (Haselton, Nettle & Murray, 2005). Social norms are the unwritten rules of how to behave in a particular social group or culture ((Cialdini et al, 1990), (McLeod, 2008)). For example, it is expected that employees arrive at an office on time and complete their work. As James (2015) puts it, 'Culture is defined as a social domain that emphasizes the practices, discourses and material expressions, which, over time, express the continuities and discontinuities of social meaning of a life held in common' (pp. 53). 'Thinking' is the mental process of manipulating information for problem solving, reasoning and decision making (Beresford & Sloper, 2008). One can say, in sum, that individuals can behave differently in particular situations, depending on their own thoughts, social norms and culture. Studies suggest that norms and culture for a large group usually do not change: ‘Culture is thus synonymous with group identity’, concludes Spencer-Oatey (2012, pp. 16), and Richardson (1999) writes, ‘Having a set of norms—or ground rules—that a group follows encourages behaviours that will help a group do its work and discourages behaviours that interfere with a group’s effectiveness’ (pp. 1). Thus, individuals born within the same culture usually accept the norms of that culture. Therefore, from this discussion it can be said that it is the human ‘thinking’ which drives individuals to act differently than others. However, various studies have suggested that human thinking is not strictly rational and may occasionally be affected by variety of cognitive biases (Tversky & Kahneman, 1974).

1.2.1 Cognitive bias

Cognitive bias can regarded as people’s tendency to create their own versions of social reality from their perception of sensory input which is not actual input and which may dictate their behaviour in the social world (Bless, Fiedler & Strack, 2004). For example, thinking that of a certain number will or will not appear in the next round of a lottery just because it appeared several times before is misguided, since the numbers appear randomly. This is called ‘the gambler’s fallacy’, and it is associated with people who gamble frequently (Coltfelter CT, Cook PJ, 1993). Cognitive biases can be classed by their effects: for example, decision-making biases (empathy gap, belief bias, etc.), social bias (the halo effect, self-serving bias, etc.), or memory bias (misattribution, the effect of humour, the bizarreness effect, etc.) (I discuss biases in Chapter 3 and add collected lists of biases in the Appendix G).

In a given context, individuals may act differently based on the effects of their biases on their thinking processes (Gigerenzer & Goldstein, 1996). Cognitive biases can help people to make faster decisions, but with some cost to the reliability of those decisions (Tversky & Kahneman, 1974). Coxall (2013) has stated that cognitive biases explain people’s limited capacity for information processing, and Baron (2008) has suggested that cognitive biases (e.g. social biases) have a reasonable amount of influence on human thinking processes, leading to misjudgements, mistakes and fallible activities. As most relevant studies have suggested that biases play crucial role in decision making, it can be assumed that such cognitive biases have influence the choices made in interactions between individuals in given contexts. Such differences in cognitive responses among individuals seem to make human interactions unique and natural to humans. For example, it is common during social interactions that one party may pay more attention to someone yawning as a sign of disinterest but not so much attention to other cues suggesting interest in a conversation (such as leaning in). Based on such incidents, an individual's behaviour towards the yawning person might change.

Biases are not the only factor which influences the human decision making, but such systematic errors in judgement and decision making are common in all individuals (Wilke & Mata, 2012). Cognitive biases play crucial roles in human thinking, effecting people's decisions about certain things; decisions are not based only on the facts but also on a person's biased thinking. The different types of biases include memory biases (which result in people misattributing and forgetting information), social biases (which effect people's social thinking and interpersonal relationships, e.g. through Dunning-Kruger effects) and many others (lists of cognitive biases are attached in the Appendix G). Similar to emotions, moods, traits and personalities, cognitive biases are one of the main components that influence each individual's thoughts. Unlike moods, emotions and traits, however, the effects of such biases are never tested in social robots for human–robot interactions.

In the social robotics research, robots can imitate human actions. For example, Kismet can understand human affection through vision and sound and express itself with emotive facial expressions (Breazeal, 2004). Recognising the wrong emotions and acting accordingly, however, is also a human behaviour which can happen in many social situations; for example, people often fail to recognise others' emotions in different situations; if someone feels secure, then it is hard for them to discount the idea that other people are anxious (e.g. hot-cold empathy gap bias) (Van Boven, Loewenstein, Dunning & Nordgren, 2013). Therefore, such humans common characteristics (i.e. faults, unintentional mistakes) are, however, absent from these social robots. Many robots can present social behaviours in human–robot interactions but are unable to show anthropomorphic cognitive biased behaviours such as faults, unintentional mistakes and task imperfectness, and people often find social robots difficult to relate to. Many researchers (e.g. Leite et al., 2016; Castellano, Pereira, Leite, Paiva & McOwan, 2009) investigate user engagement issues between human and companion robots, providing game-playing scenarios to make the responses of the companion robots more 'empathic', but in the real world, people can sometimes refuse to accept losing in games and sometimes accuse their opponents of cheating (e.g. self-serving bias effects, Campbell, Sedikides, Reeder & Elliot, 2000).

Modern social robots, such as the robot Nadine built by Prof. Nadia Thalmann from Nanyang Technological University Singapore (NTU) (Kok L, 2015), can recognise past guests and initiate conversation based on previous interactions. The robot can exhibit happiness or sadness, depending on the topic of the conversation. Thalmann stated in an interview that this robot was built to be a companion robot similar to C-3PO from *Star Wars*: ‘This is somewhat like a real companion that is always with you and conscious of what is happening. So in future, these socially intelligent robots could be like C-3PO, the iconic golden droid from Star Wars, with knowledge of language and etiquette’ (Knapton, 2015). However, C-3PO is liked not only for being an intelligent humanoid robot, but also for its fussy and worry-prone personality throughout its decades of operation (Star Wars Wikia, n.d.). In the movie, the robot fears changes and strives to maintain order and peace. Such worry-prone behaviours are common in humans and can be seen in everyday life. Hence, plausibly, if a robot could show anthropomorphically fallible behaviours (e.g. mistakes, faults and biases), it could appear more human-like, facilitating a more natural relationship between humans and robots.

People commonly forget the names and faces of others or misremember past events, brag about their achievements, or blame others for their own failures. Such behaviours arise widely between friends, family members and colleagues, but are absent in current social robots. Many studies have investigated the effects of human-like emotions and expressions (Yannakakis & Pavia, 2014), empathic behaviours (Leite et al., 2016) and personality traits (Lee, Peng, Jin & Yan, 2006) in robots, but cognitive biases are also one of the main factors responsible for human actions manifesting in specific ways (Bless et al., 2004). No research has explored in detail how human interlocutors react if a robot shows such fallible behaviours based on cognitive bias in human–robot interactions. This thesis examines the importance of the effects on people of such biases exhibited in robots, and studies whether such behaviours presented by social robots can influence human–robot interactions.

This research thus focusses on developing cognitive biases in robot behaviours, with regard to verbal and non-verbal abilities, emotions and gestures expressions, and so forth, to investigate their effects in human–robot interactions. The idea of developing cognitive biases in a robot is novel. It is expected that the use of cognitive biases in robots can result in a model which will help people to understand more easily the behaviours of robots.

1.3 Hypothesis

The research described in this thesis recognises the importance of the effects cognitive biases in humans and commits to investigating their effects in the social robots for human–robot interactions. The research examines the following hypothesis:

> *If a robot can exhibit behaviours that include selected cognitive biases when interacting with humans, then humans will have increased fondness for the robot.*

This main hypothesis addresses three key challenges:

- the development of cognitively biased behaviours in robots, which could demonstrate the bias's behaviours properly and in an understandable manner,
- study of cognitively biased behaviours in a robot to find out what kind of effects they have on the participants, and
- study of the effects of such biased behaviours in human–robot interactions for a long period of time to find out whether such influences change.

Based on such challenges, additionally I seek answers to the three research questions:

- Despite a robot's appearance, functions and features, can it interact with humans in ways similar to human-human interaction by engaging with humans in both the long term and the short term and drawing on cognitively biased behaviours?
- With cognitive biases introduced in a robot, is it possible to develop anthropomorphically biased behaviours in robots to influence human-robot interactions?
- Will cognitive biases help humans to develop human-robot interactions in the long term?

As it features in the hypothesis, the selection of cognitive biases is an important part of the current study, as explained in Chapter 3. I investigated the hypothesis and research questions in four different human–robot experiments (e.g. with conversations and game playing) based on the five different cognitive biases, e.g. misattribution, empathy gap, lack of knowledge (based on Dunning-Kruger effects), self-serving and humours effect biases. In these experiments, I developed biased and non-biased algorithms in robots to manipulate their behaviours to show the effects of such biases in interactions with participants. In the non-biased interactions, however, the robot did not show such biased behaviours and engaged in similar conversational and game playing interactions. At the end of each experiment, the participant's feedback was compared to determine the participant's preferred interactions. I further discuss the procedures and results in the later chapters in details.

1.4 Objectives

As the current research investigates a novel approach for developing human-like skills in social robots based on cognitive biases, the objectives of the research are important for the field of social robots and for that of psychology. Apart from its main objective—to develop different cognitive biased behaviours in robots to explore the effects of biases in human–robot interactions— the study has other collective objectives:

- To seek the appropriate cognitive biases to develop in social robots for human–robot interactions.

Because over 200 kinds of cognitive bias have been identified, it is important to understand which biases would be appropriate to develop in social robots and whether it is possible to create any rules or principles by which to select cognitive biases suitable for social robots.

- To determine how such cognitive biases can be developed in social robots.

As this is a novel approach to establish whether cognitive biases in robots can be useful in developing human–robot interactions, it is important to understand how such biases can be used in the robots.

- To study whether a certain cognitively biased behaviour in a robot can change the participant's behaviours towards the robot and how such behaviours change over a long-term period.

Most cognitive biases produce negative responses in human interaction; therefore such biases introduced in social robots may discourage human from interacting with these robots. For this reason, human response to certain biases must be understood.

- To discover and compare the differences in participants' preferences between a robot with a certain cognitive bias and a robot without that bias, over the long term.

Investigating this objective may reveal the effects of cognitive biases on the robot in human-robot interactions over the long term.

- To evaluate the change in the participant's response over time in conversational and game-playing interactions with the selected biased and non-biased robots.

This objective investigates the changes in response even if the users like or dislike their first interactions with the robot based on cognitive biases, since responses may change with repeated interactions over a long period of time between two interactions. It is important to know the effects of such biases over the long term.

The above objectives are addressed through this thesis. In Chapters 5 and 6, I describe the experiments and examine the statistical analysis of the collected data based on the hypothesis and research questions.

1.5 Thesis organisation

Chapters 2–8 describe the research as a whole and all of its experimental details. These chapters are organised as below.

- Chapter 2: Social Robot

The chapter describes the stand of the social robots in society. In this chapter, a brief history of social robots is offered, after which the relation of sociable robots to our current research is described.

- Chapter 3: Long-Term Human–Robot Companionship

In this chapter, the study of long-term human robot interactions is discussed. Existing studies in this field are talked about, and the problem domains are briefly described.

- Chapter 4: Cognitive Biases and Imperfectness in Humans

This chapter focusses on the biases in humans and why they are important, the common characteristics flaws in human behaviours and how such behaviours in social robot during human-robot interactions can impact the interactions between robots and humans.

- Chapter 5: Interaction Algorithms and Experimental Setup

In this chapter, the model of biased and non-biased algorithms used in the experiments and the basis of each of the different experiments are discussed. Also discussed is how such algorithms are formed and applied in the robot for specific interactions. The experimental details, participants, groupings, settings, measurements and other details are described in this chapter.

- Chapter 6: Statistics for All Experiments

In this chapter, the general statistics of the collected data from all the experiments described in the previous chapter are presented. Data are shown for each experiment as per the set dimensions, assessment of the biased and unbiased interactions and the algorithms effectiveness.

- Chapter 7: Discussion of Hypothesis and Research Questions

In this chapter, the results of the previous chapter are discussed. After that, the statistical analysis from the previous chapter, based on the hypothesis and research questions, is discussed.

- Chapter 8: Future Work

In this chapter, the possibilities for future research are briefly described, the research as a whole is discussed and the thesis concluded.

Chapter 2

Social Robots and Human–Robot Interaction

2.1 The social robots

This research investigates the effects of cognitive biases, introduced in social robots, on human–robot interactions. The aim of this study is to make social robots more cognitively anthropomorphic in order to improve users' experiences in human–robot interactions. In recent years, many studies have suggested that human-like behaviours in robots can improve the quality of human–robot interactions, as the user finds it easy to understand such behaviours, which helps them relate with the robot. In the following sections, I discuss the importance of social robots equipped with various social skills. In the discussions, I point out the skills in such systems and discuss the possibilities of improving interactions.

The concept of the social robots must be defined for the purposes of this study. Breazeal (2003) defines a social robot as a robot which can have social interactions with people. To create such 'social interactions', it is important to develop knowledge in the robot of, for example, the social norms and cues that humans follow in regular social interactions. Currently, social robots are used in many different fields, for example healthcare, tourism, reception, education, hospice and many more, where robots engage with the users in social interactions. Breazeal (2003) has proposed four subclasses of the existing social robots:

- 'socially evocative' robots—designed to resemble humans and other social entities (e.g., robot pets, animated characters);
- 'social interface' robots—designed to use anthropomorphic modes of interaction (e.g., facial expressions, eye contact, facial analysis, etc.). Robots with such social interfaces can be used as tour guides, receptionists, and so on, to engage people with understandable expressions, gestures, speech, and so forth;

- ‘socially receptive’ robots—designed to learn from their interactions with people (e.g., learning new words, geometrical shapes, gestures, etc.); and
- ‘sociable’ robots—designed to engage with people to help them to perform tasks and to update their own performances. These robots interact with people following social cues and anthropomorphic cognitive behaviours, developed from the field of psychology.

Describing the vision of a sociable robot, Breazeal (2002) states, ‘…a sociable robot is socially intelligent in a humanlike way, and interacting with it is like interacting with another person’ (pp. 1); this definition indicates that people should be able to relate to socials robot as they relate to other people or objects. Dautenhahn (2007) identifies various definitions of social robots from the literature, including Breazeal’s (2003) ‘socially evocative’ and ‘sociable’ robots. The other definitions she raises are as follows:

- ‘socially situated’—as earlier mentioned by Fong, Nourbakhsh, and Dautenhahn (2003), these robots are placed in a social environment, reacting socially to different social entities. This definition states that the social robot should be in the society and should be able to understand and react to the different objects in the environment; and
- ‘socially intelligent’ robots—Dautenhahn (1998) also describes that socially intelligent robots can be based on human-like cognition and social knowledge and should be able to demonstrate human-like social intelligence in human–robot interactions.

Synthesising the above definitions, one can say that a social robot should be able to understand and express its social intelligence in interactions, and it should be able to perceive and update its social knowledge by continuing such interactions. Researchers worldwide are studying different social skills to develop in social robots to express and understand human social interactions. Such social skills should also help humans to understand and relate to the robots as they can relate with others. So far, the various social robots show different social abilities (e.g. Advanced Step in Innovative Mobility [ASIMO] can walk, jump, talk; the Flexible LIREC[1] Autonomous Social Helper [FLASH]) humanoid robot can express emotions in interactions (Leite I *et* al, 2012); Paro robot can learn to behave in a way that the user prefers; and many other robots might also be mentioned). In the next section I discuss the importance

[1] LIREC (Living with Robots & Interactive Companions) http://lirec.eu/flash

of such social robots with different social skills in the various areas of the modern society. In the later section (2.3), I discuss the skills that the social robots use in human–robot interactions (HRI) and what more could be added in order to develop HRI more human friendly.

2.2 The importance of social robots

In the previous section, I have discussed the various definitions and subclasses of the social robot, and in this section I discuss the importance of such social robots in our society. More specifically, I explain various applications of these robots in different sectors of society, for example healthcare, education, house maintenance, hospice, retail and others.

The ageing population is a major concern in the UK. Based on the information from the UK government project called 'Future of an Ageing Population', in the coming decades the average age in the UK population is expected to increase significantly (Government Office for Science, 2017). In this scenario, the social robot could be useful in assisting elderly people in need. Robots can be programmed with caring and empathetic behaviours, obeying users unconditionally, and they can provide many entertaining applications useful in helping the elderly. In Europe, many studies are proceeding with collaboration from different countries to develop care robots for the elderly in care homes and hospitals: for instance, Hobbit[2] (Pripfl J *et* al, 2016), Robot-Era[3] (Subash S *et* al, 2013), EnrichMe[4] (Bellotto N, Fernandez-Carmona M and Cosar S, 2017), the Mario Project[5] (Casey D *et* al 2015), and more. The reasons for such interest in the use social robots in elderly care include ageing populations and new technological advancements. Projects such as HOBBIT aim to support independent living for the elderly by providing physical and cognitive assistance (Pripfl J *et* al, 2016). The information on the project website indicates that with this project the user and robot can develop companionship based on taking care of each other, so that users accept the robot in their houses. The European project ENRICHME offers an intelligent robotic system which interacts and monitors elderly users in a smart-home environment (Bellotto N, Fernandez-Carmona M and

[2] HOBBIT – The mutual care robot, http://hobbit.acin.tuwien.ac.at/index.html

[3] Robot Era (The Robot-Era Project has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement num. 288899 FP7 - ICT - Challenge 5: ICT for Health, Ageing Well, Inclusion and Governance) http://www.robot-era.eu/robotera/index.php

[4] EnrichMe - ENabling Robot and assisted living environment for Independent Care and Health Monitoring of the Elderly, http://www.enrichme.eu/wordpress/about/

[5] Mario project - Managing Active and healthy aging with use of caRing servIce robots, http://www.mario-project.eu/portal/

Cosar S, 2017). If the robot notices any emergency situation, it can directly contact the care assistant and the family members for immediate assistance. In the Robot-Era project, the robot Johnny offers various services to its elderly users: shopping, cleaning house, washing the laundry, ordering food, calling friends and family, and organising events, among other things (Subash et al, 2013.). In this project, the entire system uses three different robotic platforms (i.e. domestic, condominium and outdoor), each intended for specific tasks; for example, the user orders shopping using the interfaces of the domestic robot, the outdoor robot collects the ordered goods from the store and brings them to the condominium robot at the house, and this condominium robot then transfers the goods to the domestic robot (Subash S et al, 2013). In most of the care-robot projects, the robotic system offers multiple ways to interact with the robot, for example based on their health and preferences, the user can interact with the robot using speech, a tablet computer or even simple gestures. In these projects, the robot provides companionship to the users and monitor patients alongside the caregiver to offer the cognitive mental supports. However, since some elderly people are less technologically advanced, the interaction with the robots must be easy, so that they can relate to the robot in order to receive all the benefits. In this thesis, I study such human-like skills in robots to make their behaviours more familiar and understandable to users, so that the interactions are much easier for them.

Studies have suggested that social robots such as KASPAR can be useful in educating the autistic children to improve their social interactions (Syrdal, Lehmann, Robins, & Dautenhahn, 2014). This robot is a minimally expressive child-like humanoid robot developed by the Adaptive System Research Group at the University of Hertfordshire (Dautenhahn, 2009). The developers’ study concluded that KASPAR’s minimal expressions and imitative behaviours, playing a drum or peek-a-boo, helps autistic children to understand and learn basic behaviours, because the children tend to pay much attention to the robot (Dautenhahn, 2009). Another recent study (Valadão et al., 2016) between a robot called MARIA and children with autism spectrum disorder (ASD) found out the robot to be useful in enhancing the social skills in ASD children because they reacted well to the robot. In their research, children with ASD looked at the robot, touched it, and paid much attention to the interactions and the robot where the robot plays mediator role (Valadão et al., 2016). In recent years, studies using social robots to teach social behaviours to ASD children have become more common (among others, Bharatharaj, Huang, Mohan, Al-Jumaily & Krägeloh, 2017; Nelson, 2013; Feil-Seifer & Mataric, 2009). Ricks and Colton (2010) have argued that such robot-based therapeutic trends can indeed help autistic children, proven through the social skills children show during their interactions with the robots. However, they mention also that robot-based studies need to extend beyond the laboratory to examine how well children can interact in the real world, with the children without ASD (Ricks & Colton, 2010). Clearly, social robots can play an important role in engaging ASD children to teach them social skills. This insight has inspired this thesis to investigate the different social behaviours of humans as expressed in regular social interactions, and how such common behaviours can benefit human–robot interactions if developed in social robots.

The social robots are being used in the collaborative learning, where children can memorise and learn language, science and technology (Mubin, Stevens, Shahid, Al-Mahmud & Dong, 2013). In their review on the acceptance of the robots in the education, Mubin et al. (2013) concluded that robots can provide a tangible and physical representation of learning outcomes which could be valuable for employing the robots in education. Preschool children can be naturally attracted to some robots, for example the BeeBot robot (Janka, 2008). The BeeBot is a colourful bug-like programmable robot used in a kindergarten to study how young people learn and respond to technology with the help of a robot (Janka, 2008). Janka (2008) concluded that using robot to teach children could be appealing. Not only the robot, moreover, but robot construction kits may be useful in the classrooms to teach children about programming, technology and engineering (Sullivan & Bers, 2015). In my understanding, such collaborative lessons require children and robots to work side by side, so it is important that the actions of the robot be easily understandable for the children. To be understandable for children, a robot must present behaviours easily recognised by children. For example, in collaborative games, if a robot always loses and lets the child win, the child may get bored; on the other hand, however, bragging after success and blaming after losing are well-known human behaviours that arise even among friends and families (e.g. self-serving bias, De Michele, Gansneder & Solomon 1998). Therefore, in playing games, if the robot wins and brags, then the children may feel challenged, and they may be more likely to play again to defeat the robot. This competitive reaction could make child-robot interactions more enjoyable to the children as well.

Not only in healthcare and education can the social robots be used as alternative helpers, but also in stores and shopping malls. According to an article published in Consumer News and Business Channel (CNBC) (Kercheval, 2015) the chief executive officer (CEO) of the International Council of Shopping Centers discussed the possibilities of using the robots in shopping malls to provide useful services such as carrying products, pushing or carrying carts, helping customers navigate, assisting them in finding specific products, and so on. Based on the IEEE Spectrum article (Ackerman, 2016) Robots such as Budgee can follow customers around like a trusted friend and guide. In shopping malls, these robots may meet different types of people, of all ages and backgrounds, who would respond to such technology differently; for example, some might be too enthusiastic, while others might be confused or frightened, and still others might not interact properly, thinking that the robot might be too complex to operate. Therefore, these robots in such public places need to be more socially active by showing much more anthropomorphic behaviours.

Based on an article published in the *New York Times* (Markoff, 2013), a guard robot, Knightscope's K5 (Knightscope, n.d.), can help security guards and police do their jobs more efficiently and safely. The robot is able to detect guns and anomalies, collect video, record audio and even has forensic capabilities (Knightscope, n.d.). However, such a security guard robot needs to exhibit humanoid behaviours in order to work alongside human guards in a team, and it must also show such behaviour in communicating with the public. In this thesis, I try to determine how such robot behaviours can be made more human-like, so that working alongside people in public places can be made easier for both humans and robots.

Social robots can furthermore entertain their users, like pets in households. In a study using robotic pets in the households to understand children’s behaviours towards living and robotic dogs, Melson et al. (2009) found that children usually become attached to such robotic pets, and they sometimes even feelings as if the pets are alive. Francis and Mishra (2009) observed similar results in a study with children and the Sony AIBO robot dog: Children treated the interactive robot dog as ‘real’, with intentional qualities, even knowing that such toys are not real. Sparrow (2002) has argued that even if people enjoy having a robot pet for entertainment and social interaction, but having such artificial pet might have ethical issue in case of a socially isolated elderly person. However, as people tend to enjoy robotic pets, to develop human–pet companionship such robotic pets need to show known and understandable behaviours to their owners, in order to grow the relationship quickly and firmly. In the current thesis, I seek human behaviours that can be developed in robots so that people may easily recognise and relate to their behaviours.

The examples in this section describe the usability of social robots in various fields of society, showing the relevance and promise of social robots. I have described the different studies and applications of social robots’ various social skills in interacting with people, for example eye movement, emotion, gestures, human-like behaviour, and verbal and non-verbal communication. It is my understanding that robots' interactions with users should be as simple as human–human interactions and as easy to understand. To this end, the robots must be given various social skills. Presently, research worldwide have developed various methods and social skills for robots to make their behaviours understandable to their users (e.g. imitating human behaviours or otherwise expressing anthropomorphic behaviours). In Section 2.3, I briefly discuss the different social skills developed in the robots for human–robot social interactions.

2.3 The social skills used in robots

In the previous examples of the different social robots (e.g. Kismet, KASPAR, AIBO), the robots used various social skills, for example eye contact, emotional expression, verbal or non-verbal sounds, and so forth, to interact with people. All such skills in robots were developed based on familiar social skills among humans and other social animals. In this section, I exemplify some of the important skills of robots, including expressions of mood and emotion, trait-based personalities, and other anthropomorphic behaviours.

2.3.1 Emotions

Expressing emotions is one of the major human-like skills robots can have, and various studies have shown the importance of social robots possessing emotional understanding. For example, Breazeal's (2003) Kismet robot can express emotions through various facial and auditory expressions. Kismet is known for its anthropomorphic expressions and natural ability to engage with people. The robot has been designed to support several social cues and skills that may play important roles in socially situated learning with a human instructor (Kismet, n.d.). Kismet was an early-stage robot built by a research group at the Massachusetts Institute of Technology (MIT), and their most recently developed robot, Nexi, is capable of learning from human behaviours (Siegel M, Breazeal C and Norton MI, 2009).

Another expressive robot, KASPAR, mentioned earlier, can express emotions and helps autistic children learn social skills (Figure 2.2). This robot's face is capable of range of simplified expressions, with some of the complexities of a real human face. It has movable arms, head and eyes, which can be controlled by the teacher or the parent, but also can respond to the touch of a child. The robot was designed to be inexpensive and to explore the reactions of autistic children. It has been used in many autistic studies since its development in 2005, and KASPAR has been designed for use as a social mediator, thus encouraging and helping children with autism to interact and communicate with adults and other children (Wainer J *et* al, 2014). The robot was built to express predictable and repetitive expressions so that the autistic child finds the robot easy to interact with (Wainer J *et* al, 2014).

The robot KASPAR can show other human-like skills, such as blinking and movement in the mouth and neck and arms, but his rubbery face lacks the intricate detail of a human face. Teachers and parents have been impressed by the results that they have witnessed; some saw their child interact, mimic or make eye-contact for the first time in their lives (Wainer J *et* al, 2014).

Many robots are used in research of human–robot interactions based on emotional expressions and understanding. One of the most popular is Keepon (Figure 2.1), a small yellow robot capable of anthropomorphic expressions. Keepon has been used in studies with autistic children, in schools and in much HRI research (Kozima et al., 2009). It is touch sensitive, reacts to emotions using sound, dances upon hearing music, and reacts when someone pokes, pats or squeezes it. Keepon’s simple appearance and behaviours are intended to help children with developmental disorders such as autism to understand its attentive and emotive actions (Kozima et al., 2009).

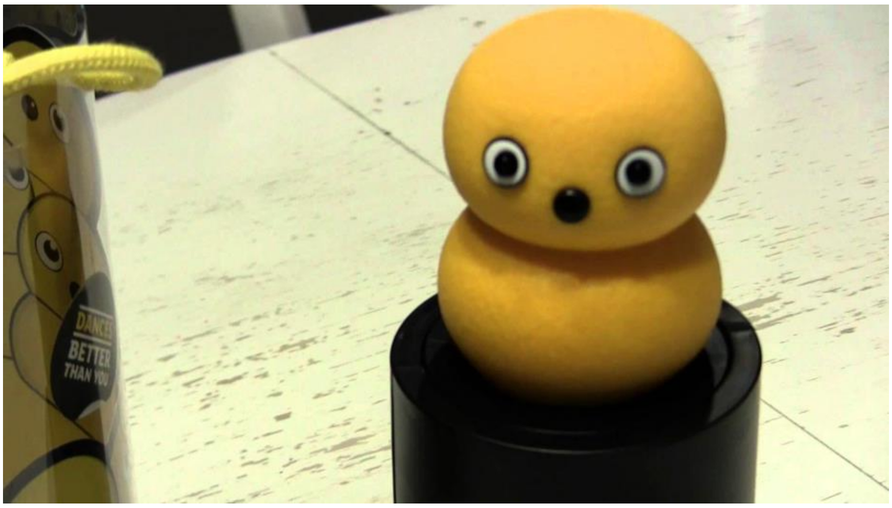


Figure 2.1. The Keepon robot

The above examples suggest that expressing emotions is an important skill, which helps robots to interact with people. The examples above show that robots expressing emotions can help to teach various aspects of communication to children with ASD and can be used as therapeutic devices for those children. In other words, it can be said that social robots expressing emotions help children to relate to the robot, so they can understand the robot and learn from the robot important social communication skills.

2.3.2 Mood and personality traits

Human personality is one of the most researched areas in psychology. Scientists study trait theory to study human personalities. Traits are the habitual patterns of people such as emotions, thoughts and behaviours (Kassin, 2003). In the current world of social media, accessible communication, global migration and the Internet, it is important to understand the behaviours and attitudes of people. Character and personality can be used to describe an individual's behaviours, but it is likely that overall character is revealed through time, while characteristics are easier to notice in the short term. Different personality traits are the reason why people differ from one another. Traits usually are consistent and stable, so they have been studied based on a set of trait dimensions such as the five-factor model and the HEXACO model of traits (Ashton & Lee, 2007), among others.

In the Robotics study, researchers have found that if a robot can interact based on human-like mood and various personality traits, similar to emotional expressions, then users find easy to interact with the robot. For social robots, Moshkina, Park, Arkin, Lee, and Jung (2011) developed a behavioural model called time-varying affective response for humanoid robots at the Samsung Research Lab. The model's complex architecture was based on traits, attitudes, moods, and emotions which could be developed in humanoid robots (which is why it is often called TAME). The model was built for effective human–robot interactions based on the robot's affective state being easily interpreted by the user (Moshkina & Arkin, 2009). Their studies with large numbers of the participants show that a story-telling, socially expressive robot is more likely to be perceived as social engaging than other story-telling robots that are less expressive or non-expressive (Moshkina et al., 2011). This model suggests that if the robot can show such anthropomorphic social skills, it can maintain more attention and engagement from its users.

Meerbeek, Saerbeck, and Bartneck (2009) have proposed and designed model for user-centred personality for domestic robots. Based on the individual user, they have proposed that personality in the robot can be developed over five steps: creating personality profile, designing the expressions, sketching scenarios, visualising objects with 3D animation and evaluating the design (Meerbeek et al.,2009). Their previous research investigated users' control over a robotic TV assistant based on personality traits, and they concluded that 'personality' in robots 'can be used as a means to increase the amount of controls that user perceives' (Meerbeek, Hoonhout, Bingley & Terken, 2006, pp. 1). Their study suggests that personality in a domestic robot could be useful for users and may support human–robot interactions.

Lee et al. (2006) have investigated behaviours of extraverted and introverted personality traits on a popular dog robot called AIBO in human–robot interactions. The robot expressed and differentiated between extraverted behaviours and introverted behaviours by manipulating its voice, verbal and non-verbal cues, LED lights, movement angle and speed. The study found that 'individuals regarded a robot with a complementary personality as being more intelligent, more attractive, and more socially present than a robot with a similar personality'. This study, much like the previous example of the robotic TV assistant (Meerbeek, Hoonhout, Bingley & Terken, 2006), shows that behaviours based on personality traits in robots may make the robots more likeable, hence may be useful in human–robot interactions.

2.3.3 Anthropomorphism

Anthropomorphism is another popular area of study in human–robot interactions, where human attributes are put into the robots. Humans interactions with other humans, animals and objects are based on their conceptions of such entities and, they try to understand such entities and interactions by projecting their existing concepts onto such animals and objects (Fussel, Kiesler, Setlock & Yew, 2008). This phenomenon can be observed in social robots and other technologies as well; for example, in a recent study Lee, Lau, Kiesler, and Chiu (2005) have observed that people create predictable mental models about robots' knowledge based on the limited information they have. Previously, Reeves and Nass (1996) showed that humans can engage in social interactions with computers and other technology. It is therefore possible that humans can easily anthropomorphise social robots, regardless their appearances (e.g. machine-like or humanoid).

In a review of the literature on anthropomorphism and human likeness, Fink (2012) has suggested that robots with social cues can hold people's attention, which could help to develop acceptances of robots by their users. In addition to the issue of acceptance, DiSalvo and Gemperle (2003) have stated that anthropomorphic forms can solve design issues and can be used to project and reflect human values in other objects and robots. Therefore, in order to gain social acceptance, the social robot should be developed based on human social skills and values so that people could easily relate to it. The reason for anthropomorphism's ability to foster such 'acceptance' and 'relatability' may be, as Duffy (2003) has suggested, that anthropomorphism provides powerful physical and social cues suggesting humanity, which can be implemented and used in social robots 'to a great extent' for human–robot interactions. However, he argued, anthropomorphism is not be the solution of the ethical and implementation issues in HRI. This argument also can be supported by another case study (Eyssel, Hegel, Horstmann & Wagner, 2010) on the emotional, non-verbal cues of robots, which found that people responded more positively to the robots showing emotional behaviours and social cues in interactions than those with only functionality.

From this discussion, it can be said that human-like behaviours, social cues and skills can be helpful in anthropomorphising robots, perhaps enabling people to relate with and accept them.

2.3.4 Faulty behaviours

Apart from mood, emotions and trait-based personalities, recent studies have focussed on faulty behaviours in robots to uncover their effects on human trust in robots in collaborative tasks. Salem, Lakatos, Amirabdollahian, and Dautenhahn (2015) have studied how the perception of erroneous robot behaviour influences human interaction choices and willingness to cooperate with the robot. The hypothesis was tested in a home environment and with a robot which had both faults and correct behaviours, each set as two different modes of interaction. Their study suggests that people do not trust to the unusual requests from robots and prefer to trust the robot in the correct conditions. Although the purpose and methods of this study were totally different than those of the present research, in the further chapters I discuss that, it is possible if the robot shows behaviours based on cognitive biases in conversation and game playing then people usually like the interaction more than the interaction without such behaviours from the robot.

The study of faulty behaviours in robots and excess trust of participants were pointed out by Robinette et al. (2016). In their study, participants followed the robot in an emergency, despite having observed the same robot performing poorly in a navigation guidance task just minutes before. They concluded that the robots have the potential to save lives in emergency scenarios, but could have an equally disastrous effect if participants trust them too much. This thesis does not address trust issues between humans and robots, however; rather, it studies making a robot's behaviours more human-like by applying biased behaviours in robot. It is noted that, as in this example, humans can trust a faulty robot in emergency situations.

As we have seen, researchers have developed various social skills in robots to make their behaviours more human-like, such as emotions, trait-based personalities and moods; such human-like components make human–robot interactions more effective. In the research presented in this thesis, I study common human behaviours such as mistake making, bragging, forgetfulness and such relatedly biased behaviours in social robots, specifically in their connection to human–robot interactions. It is hoped that such behaviours in a robot could make human–robot interaction more enjoyable, familiar and understandable for people.

2.4 Problem domain and solution angle

The studies presented in this chapter have described different methods used by researchers to develop social skills in robots for human–robot interactions. In the previous section, I presented various studies in which emotional expressions, trait-based personalities, social cues and skills, and erroneous behaviours were instilled in social robots. Researchers developed such behaviours in these robots so that users can easily anthropomorphise and relate to them. In my understanding, human behavioural characteristics also include faults, unintentional mistakes, cognitive biases, and tasks imperfectness—which are still absent from such social robots. Sometimes, a robot's social behaviours lack that of a human's common characteristics such as ideocracy, humour and common mistakes, but such behaviours are very common in people. Most of the robots described in the examples till now, show the social cues in human–robot interactions but lack human-like, naturally flawed behaviours, and sometimes human-like behaviours presented by the robots do not meet the standards expected by humans. As we have seen, researchers have developed personality traits, various moods and emotions, and other human-like behaviours in robots, to make them more sociable. These robots express emotions and interact based on the developed traits, but in human social interactions, making unintentional mistakes in judgment and showing various cognitive biased behaviours (e.g. brag, blame, ideocracy, etc.) are common. The absence of such natural biased behaviours (e.g. brag, blame, ideocracy, etc.) in the social robots interactions may be one of the many reasons that social robots remain far from an ideal state allowing 'companionship' between robots and humans. In current robotics, most of the time, it is expected that the robot could do a task repeatedly without asking questions or making mistakes and without showing any fatigue, but in reality, expressing tiredness and boredom are very common human behaviours, and such behaviours are present in all living creatures; these may be the key to fostering human–robot companionship. Therefore, this issue needs to be addressed and investigated that if a social robot shows cognitive biased behaviours during human-robot social interactions then how people responds to that.

In the current study, I examine human behaviours based on the common cognitive biases and develop such biased behaviours in robots to investigate the responses of participants in human–robot interactions. It is expected that biased behaviours in robots could help people to relate to them and make the interaction more likeable to people.

2.5 Summary

At the beginning of this chapter, I discussed briefly the different definitions of the social robots found in the literature. I then discussed the importance of the social robots, where I stated different EU-funded projects for elderly care and discussed the different roles for the social robots in the different fields, such as autism therapy, collaborative learning, robots in the shopping mall retail, robots in the security, household pets, entertainment and others. I then discussed the different social skills used in such social robots, for example expression of mood and emotion, trait-based skills and also faulty behaviours, helping people to anthropomorphize the robots. After discussing the different social skills used in social robots, I briefly stated my argument for using humanlike cognitively biased behaviours. This chapter offers, in sum, an overview of social robots, the importance of such robots and the different skills robots use to interact with people.

In the next chapter, I discuss how existing social robots are being used in various interactive studies, and I then discuss the proposed solution to the issues I have raised.

Chapter 3

Introduction to Cognitive Biases in the Social Robots

3.1 Interactions between human and social robots

Chapter has discussed the definition and importance of the social robots, along with the various social skills used in such robots and how such social skills help robots to interact and relate with people. I presented various examples and studies based on such human–robot interactions, where robots use human-like skills and social cues to interact with people. In this chapter, I further illustrate human–robot interactions with examples, parsing each of the examples. I then justify why considering cognitive bias in robots could be beneficial, after which I construct the hypothesis for the present study. I close by discussing the set of cognitive biases investigated in my study.

With the increase of robots in houses and the workplace, and with an aging population, it will become necessary to interact with and work alongside robots. Robots are already present in households in different forms to assist people in various tasks; for example, cleaning robots such as Roomba (iRobot, n.d.); lawn mowing robots such as Bosch' Indego (Bosch, n.d.); pet-care robots such as Litter Robot (Litter Robot, n.d.); personal companion robots such as Buddy (Blue Frog Robotics, n.d.); butler bots used in luxury hotels, such as Charley (Dormehl, 2016); mopping bots which mop floors, like Bimby (Super Kitchen Machine, n.d.); kitchen robots that help with cooking, like Moly (Moly, n.d.); and assistive and care robots such as Paro (PARO, n.d.) that help the elderly and children. Although such robots for helping humans in domestic areas are not meant for social interaction, if such domestics robots could have meaningful social interactions, working side by side with them for long periods of time would be much easier for both parties (Breazeal, 2003). For example, an elderly care system called Gecko CareBot has been performing in-home trials as an elderly-care personal assistant since 2002. As Gecko claims, 'trials have made a valuable contribution to the development and design of the CareBot'

(Gecko Systems, n.d.-a). Such elderly-care robots may make very useful companions, as they are capable of taking care of their uses. As the Gecko website state, 'The CareBot™ home care robot is capable of assisting in senior care in a variety of real-life situations, such as frequent welfare checks, on-time medication reminders and virtual visits by family members that give a sense of safety, security and most of all, being reminded that they are a loved and cherished member of the family' (Gecko Systems, n.d.-b). This statement suggests that the robot can be an important part of the lives of elderly people. As the GeckoSystem website claims, such care-giving robots not only help in medical emergencies, but also take care of patients' psychological comforts (Gecko Systems, n.d.-b).

However, caring for the elderly with modern technology such as robots may be complex, because of elderly people's health issues (e.g. poor eyesight, hearing problems or memory issues), so communicating with robots may be difficult for them. In these situations, robots need to feature easy-to-understand interfaces and anthropomorphic behaviours to provide physical and psychological comfort for the elderly. I investigate how it is possible to make social robot's behaviours more human-like, to provide such psychological comfort.

For instance, in the field of elderly care, one of the popular robots, Paro, interacts based on touch, sound, eye contact and memory of users' interaction choices. The robot serves as a low-maintenance alternative to cats and dogs, and studies (Jøranson , Pedersen, Rokstad & Ihlebæk, 2015) have found that Paro could effectively help dementia patients reduce agitation and depression. Robinson, MacDonald, Kerse and Broadbent (2013) found that robotic pet companions could be helpful in decreasing the loneliness of elderly people in hospitals and rest-homes. They compared the psychological effects of interactions with the robot pet Paro and a resident dog and analysed the behaviours of the elderly while they were communicating with the robot and the dog. Their study concluded that Paro was a positive addition to the hospital environment and also that it had benefits for older people in nursing homes, as the robot was able to address some of the unmet needs of older people, particularly relating to loneliness, a need which the resident animal could not meet (Robinson et al., 2013). However, Turkle, Taggart, Kidd, and Dasté (2006) have argued that such robot pets create an illusory relationship in order to develop companionship for the elderly and those who find human relationships challenging, which raise ethical problems. In their review of the literature, Shibata and Wada (2011) have argued that the organisations conducting the study with Paro were subject to ethical committees and had participants and their relatives agree to join. They focussed on the benefits of using robots in therapy and concluded that Paro could be a social mediator for the elderly and that interacting with it could improve their quality of life and reduce loneliness (Shibata & Wada 2011).

The above examples show that the social robots can help the elderly to reduce their feelings of loneliness using various skills and abilities. It is understandable that in the field of healthcare, robots should exhibit many recognisable human-like behaviours, so that relating to the robots becomes easy for elderly people. The research described in this thesis studies existing skills and human-like behaviours in social robots and investigates how they can be improved further.

Many studies support the argument that social robots may be helpful as pets and companions not only for the elderly, but also for people of all ages. Barlett, Estivill-Castro, and Seymon (2004) found that children saw AIBO as a robotic pet rather than a machine. This robot pet interacts based on dog-like barking, jumping, running and tail wagging. It also follows users around the room. Bartlett et al. (2004) note that such behaviours in this robot pet helped children to quickly become attached in a way similar to how they become attached to living dogs. Melson et al. (2009) investigated the difference in the behaviour of children towards robotic and living dogs. Their study found that children compared the live dog to AIBO; however, a surprising majority of children conceptualised and interacted with AIBO as if it were a living dog. This study suggests that it is possible that a robot dog could be conceptualised as living dog if it can show such pet-like behaviours. Using a robot pet as a companion has gained popularity in recent years, and Turkle et al. (2006) have reported that elderly people accepted the robots more than others: ‘Seniors felt social “permission” with the robots presented as a highly valued and “grown up” activity’ (pp. 7). Turkle (2006) mentioned a woman’s comment on Sony’s household entertainment robot dog AIBO, which supports the concept of having household robotic pets: ‘[AIBO] is better than a real dog … It won't do dangerous things, and it won’t betray you … Also, it won't die suddenly and make you feel very sad’ (pp. 9). It is considered that a robotic pet could achieve popularity among some people, as robots possess a type of immortality, so can therefore live with the family for a long time. Robotic pets do not have the issues of feeding and regular defecation and urination. Such robotic companions can therefore be used in the home with children and the elderly for a long period of time without the stresses of regular care. Such studies suggest that robotic systems have may be used as pets in houses and care homes. In this research, I investigate more common human behaviours (e.g. forgetfulness, bragging, misunderstanding emotions, etc.) in robots to uncover their effects in human–robot interactions.

As discussed earlier, researchers have developed different skills and abilities (e.g. emotions expressions, traits, anthropomorphic behaviours) in robots to make them much more understandable to people in order to make them acceptable in society. In 2000, there was a famous ‘baby’ robot called My Real Baby (Hafner, 2000), which used the emotive expressions of children to interact with users. This robot was described as an ‘interactive emotionally responsive doll’. It could manifest many facial expressions such as, happy, sad, angry etc. and could also blink and suck its thumb or a bottle. This robot can change its emotive expressions depending on how it is treated. In an interview, Dr Brooks (the founder of iRobot) described the robot as ‘… a toy which can enhance fantasy’ (Hafner, 2000). The robot can cry, but is able to stop if the user bounces it on a knee, and if the user rocks it to sleep, its eyes grow heavy. Such abilities helped people to anthropomorphise the robot. It basically imitates the behaviours of a child, and that made this robot very popular among children, and even adults. My Real Baby is one of the classic examples of diverse human attributes developed in a robot. This robot’s anthropomorphic behaviours made acceptable by children and adults in 1990s households. This robot made it evident that human-like abilities could make the robot more anthropomorphic, which could lead to better interactions between humans and robots, even the development of companionship. This thesis investigates further human-like behaviours and skills to develop in robots, so that people can easily anthropomorphise and relate with the robots.

The idea of interacting and developing companionship between humans and machine is far older than My Real Baby, however. Previously, a computer program called ELIZA was developed by Weizenbaum (1976), which could engage people in dialogue-based conversation similar to that of a Rogerian psychotherapist. For example, a user saying, ‘My mother is making me angry’, might make the program respond with, ‘Tell me more about your mother’, or ‘Why do you feel so negatively about your mother?’ Turkle (2006) states that Weizenbaum was disturbed that his students fully knew that they were talking with a computer program and wanted to chat with it, and indeed, wanted to be alone with it. It is surprising that ELIZA, without having any embodiment, could secure such intimate interaction among students. This observation opened the idea of development companionship with artificial objects, along with its associated benefits.

To make people more accepting of a machine or robot and more willing to interact, in this study I put imbue the robots with certain attributes, which should be familiar and easy to understand. Thus, I investigate human attributes never before explored before in robots (e.g.

cognitive biases), but common and natural among humans. I expect that such behaviours (e.g. misattribution, empathy gap) in robots make their behaviours more familiar and easy to understand for people.

Various studies have proven the usefulness of using robotic companions as helpers. In the US, an experimental robot has been used within public places to help tourists. For example, Gockley et al. (2005) examined a robot receptionist called Valerie, used to provide useful information to visitors; Valerie even had a story for its character. Frequent visitors were encouraged to repeatedly interact with the robot, so that the changes made in the robot's voices, emotions and appearances could be recognised and used in such interactions. Initially, the robot interacted with the visitors for only few seconds, without any expressions, but in a later version, Valerie could display a positive, negative or neutral mood, which people were able to accurately distinguish. Researchers then developed mood-based emotional interactions in Valerie to make it more appealing to tourists. Such positive, negative and neutral moods trigger different dialogues and emotions, which helped the robot to present itself in a more human-like way. Gockley et al. (2005) determined that the mood, emotions and other anthropomorphic behaviours were important factors in such human–robot interactions. In the research described in this thesis, I investigate what other anthropocentric factors may be important in human–robot social interactions, so that the robot can present itself as more human-like.

Another HRI project Adaptive Strategies for Sustainable Long-Term Social Interaction (ALIZ-E) investigated child-robot companionship in hospitals. This project had the specific goal of supporting children engaged in a residential diabetes-management course (Baroni et al., 2014). Their study found out that the robot could play an important role in healthcare communication in terms of supporting and educating diabetic children (Nalin, Baroni & Sanna, 2012; Blanson-Henkemans et al., 2012; Hiolle, Cañamero, Ross & Bard, 2012). Baroni et al. (2014) used an Aldebarn Nao robot which communicated with children by making conversation and playing games. This example shows that it is possible for robots to educate and help children by playing games and acting as their friends. The researchers in that project, however, pointed out a memory issue with the robot: 'the robot required the ability to support long-term interactions over multiple distinct episodes' (Baxter et al., 2011). In a long-term companionship fields, as in this case, a robot should be able to recall previous interactions with its users. However, it is also true that forgetting previous information and misattributing past events are also pervasive human behaviours (Schacter, Guerin & Jacques, 2011; Schacter & Dodson, 2001; Payne, Cheng, Govorun & Stewart, 2005); yet, people usually accepts these flaws in others and form relationships regardless. In the current, I develop such behaviours in robots analyse their impacts in human–robot interactions.

Human-like skills and behaviours in robots can help to reduce uncanny valley effects. When a humanoid robot closely resembles to human but not that convincing realistic then people feels uneasy to interact with it and this phenomenon is called uncanny valley (Zlotowski, Proudfoot, Yogeeswaran & Bartneck, 2015). Walters et al. (2008) investigated the typical links between human personalities and attributed robot personalities and how these links can be used in reducing uncanny valley. Their research showed that participants tended to prefer robots with more human-like appearances and attributes. They also found out that individual differences in the appearance of the robot were not consistent for all participants. In their experiments, introverted participants with less emotional stability tended to prefer mechanical appearances more than other participants did. This research shows that human-like attributes may be used in robots effectively to reduce uncanny valley effects. Therefore, other human-like attributes such as cognitive biases may also be used in human–robot interactions and may help to reduce such effects.

From the above discussion, it can be said that there are various situations in which human–robot interactions are possible, so it is also possible to develop companionship with social robots based on their skills and attributes. The examples of studies showed that emotions, traits and other anthropomorphic behaviours in robots make robots more appealing and improve their relations with people. The computer program ELIZA, for instance, showed an ‘interest’ in knowing participant’s problems; the receptionist Valerie expressed emotions, showed anthropomorphic behaviours and helped in collecting information for the tourists; AIBO showed adapted behaviours from dogs and played with its users accordingly. These examples show how the robots use different known and common behaviours (e.g. AIBO jumping around the user, or Valerie smiling while talking based on human-like attributes such as moods, emotions, and traits that users could easily recognise). Therefore, it can be said that expressive emotions, personality traits and anthropomorphic behaviours in robots are being used to make them more cognitive in interactions, so that the users can recognise and relate to them more easily (Duffy, 2003; Gockley et al., 2005; Lee, 2006). From the previous examples of studies in Chapter 2, we have seen that researchers studied various interactions methods where the social robots anthropomorphized and mimicked human actions, expressed emotions, used speech, made sounds and performed other human-like cognitive tasks. From this discussion, it can be concluded that human-like different cognitive behaviours are important to have in social robots in order to make HRI more familiar to people.

In this thesis, the aim is to make robot behaviours more familiar to users by creating a relatively new method of interaction which has not before been studied in detail. I propose to introduce the application of cognitive biases in the interaction process between a robot and human. To this point, when researchers have discussed robots with natural human-like behaviours, they have tended to ignore one of the most common and natural human 'behaviours': biased behaviours such as forgetting, misattributing, inability to understand other's emotions, bragging, blaming and others. Everyone makes mistakes, forgets things and shows biased behaviours; this is human nature. Social roboticists study traits, emotions and other human-like skills in robots, but the common attribute of biased behaviour has never been implemented in social robots. The work presented in this thesis addressed this important lapse. Cognitive biases are expected to make robots' behaviours more familiar to humans, helping to improve the human–robot interaction.

3.2 Why consider cognitive biases?

I aim to develop different cognitive biases in robots for human–robot interactions. To explain why I chose to develop cognitive bias in robots, this section explains the effects of cognitive biases have in humans. I discuss the importance of cognitive biases and explain their application to the research presented in this thesis, identifying their importance in developing more 'acceptable' social robots.

A cognitive bias is a type of error in thinking that occurs when people are processing and interpreting information in the world around them. Amos Tversky and Daniel Kahneman (1974) introduced the term 'cognitive bias' to describe people's systematic but purportedly flawed patterns of responses to judgement and decision problems. They were inspired by Herbert Simon in the late 1950s (Kahneman, 2002). According to these two scientists, human judgements and rational choices depend on the decision makers and the factors associated with them. Tversky & Kahneman (1981) explained such features of decision making:

> A decision problem is defined by the acts or options among which one must choose, the possible outcomes or consequences of these acts, and the contingencies or

> conditional probabilities that relate outcomes to acts. We use the term “decision frame” to refer to the decision-maker's conception of the acts, outcomes, and contingencies associated with a particular choice. The frame that a decision-maker adopts is controlled partly by the formulation of the problem and partly by the norms, habits, and personal characteristics of the decision-maker. (pp. 453)

The above statement suggests that the decisions people make can be partly influenced by their habits, characteristics and social norms and therefore such decisions are biased. Cognitive biases are often a result of an attempt to simplify information processing, which can help one to make sense of the world and reach decisions with relative speed (Bless et al., 2004). As such, people create cognitive ‘short-cuts’ to process information faster and with less effort. Such approaches to information processing are known as schemas and heuristics (Judea, 1983). Heuristics help to simplify the social worlds and, as such, can have a detrimental effect on an individual’s thinking and consequent behaviour (Ippoliti, 2015). Such behaviours are common and can be seen in others regularly, but the effects of such behaviours in social robots for human–robot interactions have yet not been explored. I have therefore aimed to study such cognitive biases and their effects in humans and develop some of the biases (e.g. misattribution, empathy gap, etc.) in robots for human–robot interactions.

Cognitive biases can arise from various processes such as social influence (Wang, Simons & Brédart, 2001), information-processing shortcuts, distortions in the mental process which is result of ‘storing and subsequently retrieving objective evidence in memory’ (Hilbert, 2012, pp. 6), limited brain capacity for information processing (Simon, 1955; Marois & Ivanov, 2005), and emotional and moral motivation (Pfister & Böhm, 2008).

In sum, cognitive biases may lead individuals to behave irrationally. The above research suggests that individuals create their own social realities from their perceptions of inputs, but that construction of the real world is not objective, and therefore such subjective states of thinking can lead to incorrect understandings of reality, inaccurate judgments and illogical interpretations. For example, a student receives an A+ on an essay and attributes it to his or her hard work and skill (this is called *internal attribution*). Next week, when he or she receives a C- on an essay and blames that grade on the teacher, claiming that the teacher did not explain the topic well enough (this is called *external attribution*). In reality however, it could be the student who might have rushed through the assignment. This tendency in people is called self-serving bias effects (Sedikides, Campbell, Reeder & Elliot, 2000). Self-serving bias is common among students, employers and in relationships. The research I undertake here develops such biases in robots for human–robot interactions and studies their effects in interactions.

Biases, however, can sometimes lead to a more effective set of actions, yielding positive outcomes (Gigerenzer & Goldstein, 1996). For example, framing effects, where people react to a choice depending on how it is presented; for example, a student feels more incentive to register for class when threatened with a late fee than she feels enticed by an early-bird registration discount. It can be said that the biases help us to reach decisions more quickly (e.g. heuristics biases, which are different strategies derived from experience with similar problems, Tversky & Kahneman, 1974). This efficiency can be vital to processing information more efficiently in certain situations, such dangerous or threatening ones, where making a quick decision is more important than careful consideration. Biases such as humour effects (e.g. people tends to remember humours events more than non-humours) can also help people to remember information (Martin et al., 2003). Such a bias may be useful in social robots, for example, if the robot adds humour to its behaviour and conversation, which could help people to remember the interaction for longer, and it is possible that the person would want to interact with the robot again.

Cognitive biases can be defined as mental shortcuts which influence people's thoughts, judgements, and behaviours. Damaging behaviours are sometimes the result of people's erroneous decisions, which are influenced by behavioural biases (Chira I, Adams M and Thornton B, 2016). Behavioural bias is closely related to economics because the investor's thinking and feeling while making major investment decisions can cause behavioural changes (Duxbury, 2015). There are several biases which can be classified as behavioural biases, for example, cognitive dissonance (i.e. a mental state when a person experiences mental discomfort for having inconsistent thoughts and beliefs which are related to their behaviours), conservatism (i.e. holding traditional ideas which oppose the welcoming of new innovation or values), confirmation (i.e. the tendency to look for information which confirms someone's own existing beliefs), representativeness (i.e. the tendency to make decisions based on patterns of events), illusion of control (i.e. the tendency to overestimate one's own ability to control an event), hindsight (i.e. a psychological phenomenon where, after an event, people think they knew the result all along).

Social bias refers to a group of biases which are related to society and how people perceive others in society. Examples of such biases include stereotyping, which is when an individual holds a perception of another individual based on his or her age, gender, location, race, skin colour, etc., but without any factual knowledge (Fiske,2010); the halo effect, which is the tendency to perceive others based on a single trait or their previous positive or negative actions; or system justification, which is a psychological theory that suggests people hold positive views of themselves and the group they belong.

Traits can be defined as qualities or characteristics which produce particular types of behaviour, and traits are typically discussed as belonging to people. Traits can be inherited from one's parents and also shaped by environmental factors and life experiences (Ali I, 2018). For example, eye and skin colours come from parents, while height can be shaped by nutrition from the environment.

In recent years, social biases have shown importance, as they can be seen in many places. Due to the rise of social media and the Internet, people are able to better communicate with others in the world and more easily express their opinions. On popular social media platforms, such as Facebook, it can be seen that people express their opinions on different matters like politics, news, sports, movies, and more, though such opinions are often not based on facts. In such cases, the posting person tries hard to prove his or her opinion based on limited knowledge or incorrect sources, which is a clear instance of Dunning-Kruger effects. Social media and popular search engines display advertisements to users based on their choices and Internet searches, which is an example of using the effects of decision-making biases, such as attentional bias. In this age of information, many people are victims of information bias, where we always seek information regardless of necessity. In our personal lives, we regularly experience various social biases. People tend to prefer themselves over others in matters of success but usually blame failures on others. This phenomenon is known as the self-serving bias, which can affect close relations, workplace dynamics, classrooms, and other situations. In relationships, people tends to favour their relatives, friends, and family members over to strangers (Sedikides, Campbell, Reeder, & Elliot, 1998). People sometimes overestimate others' positive qualities while underestimating their negatives traits, which is known as illusory superiority or the better-than-average effect (Hoorens, 1993). Conversely, people tend to believe themselves to be worse than others at performing difficult tasks, which is known as below-average effects (Kruger, 1999). There are several other biases which affect people's judgements and behaviours in different social situations (a list is provided in Appendix G, page 285). Since individuals' characteristics and traits can be shaped from their environments and life experiences (Ali I, 2018), it is possible that traits can be affected by biases as well.

Environments influence people based on their thinking, and we discussed that human thoughts are usually affected by various biases. Therefore, it can be said that cognitive biases are important for shaping individuals’ characteristics and traits, as well as their behaviours in society.

Many studies have been done on the effects of the biases, and based on how biases arise, affect and motivate humans, scientists categorized those biases. Haselton, Nettle and Andrews (2000) propose a taxonomy of cognitive biases based on how those arise:

1. Heuristic – where the biases are caused by decision-making process, e.g., uses of stereotypes (Haselton, Nettle & Andrews, 2000)
2. Error management biases – when a biased based solution for a problem resulted in less error than the unbiased solution, e.g. sexual over perception by men (Haselton, Nettle & Andrews, 2000)
3. Artifact – when the mind is not prepared for the targeted task, e.g., base-rate neglect in statistical prediction (Haselton, Nettle & Andrews, 2000)

Scientists have suggested that cognitive biases can arise from emotional and moral motivations (Loewenstein, Weber, Hsee, Welch, & Ned, 2001; Pfister & Böhm, 2008), social influences (Wang, Simons, & Bredart, 2001; Yechiam, Druyan, & Ert, 2008), and ‘noisy’ information processing (Hilbart M, 2012). Tversky and Kahneman (1974) produced a list of heuristics and biases on human judgement and decision making. He describes three heuristics related to decision making: representativeness (e.g., insensitivity of predictability, illusion of validity, etc.), availability (e.g., biases of imaginability, illusory correlation, etc.) and adjustment and anchoring (e.g., insufficient adjustment etc.) (Tversky & Kahneman, 1974). Many studies have been done on specific biases in psychology, based on their effects and importance, but in it is still unknown how many biases there are, exactly. In the current thesis, I study only the biases which may be important and relevant in the current study. Here, I list few of the most studied and discussed biases as examples in Table 3.1. I provide another exhaustive list of cognitive biases in the Appendix G.

Table 3.1: Commonly studied cognitive biases

Name	Description
Empathy gap	The tendency to underestimate the influence or strength of feelings, attitudes, preferences, and behaviours, in either oneself or others (Van Boven et al., 2013)
Misattribution	When individual fails to recall source details retained in memory and erroneously attributes a recollection or idea to the wrong source (Schacter & Dodson, 2001)
Self-serving bias	The tendency to claim more responsibility for successes than failures; also manifests as a tendency among people to evaluate ambiguous information in a way that is beneficial to their interests (Forsyth, 2008)
The lack of knowledge/ ideocracy effects	The tendency for unskilled people to overestimate their own ability and experts to underestimate theirs (Kruger & Dunning, 1999)
Framing	Using a too-narrow approach and description of the situation or issue
Humours effects	The tendency to remember humorous items more easily than non-humorous ones (Schmidt, 2002)
Hindsight bias	Sometimes called the ‘I-knew-it-all-along’ effect, the inclination to see past events as predictable (Roediger & McDermott, 1995)

Some cognitive biases can influence interpersonal relationships, such as in-group bias. People favour others in their groups more than people other groups. This bias helps illuminate the origins of prejudice and discrimination (Aronson, Wilson & Akert, 2010). Attribution bias, according to Heider and Simmel (1944), represents the systematic errors made by individuals when they try to evaluate their personal behaviour. For example, if a mother gives instructions to some children, the children might think that mother considers them stupid and feel attacked, becoming irritated in consequence. Alternatively, if the children think that the mother is trying to be helpful, then they feel caring and appreciative. The empathy gap can influence relationships, for example, if someone is happy, that person may have difficulty imagining why people would be unhappy, creating conflicts in relationships. Understanding such biases could be useful in human–robot interactions, so it would be interesting to find out how people react to such biases (e.g. empathy gap) in interactions with a robot which is unable to understand people's emotions. Thus in the Chapter 4, I develop the empathy gap and other related biases in robots to study their effects in human–robot interactions.

To understand the effect cognitive bias plays on an individual, the famous example of the 'bandwagon effect' can be discussed (Goidel & Todd, 1994). This bias is closely related to the herd mentality, which describes how people are influenced by their peers to adopt certain behaviours, follow trends and purchase items. Individuals experience the bandwagon effect in that they place much greater value on decisions that are likely to conform to current trends or please individuals within their existing (or desired) peer group. This effect may be interesting to study in social robots when the robot and user parsing same or opposite product of interest, and in making robots that 'conform to society or peers'. From a consumer's perspective, this effect can be summarised as making 'purchasing decisions' out of a desire to have and be seen with 'the next big thing', or to increase perceived social status by owning a particular product. Such bias effects are so common because of people's deep-seated need to conform and fit in. Many consumers' devotion to Apple products is a great example of the bandwagon effect in action.

From discussions so far, it can be said that cognitive biases exist in humans and have major influences on decision making and behavioural actions. Biases refer to ways of thinking or thought processes that produce errors in judgment or decision making, or at least depart from normative standards (Hahn & Harris, 2014). Biases however, can lead to good decisions in many situations, and it is possible that in some situations they can lead to better decisions than the regular approaches (Gigerenzer & Goldstein, 1996). Some biased forms of reasoning (e.g. belief bias) are often assumed to happen quickly and automatically, without working memory resources and independent of contemplative activities (Evans & Stanovich, 2013). It seems reasonable to agree that biases are indeed an important factor in developing an individual's character and are partially responsible for one's behaviours, as they affect thinking and reasoning (Stanovich & West, 1998).

3.3 Why consider cognitive biases for social robots?

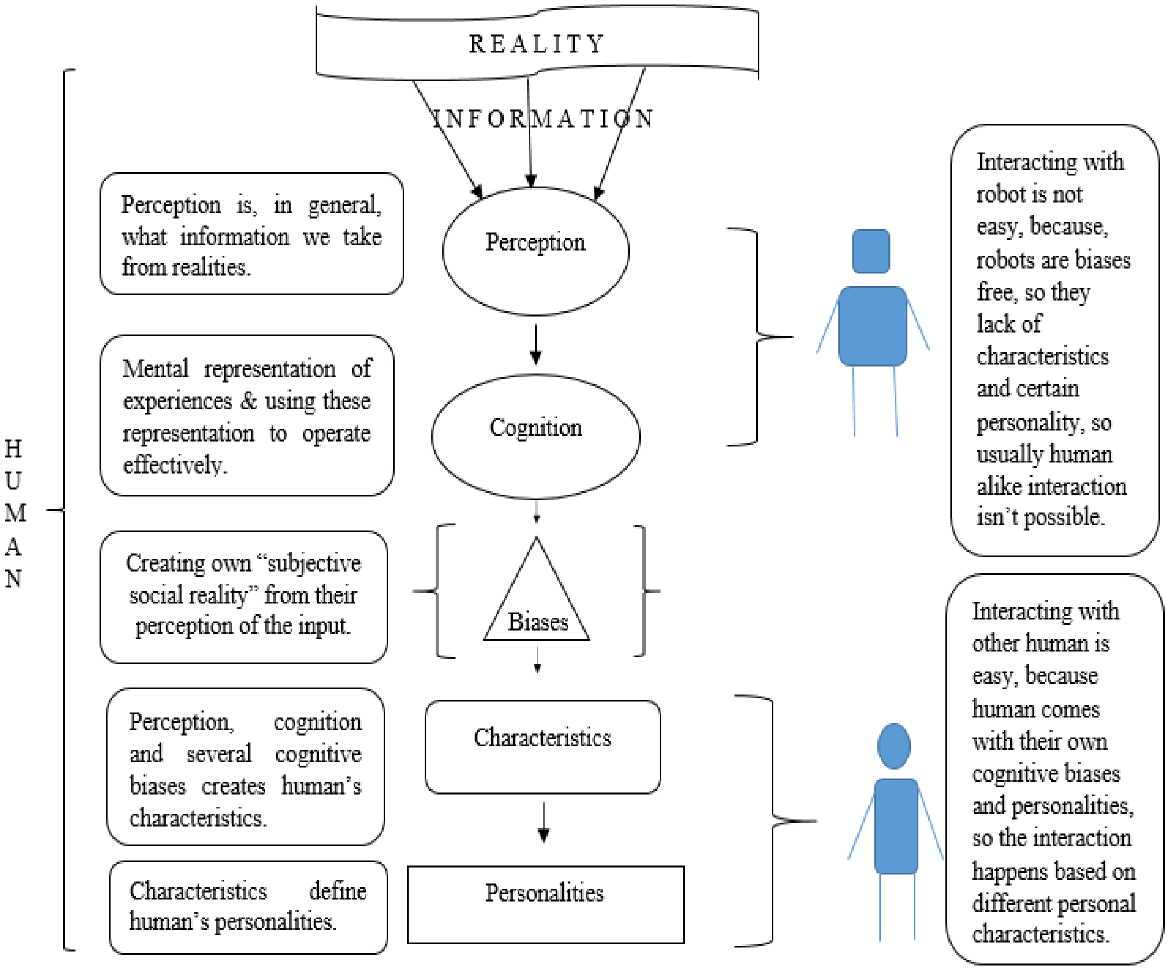

REALITY
INFORMATION
HUMAN
Perception is, in general, what information we take from realities.
Perception
Mental representation of experiences & using these representation to operate effectively.
Cognition
Creating own "subjective social reality" from their perception of the input.
Biases
Perception, cognition and several cognitive biases creates human's characteristics.
Characteristics
Characteristics define human's personalities.
Personalities
Interacting with robot is not easy, because, robots are biases free, so they lack of characteristics and certain personality, so usually human alike interaction isn't possible.
Interacting with other human is easy, because human comes with their own cognitive biases and personalities, so the interaction happens based on different personal characteristics.

Figure:3.1 Comparative diagram of the differences between people and robots, how humans process information and robots lack the ability to represent certain types of personality characteristics, which might be key to interactions between the humans and robots

3.4 Biases are common in people

Figure 3.1 shows the general differences of perceptions between people and robots, suggesting that humans perceive and process information from their surroundings using all attributes (e.g. traits, emotions, moods, and biases) but although social robots can be programmed to show traits and emotions, they sometimes lack to show human-like behavioural biases (e.g. forgetfulness, blaming others etc.). It can be agreed that robots are far from perfect, and the scientists always research to update the robot so that it could help human in many ways. The study described in this thesis does not consider the mechanics or the control systems of the robot, but discusses the sociable behaviours of a social robot. Terms such as 'perfect', 'imperfectness' are used to compare and differentiate between two set of behaviours in this thesis. This is notable that the cognitive biases do not make the human behaviours perfect and I do not claim that such behaviours can make the robot's social behaviours perfect, but this study investigates the effects of such biased behaviours in human-robot interactions.

We can agree on that humans have characteristics flaws, biases in thinking (e.g. mistake-making, forgetfulness, miscalculation, the empathy gap, self-serving bias, etc.). Such behaviours generally do not become the barrier between humans in making relationships. It is possible for individuals to relate to others in natural ways. Sometimes people forget information, or people attribute certain outcomes to the wrong causes, brag and pass blame on to others. It can be said that people are not uniform or perfect, yet people can make friendships with others, although they differ. It is usual for people to find all sorts of different imperfect behaviours in themselves and in the people around them. But these sorts of behaviours are not found in the social robots yet, it is very unusual that a robot would forget the information or blame people if lost in game. But forgetting information, blaming others are very common human behaviours. So the current research seeks answers to the question that if it is possible to develop such imperfect behaviours in the social robots, similar to those of humans, where a robot may forget information, misattribute certain outcomes, or brag, certain questions arise: How would the user perceive the robot? What influences would such behaviours in robots have in human–robot interactions? Will people like and accept the robot or reject it? It is common for people to notice all sorts of different imperfect behaviours in themselves and in the people around them. However, these sorts of behaviours are not found in social robots yet, so it is very unusual that a robot would forget information or blame people if it lost in a game; forgetting information and blaming others are very common human behaviours. Flaws and imperfections are characteristics of people, and even with flaws, people can relate to one another as socially accepted beings. Therefore, such flaws and imperfections should be investigated in social robot-human interactions to determine whether their effects can help people to better relate with the robot.

Sometimes, it is hard to relate to someone who is almost flawless and acts without mistakes. For example, the 'Mary Sue' characters created by Paula Smith in 1973 in 'A Trekkie's Tale' were illustrated as flawless (Smith, 1973). They are attractive and smart, and something convenient always happens to their advantage; in the end, it is almost a given that they will get what they want. Audiences grow tiresome reading about these flawless characters, which seem unreal (Verba, 2003). Audiences would rather hear about what happens to realistic characters and how they overcome their adversities; otherwise, characters like Mary Sue seem very mechanical, unrealistic and tough to relate to. This example is associated with the subject of the current thesis: social robots that do not exhibit human-like flaws (e.g. forgetfulness, blaming, bragging, etc.) in interactions seem very mechanical and distant, and the expectation is that if cognitive biases can be introduced in the robot, then the interactions between robot and human could be more enjoyable and easily relatable.

3.4.1 The concept of lacking 'bias' and 'imperfect' behaviours in social robot

As discussed in Chapter 2, the current social robots are able to mimic human actions, but their actions might be limited and may not provide the behavioural realism present in humans, thereby preventing people from creating and maintaining social relationships with them. In Chapter 1, the following question was raised: 'What happens in human–human interactions that is absent in human–robot interactions, thus preventing a social relationship forming between the human and the robot?' At this point, it can be said that social robots usually lack such behavioural realism in human–robot interactions (e.g. mistake making, forgetfulness, blaming, bragging etc.). Such variation in the cognitive nature of individuals might make human–robot interactions unique, natural and human-like. It can be said that the shortcomings of humans in their characteristics, including faults, mistakes, and other common human-like characteristics are needed, whereas robots are too mechanical and artificial. In existing social robotics, the robots imitate human social queues, like eye contact, human-like walk, talk and body movements, and so forth. The previously discussed human traits (e.g. faults, unintentional mistakes, task imperfectness, etc.) are not intentionally present in these current social robots, so this thesis is focusses on different common cognitive biases (e.g. misattribution, empathic gap, the lack of knowledge or ideocracy effects, etc.), which can influence human thoughts in an illogical fashion.

Hayoun (2014) raised an important question, 'How can we interact with something "more perfect" than we are?' By acknowledging that robots are far from perfect in terms of development, functionalities, control systems and comparability to any living being, the current thesis addresses the issue of the lack of human-like cognitively biased behaviours in social robot behaviours in human-robot interactions. As has already been shown, faults and misjudgements are common human characteristics which create openings for new conversations or interactions based on particular individuals and their situations. As discussed earlier, research has shown that human-like trait-based personalities and emotive expressions can make human-robot interactions more enjoyable for humans (Lee et al., 2006; Walters, 2008). Similar to such characteristic traits and personalities, cognitive biases are important components in human characters, and biases are widely present in all humans. Developing cognitive biases in robots may allow robots to make common mistakes, like humans do, which is an idea that has not been properly tested yet. If the robot exhibits fatigue and stress after repeated tasks and complains to its owner, what will the owners do? Will they get annoyed and replace the robot, or accept it and console it in a human-like manner? Either way, this behaviour gives the user ideas to ponder over, which may lead to a certain type of relationship. Therefore, the notion of how such cognitive bias behaviours can affect human-robot interactions needs to be investigated.

Research (Lee et al., 2006) shows that developing personalities and trait attributes in robots can make them more acceptable to humans. Also, expressing emotions and moods in interactions can help to make the attachments stronger between users and robots (Dautenhahn et al., 2009). The reason could be human-like personality traits; emotional expressions help people to anthropomorphise the robot, and as a result, people can relate with the robot. Similarly, cognitive bias behaviours like forgetfulness, bragging, blaming, and telling jokes in conversation are common human characteristics which could help people to anthropomorphise the robot even more.

As discussed in the sections 3.1, 3.2, cognitive biases effect on human thinking processes which leads to making mistakes, misjudgements and fallible activities (Baron, 2008). Such misjudgements, mistakes and fallible activities could be important factors in creating individual's social behaviours. McKay and Dennett (2009) have shown that 'misbelief' could be adaptive in human behaviours and may be expressed by the people daily. The current research investigates such cognitive biases in social robots in the context of human–robot interactions, studying the responses from the participants. It is expected that cognitive biases will make robots' behaviours more anthropocentrically fallible, and thus more recognisable to humans. These familiarly biased behaviours presented by robots during the interaction could help humans to understand and relate to robots. The experiments investigating this possibility are detailed in Chapter 4.

Haselton et al. (2005) has suggested that humans construct objective images of social realities as they perceive them. The human brain is both remarkable and powerful, but subject to limitations (Kanwisher, 2010). The human mind translates the properties from the world and draws its own conclusions in an illogical manner. Haselton et al. (2005) describes such fundamental limitations onto human thinking as cognitive bias. Cognitive biases represent an individual’s own sense of reality, not the objective input, which sometimes lead to perceptual distortion, inaccurate judgment and illogical interpretation (Tversky & Kahneman 1974). When people make judgments and decisions about the world around them, they like to think that they are objective, logical and able to evaluate all the available information (Gutnik et al., 2006). The reality is, however, that an individual’s judgments and decisions are often riddled with errors and influenced by a wide variety of biases (Tversky & Kahneman 1974; Haselton et al., 2005). This point is articulated by Ariely (2008):

> We usually think of ourselves as sitting in the driver's seat, with ultimate control over the decisions we made and the direction our life takes; but, alas, this perception has more to do with our desires—with how we want to view ourselves—than with reality. (pp. 243)

According to Ariely (2008), cognitive biases can affect a range of thoughts in the human brain, such as its effects on thinking to about how to handle money, how to relate to other people, to even how memories of the past are formed.

Many studies have proven that cognitive biases have considerable effects in the fields of investments, business and marketing. The cognitive bias of ‘loss aversion’, for example, makes people strongly prefer to avoid losses over acquiring gains, which sometimes causes people to make the wrong economic decisions (Tversky & Kahneman, 1992). Another bias, ‘hindsight bias’, occurs when an individual gives a much higher estimation of the predictability of a newly learned outcome than the individual who can predict the outcome without advanced knowledge (Pohl, 2007; Yudkowsky, 2008). Because of the availability of various technologies and mediums for advertisement, studies of cognitive biases are now an essential part of business and marketing (Bolton et al., 2013).

Cognitive biases, such as misattribution, can affect the ability to recall past events from memory. As the result of misattribution of memory bias, people can recall past events but usually forget and state the wrong source of that event. Schacter and Dodson (2001) identified three kinds of misattributions which could happen in people: source confusion, cryptomnesia and false recognition. Misattribution is very common in people, and I discuss this bias in much detail in Section 3.5.1. It is common in society that people forget names and past events or unintentionally make mistakes influenced by misattribution bias. In the current thesis, I investigate the forgetfulness effects of this bias in the robot such that the robot could not recall previous interactions with the participants.

In many cases, cognitive biases such as 'self-serving bias' allow people to protect their self-esteem (Crocker & Major, 1989); for example, people sometimes get a boost in confidence by attributing positive events to personal characteristics, and they protect their self-esteem and absolve themselves from personal responsibility by blaming outside forces for failures. Biases like the 'halo effect' influence the human mind to create an overall impression of another person, depending on how that individual feels and thinks about that person. Standing (2004) defines the halo effect as follows: 'Also known as the physical attractiveness stereotype and the "what is beautiful is good" principle, the halo effect, at the most specific level, refers to the habitual tendency of people to rate attractive individuals more favourably for their personality traits or characteristics than those who are less attractive. Halo effect is also used in a more general sense to describe the global impact of likeable personality, or some specific desirable trait, in creating biased judgments of the target person on any dimension. Thus, feelings generally overcome cognitions when we appraise others.' At this point, it can be said that biases are common in all humans and, knowingly or unknowingly, we are expressing them in many ways.

3.4.2 Biases and imperfectness in modern technologies

Even in the development of new technologies, people find different biases; for example, Siri, the personal assistant software developed by Apple (n.d.), was accused of having a bias called 'abortion bias'. In a website blogpost of Abortioneers (abortioneers.blogspot.co.uk 2011), many users found such bias while interacting with Siri, and they presented various phone screenshots and videos as proof. This news was reported in many online medias including The Guardian (Rushe D, 2011).

This bias was described as Siri could not find anything when they asked it to direct them to an abortion clinic, despite the facts that abortion is legal in the US and UK, but the program has no problem finding pregnancy crisis centres, 'which often are run by churches and right-wing groups that are very clearly against abortion', as one user complained (Wilson, 2011). According to BBC (BBC, 2011), Apple denied such accusation and issued an apology statement: 'These are not intentional omissions meant to offend anyone, it simply means that as we bring Siri from beta to a final product, we find places where we can do better and we will in the coming weeks'. In our recent technologies, most of the voice-command assistants can be considered biased; for instance, Amazon's Alexa comes with only a female voice option. Even Siri was initially available only with a female voice, and coincidently, the name Siri means 'beautiful woman who leads you to victory' in Norse (The Week, 2012). Similar to the voice assistant programs, the computer systems are not free from the biases as well. Friedman & Nissenbaum (1996) studied 17 computer systems which are from different fields such as education, banking, science, medicine, law and others and categorized biases as technical, pre-existing and emergent. They found out that the pre-existing biases come from society, practices and attitudes; technical biases arise from technical consideration and emergent biases happen when the systems are used by the real users (Friedman & Nissenbaum, 1996). They concluded that it is difficult to minimise the biases in practice in the computer design but not impossible (Friedman & Nissenbaum, 1996). It is possible that in near future in the case of technological personifications, people will want their technologies to be based on specific traits and behaviours, which might introduce certain biases in the technological devices people desire. For example, Google and Facebook extract individuals' information in the form of preferences to create user profiles; based on such profiles, selected advertisements appear (Esteve, 2017). In my understanding, the choices to show such advertisements can be called bias because the choices and information the individual put on the pages might not be accurate, and overall an individual might not like those ads at all.

Modern personal voice assistants such as Siri, Microsoft's Cortana, Amazon's Alexa can be more advanced than the early automated computer chat program Eliza, which was created by Weizenbaum (1976), but these voice assistants are still considered as weak artificial intelligence (AI). Weak AI is focused on a specific task without having any consciousness, skill or learning ability, and it is performed based on how it has been programmed. For example, if I ask Siri to find nearby restaurants, Siri can respond by searching the Internet using my Wi-Fi location and is able to list a range of places using the pre-programmed algorithm in the software. Most modern technologies that include voice assistants, computers, robots and automated machines are considered as having weak AI.

Strong AI, on the other hand, is when a machine can understand and be conscious like brain. Hypothetically, a machine with strong AI can think, learn and solve problem like humans. For example, a personal robot that can reason with its user, understand problems and its user's mental state, and knows what to suggest – just like having a close companion –

exemplifies strong AI. Strong AI still exists only in movies and science fiction, such as A.I. Artificial Intelligence directed by Steven Spielberg (2001) or Ex Machina directed by Alex Garland (2015). Searle (1980) argued about strong AI in machine in his famous the 'Chinese Room Argument', in which he showed that AI systems only can act as if it has a mind or consciousness without knowing that it originally does not.

There are concerns regarding both strong and weak AIs. The former CEO of Microsoft, Bill Gates, wrote in a Reddit 'Ask Me Anything' interview (2015) in response to a question regarding the threats of 'machine superintelligence':

> I am in the camp that is concerned about super intelligence. First the machines will do a lot of jobs for us and not be super intelligent. That should be positive if we manage it well. A few decades after that though the intelligence is strong enough to be a concern. I agree with Elon Musk and some others on this and don't understand why some people are not concerned. (Reddit)

Such concerns exist not only for strong AI, as there are many concern regarding weak AI as well in recent years and people express their concern in online articles. In an article called, "How much Longer before our first catastrophe?" in iO9, the writer George Dvorsky (2013), concerned about the growing uses of weak AI in society, mentioned, "Our infrastructure is becoming increasingly digital and interconnected – and by consequence, increasingly vulnerable". He further mentioned that weak or narrow AI could damage electric grids, nuclear power plants and even cause economic crises on the global level. The automated voices of the customer care services in banks and other organisations, spy cameras, computer visions, major websites' algorithms (YouTube, Facebook etc.) and many other automated programs assess different situations and process the data to choose where the reward is maximised (Saenz, 2010). In our everyday life, when someone call to a bank, phone company or any other customer care service, the customer usually talks to an automated voice service which suggests a set of instructions to the caller. These voice services are usually very insentient, unconscious and lifeless – where the current research might prove useful. If it is possible to make such automated voices more empathic, emotionally precise and tonally accurate, then people may find those services more helpful and feel less concerned about the systems.

Studies have found various biases and errors in Satellite Navigation systems as well. Sardon, Rius, and Zarraoa (1999) showed that there is stable receiver bias in global positioning systems (GPSs) which causes delay between two frequencies. Cornely (2013) indicates that such a delay could be introduced either by the sender or the receiver, but it should be removed in order to resolve the total electron content (TEC) accurately. In technologically advanced satellite navigation systems, there are other errors and biases; for example, the GPS ephemeris error, selective availability, satellite and receiver clock errors, the multipath error, antenna-phase-centre variation, receiver measurement noise, ionospheric delay and others (Tiwari, Arora & Kumar, 2010). As the current experiment focusses on people's cognitive biases, I go into no further details about such errors of the satellite navigation system, but it can be understood that even such modern technology is not free from errors and bias.

However, both careful and careless people show their cognitive biases in interactions. Individuals interact with their own cognitively imperfect characteristics, presented by maintaining social norms and following social cues. Common social norms and humans' generally imperfect behaviours make the interactions natural. Society is an example of each person's behaviour differing, and each having unique characteristics. Myriad human characteristics and behaviours are found within any society, and sometimes they result in the formation of relationships. My research suggests these cognitive biases may have a role to play in social robots, examining how they can affect human–robot interaction, for better or worse. It is possible that the robot would get its very own individual set of behaviours with human-like faults and mistakes which would be familiar to humans, and that familiarity might help humans to relate to it or to reject it. It is expected that cognitive biases will make the robots cognitively imperfect, which would allow the robots to make mistakes, like humans do. Using specific cognitive biases in social and companion robots could therefore prove very useful in social interactions between robots and humans.

3.5 Hypothesis and research questions

As it is evident that various biases can affect humans in different ways, such as in decision making, social interaction, memory, brain function, emotional response and so on, it is reasonable to find out the effects of such biases in social robots. Social robot–human interactions are actually based on the human–human interactions discussed in Chapter 2. From the discussions in this chapter, it is concluded that various cognitive biases have a reasonable level of influence on humans, and such biases might be proven useful in robots for human–robot interaction. Therefore, I propose the following hypothesis to examine this matter:

If a robot can exhibit behaviours that include selected cognitive biases when interacting with humans, then humans will have increased fondness for the robot.

The current research seeks answers to the following questions:

- Despite a robot's appearance, functions and features, can it interact with humans in ways similar to human-human interaction by engaging with humans in both the long term and the short term and drawing on cognitively biased behaviours?
- With cognitive biases introduced in a robot, is it possible to develop anthropomorphically biased behaviours in robots to influence human-robot interactions?
- Will cognitive biases help humans to develop human-robot interactions in the long term?

To study the hypothesis and seek answers to the research questions, two sets of behaviours were developed for robots: non-biased or baseline behaviours and biased behaviours based on selected cognitive biases. As the main aim of this study is to determine the effects of cognitive biases in robots' behaviours in human-robot interactions, comparing these two sets of behaviours seems reasonable in order to distinguish the effects of the biases.

In the current study, robots interacted with the participants in conversational and game-playing interactions, where the robot showed developed biased or non-biased behaviours throughout the interactions. To exhibit each of the selected biased and non-biased behaviours effectively so that participants can recognise and understand the differences between the two behaviours, the interactions were divided into several stages according to the mode of the interaction: that is, conversational interactions over three stages (meet and greet, topic-based discussion and farewell), and interactions in game playing over five stages (meet and greet, rules, gameplay, results and farewell). In the interactions, biased and non-biased behaviours were shown in each of the stages through speech and actions. The specific behaviours of the robots were chosen after several mock experiments and, after consulting with my supervisor, I ensured sure that the targeted biased behaviours were positively shown by using speech and actions from the robot which could be easily understood by people. I discuss this topic further in Chapter 4.

In terms of data collection, I considered measuring the participants' preferences of interactions with the robot, where the robot's behaviours were developed based on the biased and non-biased algorithms. Mainly because I compare two sets of interactions (i.e. biased and non-biased) to find out which the participants prefer, data were collected based on mainly four factors:

- how much the participants like the interaction with the robot (i.e. likeability);
- whether the participants were comfortable during the biased interactions (i.e. comfort);
- whether the participants were pleased during interactions with the robot (i.e. pleasure); and
- The developed rapport between the participant and robot (i.e. rapport).

The four dimensions above are reasonable, as the experiments are based on comparisons between two sets of behaviours (e.g. biased and non-biased) and discovering participants' preferred versions of the interactions; in understanding such preferences, the four above factors are the most crucial. It is further possible to know participants' preferred interactions at a certain time and over time based on these four such factors. I collected data in the form of rating-based questionnaires. In the next chapter, I detail the experimental methods, questionnaires and algorithms.

3.6 The cognitive biases used in the research

In the previous section, the importance of using cognitive biases in social robots was discussed. I also formed a hypothesis interrogate biases and imperfectness in robot behaviours and determine their effects in human–robot interactions. To explore the hypothesis, five cognitive biases were chosen to develop on three different robots for use in human–robot interactions in short-term scenarios. This section describes the five cognitive biases that have been used in our experiments.

As I have mentioned previously, to this point, more than 200 cognitive biases have been listed, each with different effects on humans in different situations. Some lists of biases are offered on various websites (see appendix G). However, in my research, I introduce the novel idea of developing the effects of cognitive biases in robots for human–robot interactions. Based on their effects on human's, the chosen cognitive biases are categorised as follows:

1. Decision-making, belief, and behavioural biases;
2. Social biases; and
3. Memory error and biases.

For my experiment, I chose from each category to test in human-robot interactions to find out people's reactions to such different bias groups when developed in a robot. For example, from the memory error and biases group, the misattribution and humour effects were selected; the self-serving effect was chosen from the social bias group, and from the decision-making groups, the empathy gap and idiocrasy (influenced by Dunning-Kruger bias) were selected.

In terms of selecting the biases from each group, I firstly planned to reject any that provoked racial hatred, violence or other harmful biases to consider in the experiments such as implicit social cognition or aversive racism. Then, the biases were selected mainly based on their importance, frequency, impact of the effects, understandability and suitability in the experiments. The detailed selection criteria are given below:

- Importance – how important is that bias in the group. 'Importance' here suggests both commonness and impact of the bias on people.
- Frequency – how frequently such a bias can be noticed in people.
- Impact of effects – how such a bias can influence people and how often such influences can be noticed.
- Understandability – How easily such bias effects can be noticed and understood by others, and also how long it takes to understand such effects. Some of the bias effects could take time to realise (e.g. availability cascade, backfire effect, childhood amnesia), but in the experiment, I wanted to show bias effects quickly so that the effects could be easily recognised.

- How a certain bias fits in the experiment – this point can be divided into several other issues, for example,
    - The bias can be easily developed in the robot's behaviours,
    - The bias must be suitable for conversational and game-playing interactions,
    - The bias's effects can be easily and quickly spotted,
    - The bias is not so offensive so that participant do not feel comfortable during interactions,
    - The bias must be experienced by individuals on a daily basis so that effects are known to participants, and
    - The effects of such a bias can be measured using a questionnaire.

The cognitive biases such as misattribution, empathy gap, lack of knowledge or ideocracy (based on Dunning-Kruger effects), self-serving and humour effects all satisfy the above criteria and thus were selected for the experiments. There are other biases which satisfy such criteria and have been researched extensively in psychology, for example, actor-observer bias, halo effect, group attribution and others. However, to implement these in human-robot interactions would require careful development and more complex methodologies to test rather than a comparison between two behaviours. Those biases are recommended to be researched in the future, but for this first experiment with cognitive biases, I chose the five aforementioned cognitive biases to compare with non-biased interactions to uncover the participants' reactions.

3.6.1 Misattribution bias

Misattribution bias arises when someone incorrectly attributes an effect to something with which that person has no connection or association. Misattributions can be divided into two categories: misattribution of arousal and misattribution of memory.

Misattribution of arousal is psychological situation in which people make a mistake in assuming what is causing them to feel aroused (White, Fishbein & Rutsein 1981). In my study, however, the focus is on the misattribution of memory, which involves the details retained in memory but attributed to the wrong source (Schacter & Dodson, 2001). For example, person

A tells a joke to person B, but person B tells person A that person B was the one who told that joke to A in the first place. Schacter and Dodson (2001) have described misattribution as a type of memory distortion or inaccuracy. Misattributions can be a daily occurrence in people where they forget meeting time, names and other common information. Some examples of misattribution that have been studied are as follows:

a. *Misattributing the source of memories:* This is commonly happens when a person says something they read in newspaper, but in reality it was a friend who told them or they saw it in an advert or another source. Schacter (1999) has shown that participants with 'normal' memories regularly make similar mistake of thinking they acquired a trivial fact from newspaper, when in reality the experimenters had actually supplied it. This type of misattribution is very common and fairly important at this time and also can be tested in human–robot interactions.

b. *Misattributing a face to the wrong context:* Studies have shown that memories can become blended together, so that faces and circumstances are merged (Harris et al., 2008; Galton, 1879). For example, people may claim that they saw a face in one context when they actually encountered it in another. In the experiments in this thesis, this misattribution was tested several times with the participants, where the robot was made to forget the face of the participants it previously interacted with; therefore, in the second interaction with the same participant, the robot confuses the participants face and struggles to remember. I discuss the method further in the next chapter.

c. *Misattributing an imagined event to reality:* Goff and Roediger (1998) demonstrate how easily the memory can transform fantasy into reality. In their experiment, participants were asked either to imagine performing an action or to actually perform the action, for example, to break a toothpick. The participants participated in three sessions for approximately three hours over two weeks period where in the first session they had to listen to 72 action statements (encoding session), in the 2nd session they had to imagine performing action statements from previous session and with some new statements (imagining session) and in the 3rd session the participants had to recognise and monitor the sources for the encoding session; and they were asked whether they had performed that action or just imagined it (Goff and Roediger, 1998). Those who imagined the actions more frequently were more likely to think they had actually performed the actions. This experiment suggests that imagined events could misattributed in memory. Similarly, in the current experiments, the robot states some imaginary events which never took place and tries to convince the participants that they happened. The robot creates different topics based on this type of misattribution; for example, it says ‘I remember when you came last time, you sang for me’, ‘Do you remember last time when we played card games and I won?’ and similar.

d. Unintentional plagiarism: Schacter (1999) identified this type of misattribution, when an individual attributes an idea or memory to themselves that really belongs to someone else. Such misattribution helps to explain why there is so much unintentional repetition across many different areas of human culture, such as between musicians, writers and artists. In the current experiment, the robot often quotes famous lines and claims that it made them it up, and if the participant opposes, then the robot admits the misattribution and apologises for its confusion.

The above examples span the three main components of memory misattribution: cryptomnesia, false memory and source confusion.

Cryptomnesia happens when individuals strongly believe that they are the original source of a thought. Some individuals fail to establish memories with enough detail to generate a source attribution, causing a misattribution of memory to the wrong source (Johnson, Hashtroudi & Lindsay, 1993). People often truly believe that the information they plagiarised was in fact, their own. In this study, such components of misattribution were experimented upon with the robots ERWIN (Emotional Robot with Intelligent Network) and MARC (Multi-Actuated Robotic Companion) (Figure 3.2). Both robots made misattributed arguments about various pieces of information in conversation with participants. These robots were used in the research currently described in this thesis, and the experiments are part of this thesis as well. The experiments are described in the Chapter 4.

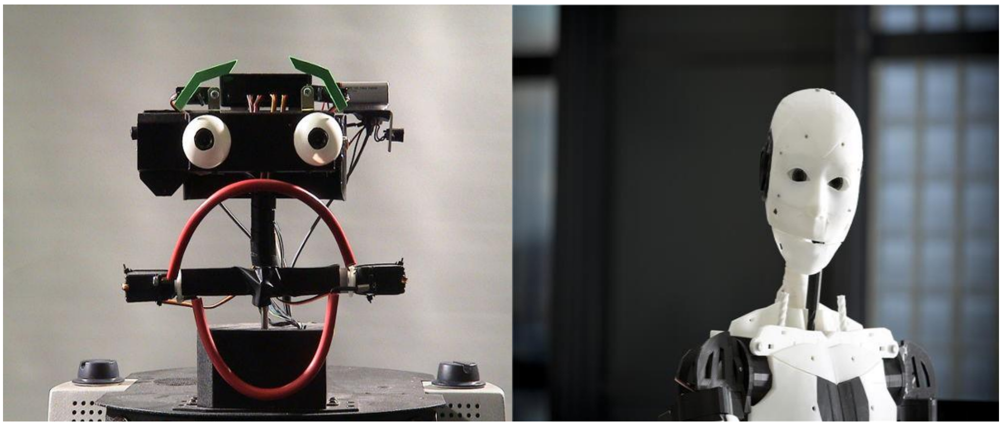


Figure 3.2: From left to right: ERWIN and MARC

False memories occur when a person's identity and interpersonal relationships are strongly centred on a memory of an experience that did not actually take place (McHugh, 2008). Roediger and McDermott (1995) have studied this subject and proved that people could provide false recall without noticing their errors.

Source confusion is another type of misattribution where an individual misattributes or forgets the source of a memory. For example, suppose the parents tell stories about their own childhood incidents to the children at bedtime every night but when the children grow up, they forgot the sources of such stories and thought such story incidents were from their own childhood. Memories arise both from perceptual experiences and from an individual’s thoughts, feelings, inferences and imagination (Henkel & Coffman 2004). Source confusion, however, is also common in people and can be encountered in daily life. In my research, I wanted to find out how people would react if a robot showed this source confusion behaviour in human–robot interactions.

There are many studies on the misattribution of memories which suggests that memories can easily become confused and distorted. As psychologist William James (1984) points out, ’Most people, probably, are in doubt about certain matters ascribed to their past. They may have seen them, may have said them, done them, or they may only have dreamed or imagined they did so.’ The misattribution bias is very common in people, and many studies have been done on this subject. Also, the effects of this bias can be developed in the robot, and it is expected that participants would understand the misattributed behaviours of the robot.

Misattribution is one of the most researched and important biases in the memory bias group. The various effects of this bias can be frequently found in people. People are easily influenced by this bias which often can be noticed by others as well. This bias fits into my experiment as it can be developed easily in a robot in a conversational interaction where the robot can show its forgetfulness to the participants; therefore, I selected this bias.

3.6.2 Empathy gap

The empathy gap is a cognitive bias which influences people to misunderstand the power over their behaviour that urges and feelings have, such as pain (Nordgren, Banas & MacDonald, 2010), hunger (Nordgren, Van der Pligt & Van Harreveld, 2009), sexual arousal (Ariely & Loewenstein, 2006), fatigue (Nordgren, Van der Pligt & Van Harreveld, 2006) and cravings (Sayette et al., 2008). For example, when someone is angry, it is difficult to understand what it is like for one to be happy, and vice versa; when someone is blindly in love with another, it is difficult to understand what it is like for one not to be (as well as to imagine the possibility of not being blindly in love in the future). It is possible that, in the medical settings and the workplace environment, empathy gap bias can cause serious damage, for example, when a doctor needs to accurately diagnose the physical pain of a patient (Loewenstein, 2005) or when an employer needs to assess the need for an employee's bereavement leave (Nordgren et al. 2011). These examples of medical and workplace settings are known as the effects of the hot-cold empathy gap (Loewenstein, 1996). Loewenstein (1996) has stated that when an urge is activated, individuals cannot evoke the visceral experience of this urge. For example, individuals cannot invoke a feeling of pain that they had earlier experienced. When one's hunger is excessive, one cannot readily elicit the visceral experiences of hunger—the mild throbbing around the stomach region or tightness around the chest. As the other individuals cannot readily invoke these visceral experiences, they underestimate the intensity of these urges, and another individual thinks such urges in others can be easily suppressed or retrained. Therefore. It can be said that empathy gap bias may influence myriad areas of life, from decision making to bullying, smoking, marketing and more.

Over the last decade, the empathy gap bias gained interest among psychologists. Nordgren et al. (2009) have studied how physical pain could be underestimated by other individuals if they have not experienced the same kind of pain. Later, Nordgren et al. (2010) showed that because of the empathy gap, individuals tend to underestimate the pain that other people felt when they are not in such pain. They demonstrated the empathy gap bias in social exclusion situations in a series of studies where some participants were either excluded or not excluded in a social setting. Next, in that setting, they rated the degree to which other people were likely to feel negative emotions when excluded. Relative to participants who were excluded, participants who were not excluded underestimated the negative emotions that other people would feel when shunned. In this study, participants who felt included now underestimated the pain they experienced when they were previously excluded. In another experiment, Loewenstein (1999) explored visceral factors related to addictions like smoking, which are associated with states essential for living (e.g. fatigue and hunger). Their findings yielded new discoveries about the hot-cold empathy gap and its important role in drug addictions.

A hot-cold empathy gap occurs when people are unable to understand others' emotions under the influence of anger, pain, sorrow or hunger in their own behaviours. This bias can affect people's judgement in various fields, such as medical decisions, bullying in schools, addictions, etc. On the other hand, empathy is the capability to understand what others feel; namely, the capability of projecting oneself into another's position (Bellet & Maloney, 1991). People with empathetic understanding can usually process many factors of decision making and problem solving. Past experiences such as childhood trauma, lack of parenting or accidents can influence such decision-making and problem-solving skills to differ from most individuals, who would respond with a seemingly obvious response (Dietrich, 2017). With the influence of empathy gap bias, however, people are unable to understand other's true emotions, but if the other individual has true empathic understanding, then he or she might recognize the gap in the other's lack of understanding. So, it is arguable that the empathy gap occurs when one objects is genuinely capable of empathy but the other is not because the 'other' in this scenario is always influenced by the bias but unaware of that influence. Meanwhile, the empathic individual might understand the 'other' individual's situation, as it is possible for empathic people to project themselves in others and understand why they act in that certain way. Therefore, the empathy gap might not refer to the one who is truly capable of empathy, but the empathy gap bias affects the other subject who is not empathic.

In the current experiment, I tested this effect of the empathy gap bias in robots by setting up its behaviours in hot or cold state at the beginning of the conversational and game-playing interactions. I allowed the responses based on this set bias for the whole interaction to find out how the participants reacted.

It can be seen that the empathy gap is a well-studied cognitive bias, which is also very common in humans, and it is possible that the effects of such bias can be developed in robots. It is usually expected that the robots are machines, and thus they do not feel pain or emotions and can perform tirelessly. As I have discussed earlier, however, such feelings of pain, urges and misunderstandings are common human behaviours and can be seen in daily life; therefore, they should be tested in social robots' interactions with humans. More specifically, some investigation of the user-impacts of a robot expressing behaviours such as tiredness, sadness or excessive happiness, or exhibiting differences from the user's emotions.. To evaluate conditions, I used hot-state and cold-state empathy gap biases in robots. These two types of empathy gaps were developed on two robots: MyKeepon (Figure 3.7) and MARC (Figure 3.3).

MyKeepon expresses the behaviours of the hot and cold states of empathy gap biases by not behaving as the participant expects in a non-conversational interaction. The robot maintains a specific tone in interactions by expressing excessive happiness or remaining calm to show excessive sadness. The robot MARC has conversational interactions and shows the same level of excessive happiness or sadness in different interactions. Both of the robots show expressions as they misunderstand a participant's current state of emotions and vice versa, and thus continue to interact based on their own mood states. The developed algorithms aim to deliver the impression of a robot with empathy gap bias, which fails to understand the participant's state of emotions and does not try to understand at all. It was expected that the participants would understand such biased and non-biased behaviours of robots, and respond based on their experiences in the interactions.

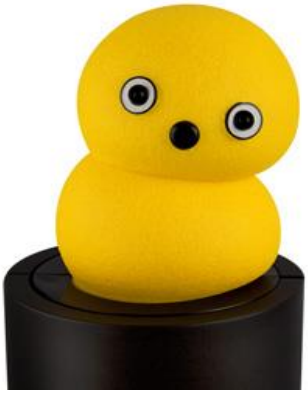


Figure 3.3: MyKeepon robot

3.6.3 The lack of knowledge, or the ideocracy effect

The lack of knowledge, or the ideocracy effect, is another human imperfection, based on the Dunning-Kruger effects, where relatively unskilled individuals suffer from illusory superiority, mistakenly assessing their ability to be much greater than is accurate. Kruger and Dunning (1999) described this bias effect as follows: '…incompetent people do not recognize—scratch that, cannot recognize—just how incompetent they are'. Their research also suggested that highly skilled individuals may underestimate their relative competence, erroneously assuming that tasks which are easy for them are also easy for others, and they may incorrectly suppose that their competence in a particular field extends to other fields in which they are less competent. These two psychologists found four main components of this bias which, for a given skill, they proposed incompetent people would demonstrate (Lee, 2012):

a. failure to recognise their own lack of skill,

b. failure to recognise the extent of their inadequacy,

c. failure to recognise genuine skill in others, and

d. recognition and acknowledgement of their own lack of skill, after they are exposed to training for that skill.

Dunning (2005) drew an analogy—'the anosognosia of everyday life'—in which he showed that a physical condition in a person who experiences a disability because of brain injury seems unaware of that disability, even with dramatic impairments such as blindness or paralysis: 'If you're incompetent, you can't know you're incompetent.… The skills you need to produce a right answer are exactly the skills you need to recognize what a right answer is' (Morris, 2010).

The Dunning-Kruger effect touches on a critical aspect of the default mode of human thought and a major flaw in human thinking, and it also applies to everyone (i.e. all human beings at various places on that curve with respect to different areas of knowledge) (Kruger & Dunning, 1999). People may be an expert in some things and competent in others, but will also be near the bottom of the curve in some areas of knowledge (Novella, 2014).

The effects of Dunning-Kruger bias can be explained in many ways. Novella (2014) on the Theness's Neurologica blog explains such effects, suggesting that when an individual tries to make sense of the world where he or she works with existing knowledge and paradigms, the individual also reconstructs ideas, then systematically seeks information that confirms those ideas. Usually, they dismiss contrary information as exceptions in the process and interpret ambiguous experiences in line with their own theories. They then remember their memories and tweak any experience that seems to confirm what they believe.

The results of the Dunning-Kruger effect show intelligent people falsely assume that others have equal understanding, which is why being smart or talented can actually result in lower self-confidence. The outcome of all of this insight is that less fortunate people feel more confident about their abilities than those who actually excel in those areas. A famous quote from Charles Darwin perfectly describes this situation: 'Ignorance more frequently begets confidence than does knowledge' (Charles Darwin Quotes, n.d.). A real-world example may be a situation in a bar: Two guys in a bar are both looking at the same girl. The smarter guy doubts his ability to talk to women, despite his eloquence. Meanwhile, the less gifted guy gives no second thought to his ability and walks straight over to the girl. Ignorance wins in this situation. In general, people with higher level of intellectual tend to over-think everything, and usually those thoughts are more critical than uplifting. Instead of logically accepting that he or she is more skilful than others at a certain task, the high intellectual person's overthinking only sometimes leads to a decrease in confidence. As Dunning and Kruger (1999) concluded, 'the miscalibration of the highly competent stems from an error about others.'

The Dunning-Kruger effects is a recently discovered cognitive bias that is very common among people. The experiments described in Section 5.3, the lack of knowledge or idiocrasy effects, were created based on such Dunning-Kruger effects. These effects were developed in the robot MARC and subsequently observed in various conversational interactions. People who are influenced by this bias can often be noticed in daily life, social media, television, work places etc. This bias can be developed in a robot's behaviours as it can be easily expressed and is understandable by participants.

The lack of knowledge or the ideocracy effect were developed based on three main components, where the robot fails to recognise its own lack of knowledge of a given subject, fails to recognise true knowledge of the participants on that given subject, and in the end realises and admits its incompetence. The lack of knowledge or ideocracy behaviours algorithm was developed for the robot to express stupidity, lack of knowledge on a conversational subject, falsely confidence and ignorance. However, at the end of the interaction MARC realises its mistakes and acknowledges them appropriately.

3.6.4 Self-serving bias

The self-serving bias is another commonly found cognitive bias in people, which relates to attribution for their personal outcomes. Self-serving bias happens when people make internal attributions of responsibility for desired outcomes and external attributions for undesired outcomes (Shepperd, Malone & Sweeny, 2008). A classic example of self-serving bias is a student taking an exam. If the student does well on the test, he or she is more likely to believe that his or her own ability and effort (i.e. things under the student's control) were the reasons for her success. However, if he or she receives a poor grade on the test, the blame falls on external factors such as luck, the difficulty of the task, or the lack of cooperation of others (Campbell & Sedikides, 1999). The student might claim that the professor made up an unfair test, or that the lighting in the room was too dim, so the student could not focus.

In the workplace, workers attribute their promotions to their own hard work and exceptional skill, but they usually attribute denial of promotions to unfair bosses or other external causes. Athletes sometimes credit themselves for performing well in the sports arena, but when they perform poorly, they blame external causes (DeMichele et al., 1998). It is even evidenced in drivers who attribute accidents to external factors, such as the weather, the condition of their car, and other drivers, yet attribute the narrow avoidance of an accident to their alertness and finely honed driving skills (Stewart, 2005). It has also been noted that self-serving biases critically affect interpersonal relationships, where both parties in relationships blame each other (Campbell et al., 2000). Shepperd et al. (2008) have shown that people's motivational processes and cognitive processes can be influenced by the effects of the self-serving bias.

The concept of the self-serving bias was introduced in late 1960s (Miller & Ross, 1975), and since then its effects have been observed in many studies. Scientists argue that some of the most noticeable and major differences in components include roles (actor or observer), task importance, outcome expectations, self-esteem, achievement motivation, self-focussed attention, task choice, perceived task difficulty, interpersonal orientation, status, affect, locus of control, gender and task type (Campbell & Sedikides, 1999). Campbell and Sedikides (1999) examine the above factors and in 13 of 14 categories, the condition that was hypothesised to give a greater sense of self-threat caused a self-serving bias of a greater magnitude. Self-serving bias is triggers prevalently when people feel that their concept of self is under threat. Some people, however, are less susceptible to this bias than others. Individuals who are depressed or suffer from chronically low self-esteem usually tend to take the blame for their failures but deny responsibility for their successes (Rotter, 1966). Rotter (1966) noticed that individuals with an external locus of control consider both positive and negative outcomes to be caused by factors beyond their control, and those with an internal locus of control assume that they, personally, are responsible for their outcomes. Miller and Ross (1975) have suggested that people who blame others for failures and attributing their successes to themselves are usually very optimistic when predicting the outcomes and experiences. Piehlmaier (2014) has suggested that people usually optimistic in nature (e.g. who expect to pass tests, get good jobs and have good relationships rather than failures) get fired from jobs or divorced, and that the reason may be that negative experiences are generally unexpected, but on the other hand, it is also true that people who tend to attribute negative experiences to external factors are rather unstable and are more depressed than others.

Self-serving bias may depend on interpersonal closeness and relationships in a social context. Campbell et al. (2000) have investigated groups working in pairs to complete interdependent outcome tasks and found out that the relationally close pairs usually do not show a self-serving bias, but relationally distant pairs do. Sedikides et al. (1998) arrived at similar results, and suggested that the reason could be that close relationships place limits on an individual’s self-enhancement tendencies. Campbell et al. (2000) investigated pairs who performed an interdependent-outcomes creativity test and were then given a bogus pipeline for success or failure outcome. Strangers exhibited the self-serving bias in responsibility attributions, but friends tended to make joint attributions for both success and failure. Researchers have taken this as evidence of ‘boundaries on self-enhancement’ (Campbell et al. 2000).

From the discussion, it can be said that evidence of this bias is well studied, and the behaviours are easily recognisable in people. This bias can easily influence people and affect relations between friends and family members. Self-serving bias can be easily developed in robots to express its effects in game-playing scenarios and then be compared with a non-biased game playing interactions. Therefore, I chose the self-serving biases to be used in the experiments on game-playing interactions with the robot MARC. The self-serving algorithm was created based on the bias’s general principles (i.e. the robot does not like to lose in games against participants). The robot takes full credit for winning a game and happily brags about the win; however, if it loses, it then makes several excuses and even accuses the opponent of cheating. The algorithm’s development and other experiment details are described in the next chapter.

3.6.5 Humour effect

Humour effect bias is a known cognitive bias of memory. It has been studied that humorous items are more easily remembered than non-humorous ones. This tendency could be explained by the distinctive effects of humour on the human brain, the increased time to process and understand the humour’s contents, and the emotional arousal caused by the humour (Seymour, n.d.). The beneficial effect of humour on experienced emotions may be based on the mechanism that processing humour requires cognitive resources, so that people are distracted from negative stimuli (Strick et al., 2009).

Humour is a part of everyday life. Humans inherit their desire to laugh, and that motivates various social actions, such as sharing funny YouTube videos or responding to text messages with a ‘LOL’ or emoticons. People even choose to get their daily news in comedic form from outlets like *The Daily Show, The Colbert Report* or *The Onion* (Jasheway, 2016).

Various studies have suggested the benefit of humour: For example, people who are associated humour in relationships are usually in satisfying and healthy interpersonal relationships (Salovey & Mayer 1990; Salovey, Mayer & Caruso, 2002). Other studies found that humour can be associated with social competence (Levine & Zigler, 1976), intimacy (Mutthaya, 1987; Hampes, 1992), trust (Hampes, 1999) and strong interpersonal relationship in marriages (Ziv 1988; Rust & Goldstein, 1989; Ziv & Gadish, 1989, Lauer, Lauer & Kerr, 1990; Kuiper, Martin & Olinger, 1993; Kuiper, McKenzie & Belanger, 1995)found that more humorous individuals are more likely to perceive potentially threatening events in a positive manner than less humorous ones. This trend has been used in this study’s experiments to represent instil humour in the MARC robot (Chapter 5.4.5).

Martin et al. (2003) found that an interpersonal form of humour, called affiliative humour, involves telling jokes, saying funny things, or witty banter, to put others at ease as well as to amuse and to improve relationships. Self-enhancing humour ‘involves a generally humorous outlook on life, a tendency to be frequently amused by the incongruities of life, and to maintain a humorous perspective even in the face of stress or of adversity’ (Martin et al., 2003). Both affiliative and self-enhancing humour are positively correlated with intimacy, extraversion and openness to experience (Martin et al., 2003; Saraglou & Scariot, 2002).

Weinberger and Gulas (1992) studied how humour can attract attention. In a practical example within a classroom, a student tends to pay more attention to the lectures when professors use humour in their talks than when they do not. This tendency could be helpful to remember in certain lectures, since a range of humour can be used in the classroom, such as satire, irony and sarcasm. Humour comes in many forms, though, such as in misunderstanding, mocking, puns, wordplay, surrealism, black humour and more.

Overall, the effect of humour is a popular area of research, and these effects are common in people's lives. This thesis used bias associated with humour to points the participants' attention to the robot. The initial goal of choosing this bias was to explore whether participants could remember such light interactions with the robot, in which the robot made several jokes throughout the entire interaction. The humour effects algorithm was developed in the robot MARC for game-playing interactions against participants. To show the effects of humours, MARC does not care if it loses or wins the game; instead, it makes several funny comments on both situations of winning and losing, so that if a participant loses, they don't feel bad about it. Such 'funny comments' (e.g. small jokes) were gathered based on situations, such as if the robot wins, the jokes are based on self-defeating humour where the robot tries to amuse the participant by showing that the win does not matter; similarly, if it loses a game, it makes a comment based on affiliative humour and self-enhancing humour, which shows that the robot is amused by the participant's wins. All the comments and jokes were made to be light and short; therefore, the participants were expected to pay more attention and enjoy the game-playing interaction. The robot poses its humour such that it can create the impression of a positive outcome for the game even when the robot loses. The development of this algorithm is discussed in the next chapters (Chapter 4 and 5).

3.7 Summary

In the first section of this chapter, I mainly described important and relative studies on human–robot interactions in different fields and using different ranges of social skills. The main aim of this section has been to give readers an idea of social-robot interaction, how it differs from other kinds of interaction, and most importantly, how people understand interactions with machines where the machines show various social behaviours. I discussed the reasons for considering cognitive biases in my experiments to develop in robot’s interactive behaviours, and to do that I explained how cognitive biases play important roles in human nature and such biases’ effects on the human brain, decision making and social relations. I posed my arguments as a hypothesis and in research questions, and I stated briefly how I study biases in social robots. I then discussed the cognitive biases I selected for my study and stated why I choose such biases. This chapter offers, in sum, information on different application areas for social robots, the role and importance of cognitive biases in humans, and my argument on developing some of these biases in the social robots.

The next chapter discusses the procedures I followed to develop such biases as algorithms in the robots, the experimental methods, data collection metrics and methods, and participants.

Chapter 4:
Methodology and Setup

4.1 Introduction

In this chapter I describe how different cognitive biases were developed in the robots and the methods used to set up the interaction experiments between the robots and participants.

All the described experiments have structural similarity, where biased version of robot interaction with participants are compared non-biased interactions. Both of such biased and non-biased algorithms were programmed and developed in robots separately to compare both types of interactions and find the participant's preferences between biased and non-biased algorithms.

In the next section I briefly discuss the pre-experimental setups, considering the details of the robots, how the algorithms were made, how the experiments were performed, the data collection methods and other related procedures.

4.2 Pre-experimental setup

To experiment the research hypothesis, I needed to develop the effects of the selected cognitive biases in the robot and prepare human-robot interactions where participants would experience such biases in the robot. To determine the effects of the biases, I needed to differentiate such bias effects with another interaction where participants experience an interaction without the selected biases. At the end, if it is possible to compare their experiences between biased based interactions and non-biased interactions, then it would be possible to find out the effects of the cognitive biases in the robot in human-robot interactions.

It is important to mention that in the current experiment, the focus was not on the robot's performance in human-robot interactions. Rather, I aimed to investigate the effects of cognitive bias in a robot's behaviours in human-robot interactions. Therefore, the most important aspect of the experiment is to measure the users' experiences with the robot (Kim & Mutlu, 2014), such as how the participants felt in interaction. The manipulation of the robot's behaviours was performed by the 'Wizard' during the experiments, and therefore, the same human error is possible in all of the experiments. Consequently, I did not determine a need to measure human error in the current experimental scenarios. If any human error occurred during manipulation of the robot's behaviours, the same error was common for all of the biased and non-biased interactions and therefore had very little influence on the participants. To measure the user experience with the robot, various questions were asked, including the scale to measure the aforementioned factors such as likeability, comfort, pleasure and rapport. For the likeability scale, a minimum of 5 to a maximum of 12 questions were asked, and the number varied based on the procedure of the experiment. For the comfort, pleasure and rapport scales, 6, 8 and 25 questions were asked, respectively. The likeability scale questioned, "Was the game easy to play?" and "How easy was the conversation with the robot?", for which the participant answered with a rating from 1 (did not it like at all) to 7 (liked it very much) to measure the 'likeability' of conversational and game-playing interactions (Hone & Graham, 2000). The pleasure scale asked the participant to rate phrases such as, "Playing the game with robot is enjoyable to me" on a scale from 1 to 7, which measured how satisfied the participants were with their experience in hedonic (Hassenzahl, 2003) and emotional terms (Norman, 2004). In the comfort scale, participants were given statements such as, "Playing the game with the robot is uncomfortable for me." The answers used a similar rating system from 1 to 7 to understand participants' level of cognitive comfort in playing the game with the robot.

In addition to the above three scales, another important scale was added to measure the level of rapport which participants experienced with the robot. The questions for this scale were developed with the help of a previous study on human-robot proxemics (Mumm and Mutlu, 2011). The questions for this scale were statements to be rated and included, "The robot has a great deal of influence over my behaviour" and "The robot and I are very close to each other". Participants rated these ideas on a scale from 1 to 7. All of the scales and questions are provided in Appendix B, page 263.

To accomplish the abovementioned task, I planned all of the experiment methodologies based on a comparison of two sets of interactions such that it is possible to compare an interaction based on selected biased and non-biased behaviours in the robot. As I am comparing biased and non-biased behaviours in interactions, it was therefore important that the selected biases can be thoroughly expressed by the robot so they can be easily understood by participants. Some of the selected biases can be expressed better through conversation or by game playing, or by a mix of both conversational and game-playing interactions.

1. Misattribution: To show this bias effects the robot should display forgetfulness, which can be expressed in conversational interactions, but there should be multiple interactions to gather information so that the robot can forget.

2. Empathy gap: The robot could express this bias effect in both game playing and conversation. So, in the first experiment, the robot showed empathy gap expressions in game playing interactions, but in the second experiment, the robot expressed the effects in conversational interactions.

3. Lack of knowledge (based on Dunning-Kruger): In this case, the effects of the biases can be expressed in a conversational interaction where the robot shows its lack of knowledge of different subjects.

4. Self-serving bias: The effects of self-service can be better expressed in a game-playing situation where the robot refuses to lose a game or brags for winning a game.

5. Humour effects: This bias can be expressed by robot in a game-playing and conversational experiment in which the robot tells a joke after a game, regardless of winning or losing, so that the interaction can have an impact on the participant and the bias effects can be established properly.

To examine the chosen biases in robots I needed to set up experiments where people could interact with the robots and give feedback. I needed to prepare the robots for the interactions, plan the interactions between robots and participants, configure the algorithms based on biases and the lack thereof and choose the appropriate measurements so that participant feedback could be measured, advertise the study, and select the participants. Therefore, to examine the effects of the biases and imperfectness in robot in human–robot interactions, I had to consider the following pre-experimental standards for each of the experiments:

1. preparation of the robots,

2. interaction designs,

3. algorithms,

4. participants and groupings, and

5. data collection and measurements.

4.2.1 Preparing the robots

To test ‘social robots’, as per the hypothesis, I set up interactions between people and different kind of robots are able to interact following social rules. In the experiments, three robots were used: ERWIN, MyKeepon and MARC (Figure 4.1). The first, ERWIN, is a robot head capable of expressing five basic emotions. MyKeepon is a small, toy-like robot, and MARC is humanoid and interacts similar to humans. The robots differ in appearance, as well, enabling me to investigate the effects of biases in interactions between people and robots with different appearances.

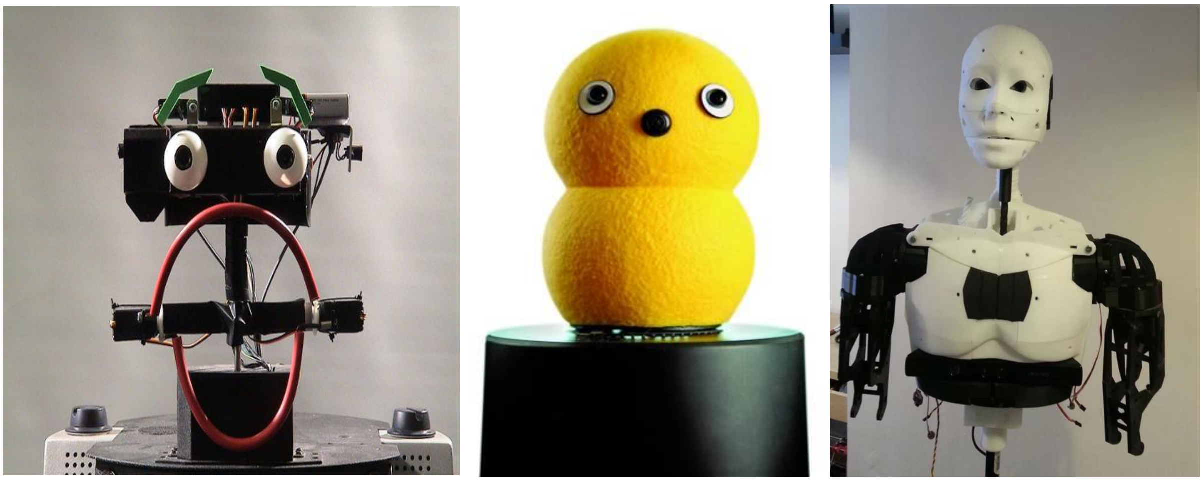


Figure 4.1: ERWIN, MyKeepon and MARC (from left to right).

As the robots have different functionalities, the design of the interactions could be different for the robots. For example, MARC can interact based on verbal abilities and gestures, while ERWIN interacts based on verbal and emotional expressions, and MyKeepon interacts based on different noises, jumping, and dancing. However, the design of the interactions followed two basic principles:

a. The interactions should allow the robot to engage with the participants in conversation, game playing or both.

b. In the interaction scenarios, the robot should be able to express its biased or non-biased behaviours by using its own functionalities (e.g. speech, noises, gesture, emotions, etc.).

In the following sections, the different algorithms that robots followed in interactions are discussed.

Control system was not affected

It is important to understand that the biased and non-biased algorithms are a set of scripted behaviours created depending on the functionalities of the robots (e.g. verbal or non-verbal abilities, gestures performing abilities, emotions expressions abilities etc.) and the main components of the selected biases. In the experiments, the developed algorithms did not change the control systems of the robots; rather, they manipulated the behaviours so the robot could show a set of selected behaviours based on cognitive bias. For example, the ERWIN (Figure 4.1 far left) robot which I used in my first experiment can express five prototypical emotions and can speak; therefore, the developed misattribution-biased behaviours were expressed by misattributed speech, followed by emotional expressions. Therefore, it can be said that the cognitive biased behaviours I used in the robots did not modify the robot's control systems, but were programmed using software based on the robot's own functionalities, and in the experiment I controlled the robots by giving various inputs through the software to the robots to enact various behaviours during the interactions.

Assessing robot's behaviours

In the current research, I chose to study five cognitive biases: misattribution, empathy gap, idiocrasy effect, self-serving bias and humour effect. The reason for the choice of these biases has been stated earlier (Chapter 3.5, page 60). Before starting the main experiments, all of the biased behaviours were checked in several mock interactions. To verify if the behaviours shown by the robot were understandable and appropriate according to the selected bias, several mock interactions were performed between the robot and my supervisor, graduate students and lab colleagues. After ensuring that the robot expressed the correct behaviours and the participants could understand what the robot showed, then the actual experiments were carried out. Based on the received feedbacks and suggestions from those mock interactions several changes were made for the actual interactions such as focusing on developing main components of the biased behaviours (e.g. for the misattribution bias, the dialogues were designed based on three main components false memory, source confusion and total forgetfulness), interactions has divided into several stages (e.g. stages for conversational interaction are meet and greet, conversation, and farewell).

It should be noted that the behaviours (biased and non-biased) were designed in the robots without considering the robot's own existing functional faults and biases (if any). The reason of not considering the robot's own faults and biases is that both of the types of behaviours (i.e. biased and non-biased) were developed in the same robot, so in both interactions, such functional faults and biases (if any) are present, therefore not making any difference in terms of the findings of the experiments.

4.2.1.1 ERWIN

The Emotional Robot with Intelligent Network (ERWIN) was used in the first experiment to test misattribution bias, and ERWIN is a robot head placed on a metal base. The robot is around 40 cm tall, including the height of the base.

Hardware modification

ERWIN (Figure 4.2) has a total of four servos, two for the jaws and another two for the eyebrows. Two cameras of 2.0 MP are placed in its eyes, so that it can record user expressions face to face. Two lips and adjacent motors are connected in a metal plate, and similarly, the eyes and eyebrows are placed and connected by another metal plate. These two plates are vertically connected with a rod, placed on a metal base. The whole hardware system looks like a mechanical face with prototype expressions. Motors are controlled by a Pololu microcontroller placed on the back of the head. The robot can rotate 90 degrees on its base.

Software used

ERWIN was programmed using C. Its voice was made using text to speech software called 'Speakonia', for the following reasons:

i. more real time than typing the robot's speech and selecting from the different voices to speak on the spot,

ii. different tones and voices to choose from, and

iii. ability to overcome the delay between participants' and ERWIN's responses.

Speakonia has eight different voices, and it can also make prosodic sounds in longer sentences.

Function and features

The robot can move its jaws and eyebrows, which show five basic expressions. Such expressions are happy, sad, surprised, angry and shocked or afraid. Figure 4.3 shows the different emotions of ERWIN.

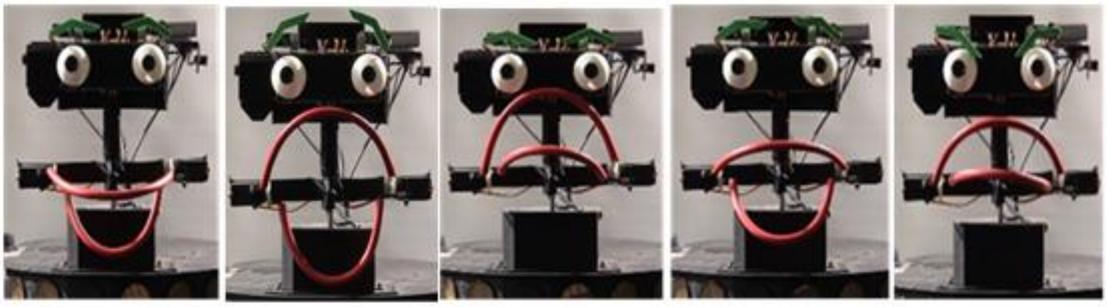


Figure 4.2: Different emotions expressions of ERWIN (from left to right: happy, surprised, angry, shocked and sad).

In this experiment, the emotive expressions were used as a tool for interacting with the participants, but the main goal was to discover the effects of the misattribution bias and forgetfulness in the robot's interactions and how hey effects the participant's *fondness or likability* of the robot.

4.2.1.2 MyKeepon

MyKeepon is a small, yellow robot which I used in the second experiment to test the empathy gap bias. Kozima et al. (2009) designed and developed the robot at the National Institute of Information and Communications Technology (NICT) in Kyoto, Japan. The robot's simple appearance and behaviour are suitable to aid children with developmental disorders such as autism to understand attentive and emotive actions.

Keepon, which is shown in Figure 4.3, is a small toy-like robot containing two yellow blobs on a black cylindrical base. It has two interactive modes: touch and dance. The robot has limited vocal ability but can make noises representing various moods and emotions. If the user touches the robot or taps on its head, Keepon usually responds by circling around towards its touch side and making noises. The robot is able to detect music and sounds and responds by dancing or jumping. Because of the robot's shape, size, colour and interactions, it is very popular among children, and it has been used in many human–robot interaction experiments in recent years.

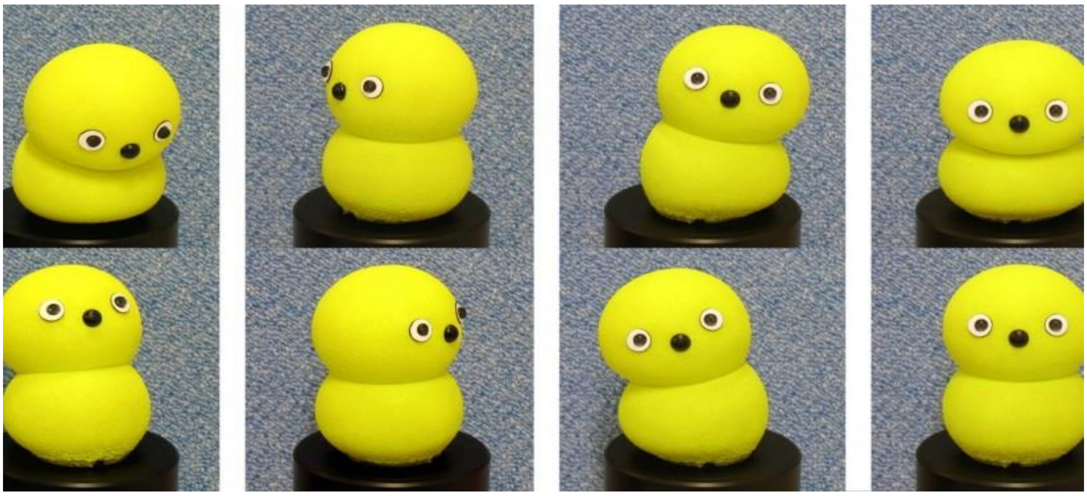


Figure 4-3: The robot Keepon showing different positions.

Hardware modification

MyKeepon is the beta version of the robot Keepon Pro, and it can do almost everything the Pro version could. It has a touch mode and dance mode, which are activated by two buttons located at the front of its platform (Figure 4.5). In the experiment, I modified MyKeepon using Arduino Uno (Figure 4.4) and in this modified version, MyKeepon is fully functional and can be used and controlled remotely using a laptop.

As soon as it is active, MyKeepon tries to attract the participant's attention. Similar to the original, MyKeepon responds to squeezes, tickles and taps and also has own noises to attract attention and expressing emotions.

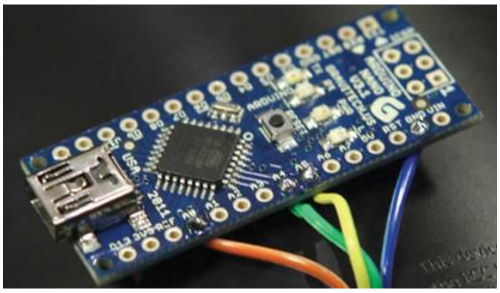


Figure 4.4: Arduino Uno.

In its dance mode, a sensitive microphone allows it to hear the music being played, and it dances to it. The microphone is located on its nose so that it can hear be better.

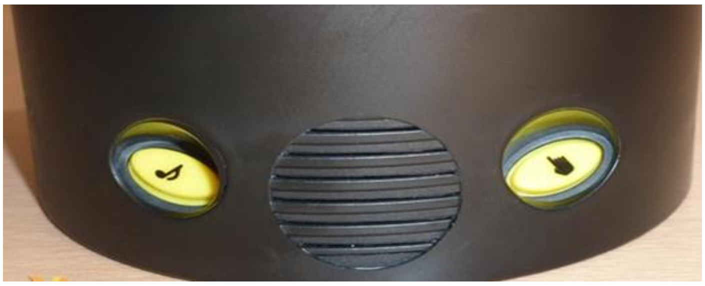


Figure 4.5: Music mode and touch mode buttons of MyKeepon. The music note symbol is for dance mode and finger symbol is for touch mode.

In the touch mode, MyKeepon responds to any touch, poke or tap on its body and head, by jumping, dancing and moving in the direction of the stimulus.

Software used:

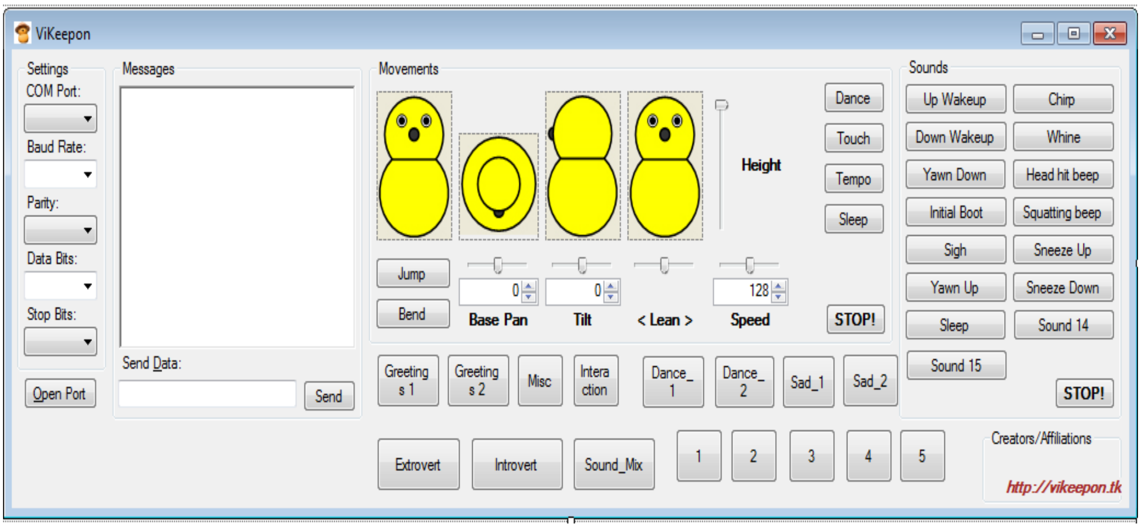


Figure 4.6: The ViKeepon interface.

An open source software interface called ViKeepon (Figure 4.6) was used to control the robot remotely during the interactions. ViKeepon allows the creation of custom movements and sounds, which can be used remotely just by clicking on the buttons. As seen in Figure 4.6, each button represents a set of movements to perform a particular task, such as ‘Greetings1’, which has three movements and two sounds to perform a warm welcome to the participant.

ViKeepon supports a minimum of 15 different sounds, including wake up, yawn, sleep, chimp, sneeze and others. In the experiment, it was possible to make different create different expressions by combining such custom moves and sounds together, for example greetings, sadness, great sadness, joy, concentration and others.

4.2.1.3 MARC

A 3D-printed humanoid robot, MARC was used in the third and fourth experiments to find the effects of the misattribution, empathy gap, self-serving and humour-related cognitive biases, as well as certain idiotic behaviours showed by the robots. Much research suggests that the human-like body of a humanoid robot helps its users to understand the robot's gestures intuitively (Hayashi et al., 2005). A recent study suggests that human brain can equally well perceive the actions of non-humans and humans (Wykowska et al., 2015). The reason for this ability could be that the actions of general gestures which evolved in socio-cultural context of human–human interaction allow also for intuitive human–robot interactions. Such actions might include the specific speech and gestures with arms and hands and body language to perform some specific expressions. These expressions are easy to interpret by the human sensory system, because it has been practiced since childhood. In the same way, the humanoid robot with anthropomorphic abilities can express human-like gestures, which can be known to its users and also easy to understand. Such gestures can be helpful to develop companion robots even for everyday social uses, as humans have a specific shape and size and have built the home environment accordingly, and a humanoid robot can easily fit into our houses. Humanoid robots are an important part of robotics, and they are very popular as entertainment; for example, Titan the robot (Titan The Robot, n.d.) is a giant, remotely controlled robot that can sing, dance, make conversation and entertain people. In the current experiments using MARC, the robot was controlled remotely and interacted based on speech and gestures using its hands, head, and neck.

The version of MARC used in this study (Figure 4.7) was fully 3D-printed based on the open source project called InMoov created by Gael Langevn (InMoov, n.d.). It was built from head to torso, and it can perform various human-like gestures using hand, head and neck movements. One of the important reasons MARC was chosen over other humanoid robots was the cost efficiency. The expenses of making MARC were minimal compared to other available humanoid robots. Each part of MARC was completely built with a 3D printer, and the total cost was approximately £700.00–£1000.00 to make this life-sized humanoid robot. The robot is around 170 cm tall and weighs nearly 12 kg.

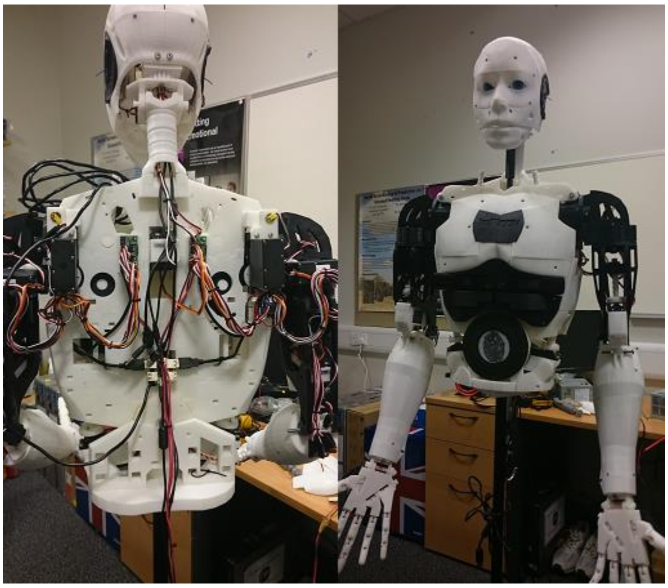


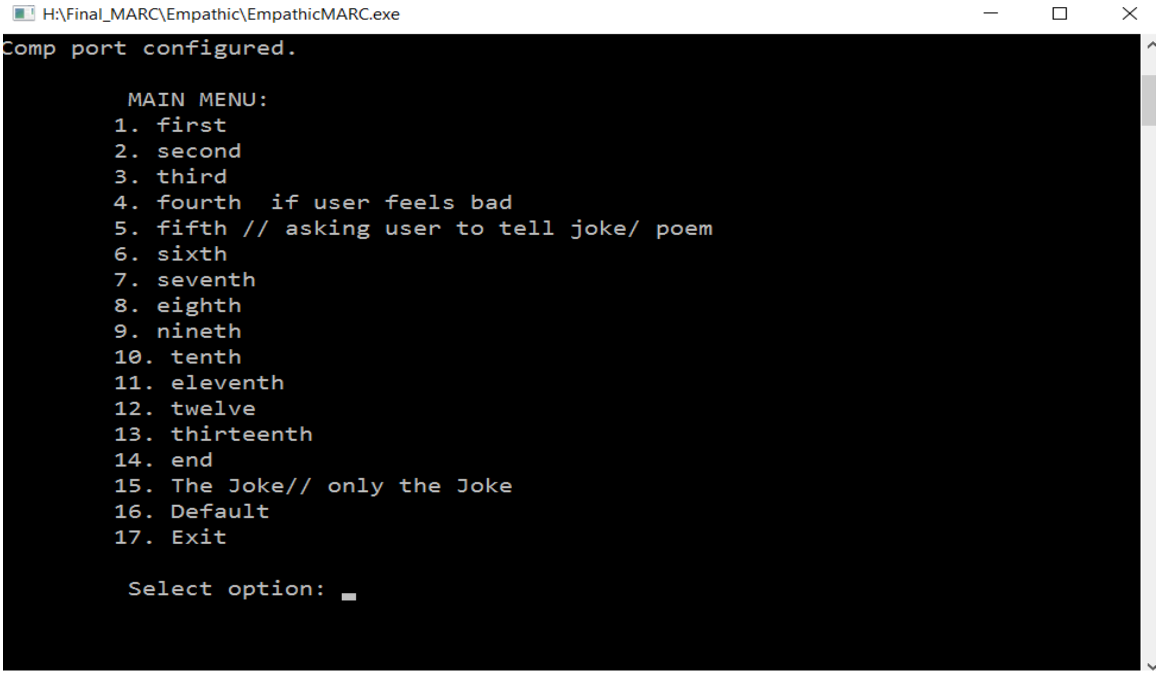
Figure 4.7 The back and front of MARC the humanoid robot.

Hardware modification

The various parts of MARC were designed using Google Sketch and then printed using a 3D printer. In total, nearly 200 parts were printed, and it took nearly three months to print them all. Acrylonitrile Butadiene Styrene (ABS) plastic was used to print the different body parts because of its high melting point and the fact that it is sturdier and harder than other printing plastics on the market. To print the forearms, approximately 3 kg of variously coloured natural ABS filament was used. Most of the body-part sketches were downloaded from InMoov, but some were created independently. Body parts were assembled after the printing of each of the body parts was complete. There are nearly 30 servos in MARC's body which help the robot to perform gestures, and MARC has two cameras in its eyes which help to watch and record the interactions. The robot can move its wrist rotationally, and both hands are able to move up, down, front and back. Its head can move up and down and sideways, and can tilt frontwards and backwards, which helps it to look up at the user to make eye contacts within the range.

Software used

MARC's firmware was programmed using Raspberry Pi. The Pololu microcontroller was used to control all servos. Servos and microcontrollers were programmed using C. An interface was created using C++ to use in the experiments; Figure 4.8 shows the software interface where different inputs can be given for MARC to perform various actions. The voice of the robot was created using software called Audacity. At first the sample voice files were created using an online text-to-speech editor. Then, the voice files were edited using Audacity to make changes in pitch, increases in tempo and to make the voice sound bit robotic. As MARC looks like a male robot, the voice was chosen to be a masculine robotic voice. MARC can move its jaws, so with the robotic voice it looks like the robot is talking.



```
Comp port configured.

         MAIN MENU:
        1. first
        2. second
        3. third
        4. fourth  if user feels bad
        5. fifth // asking user to tell joke/ poem
        6. sixth
        7. seventh
        8. eighth
        9. nineth
        10. tenth
        11. eleventh
        12. twelve
        13. thirteenth
        14. end
        15. The Joke// only the Joke
        16. Default
        17. Exit

         Select option: _
```

Figure 4.8: Control interface of the robot command prompt.

4.2.2 Interaction design

The design of the experiments allowed a comparison between two interactions: an interaction where the robot interacted based on the effects of the developed cognitive bias, and one in which the robot interacted without the effects of such bias (this latter is referred to as non-biased, as unbiased or as the baseline).

Based on the selected biases, different algorithms were developed where the interactions were set for either conversation or playing games (e.g. rock-paper-scissors, tap-on-head) between the robot and participants. The conversational engagements were based on common topics often discussed between two people at an early stage of their friendship (e.g. favourite colour, favourite food, favourite sports etc.). Such conversational interactions were used to test biases related to memory or those which were most suited to conversation. For example, to show the effects of misattribution bias, I chose conversation between the robot and participants. To show the effects of lack of knowledge or ideocracy effects and empathy gap biases, conversational interactions were used. However, selection of conversational interactions depended partly on the robot as well. For example, empathy gap bias was tested in game-playing interaction with MyKeepon (tap-on-head), but also in a conversational interaction with MARC. I used a game (rock-paper-scissors) as a mode of interaction for the self-serving and humour biases with MARC, because to express the effects of such biases the results of a game are necessary (details in Chapter 5).

To show the cognitive bias-based behaviours effectively, so that the participant could recognise and understand, the interactions were divided into several stages. For example, conversational experiments were divided into three stages (e.g. meet and greet, topic-based conversation and farewell), and game playing experiments were divided into five stages (e.g. meet and greet, explaining rules, playing games, overall score and farewell). Based on the selected biases, the conversational dialogues change in different stages; for example, with the misattribution-biased algorithm, the robot would forgets the participant's name and therefore greets the participant with pseudo names, but in the empathy gap algorithm, the robot does not need to forget the participant's names in order to express the bias, therefore robot remembers, and the effects of the bias are shown in later stages.

Table 4.1: The stages of the conversational and game-playing interactions

Conversational Interactions	**Game-Playing Interactions**
Meet and greet	Meet and greet
Topic-based conversation	Game rules explain
Farewell	Game play
	Result
	Farewell

As seen in the Table 4.1, each of the conversational interactions were divided into three stages, such as when the participant and robot meet and begin conversing, the middle of the conversation where they discuss various topics, and the end, when the participant leaves. For example, the interaction was divided into three stages to test the misattribution cognitive bias:

a. meet and greet, where participants meet with the robot;

b. topic-based conversation, where both parties make conversation on various topics; and

c. farewell, where interaction ends and the participant leaves.

Such three stages are present in both misattribution and non-misattribution conversational interactions.

As seen in the Table 4.1, the gameplay interaction was divided into five stages:

a. meet and greet,

b. rule explanation,

c. game playing,

d. final result, and

e. farewell.

These stages are meant to differentiate between biased and non-biased interactions and to effectively show the biased behaviours so that participants can understand the differences between interactions throughout the exchange (see Figure 4.9).

After meeting and greeting the participant, the robot requested that the participant play rock-paper scissor and explained the rules and. If the participants wanted to know the rules again, the robot would respond based on the developed bias. In the game play, one of three outcomes would result from a single round of gameplay:

i. robot wins,

ii. robot loses, and

iii. draw.

Based on such outcomes the robot responds differently, and such different responses are the key to making the interactions centre on the developed bias. The ‘final result’ is a state where the robot calculates results based on all rounds played and declares the winner. MARC's dialogues would be different in this stage, depending on the algorithm.

For the self-serving algorithm, the robot praises itself for winning, but blames other causes for losing. For humour effects, the robot offers various kinds of humour in the event of losing or winning. The game hands were drawn at random, so the outcomes were not fixed, because the goal of the experiments was to measure participant's responses to the biased or non-biased behaviours from the robot, not to measure win rates.

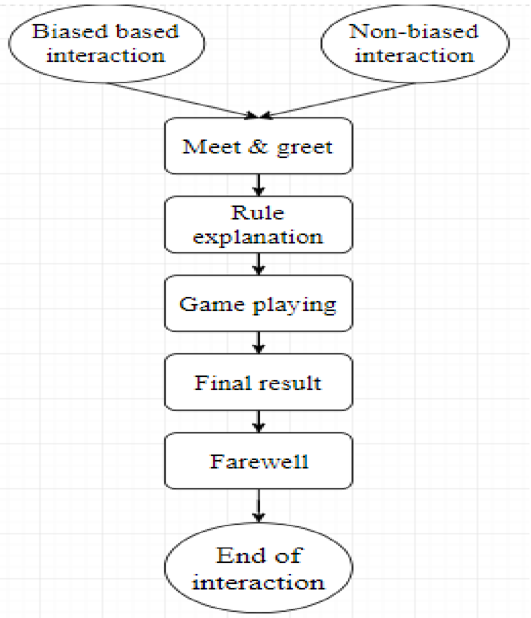


Figure 4-9: Theoretical stages in an interaction.

In the next section, I discuss the design of baseline, self-serving and humour effects algorithms, explaining the differences between the three algorithms and how they were performed by the robots in the interactions.

4.2.2.1 Conversation design

The conversational design for each topic was simple and question-answer based. In the experiment, the dialogue design of the general conversation was based on four steps (Figure 4.10), as follows:

i. robot asks a question or says something,

ii. participant responds,

iii. robot states its own opinion, and

iv. robot waits for participant's response or moves to next dialogue.

For example, MARC asks, 'Do you like football?' The participant can respond, 'yes' or 'no', and can extend their responses, but whatever the participant's responses, the robot would say something after that response based on the algorithm (e.g. biased or non-biased). The robot would then wait for few moments to check whether the participant wants to say something, otherwise it moves to the next line of dialogue. The differences in biased and baseline conversations arise in the third step, where the robot says something after the participant responds. In the baseline condition, the robot mainly says 'Okay' or 'That is great', which represents the end of that particular discussion and movement to the next topic, but in the biased interaction, the robot's dialogue reflects the bias effects, and the topic of conversation may be further pursued, depending on the participant's responses (See Figure 4.10).

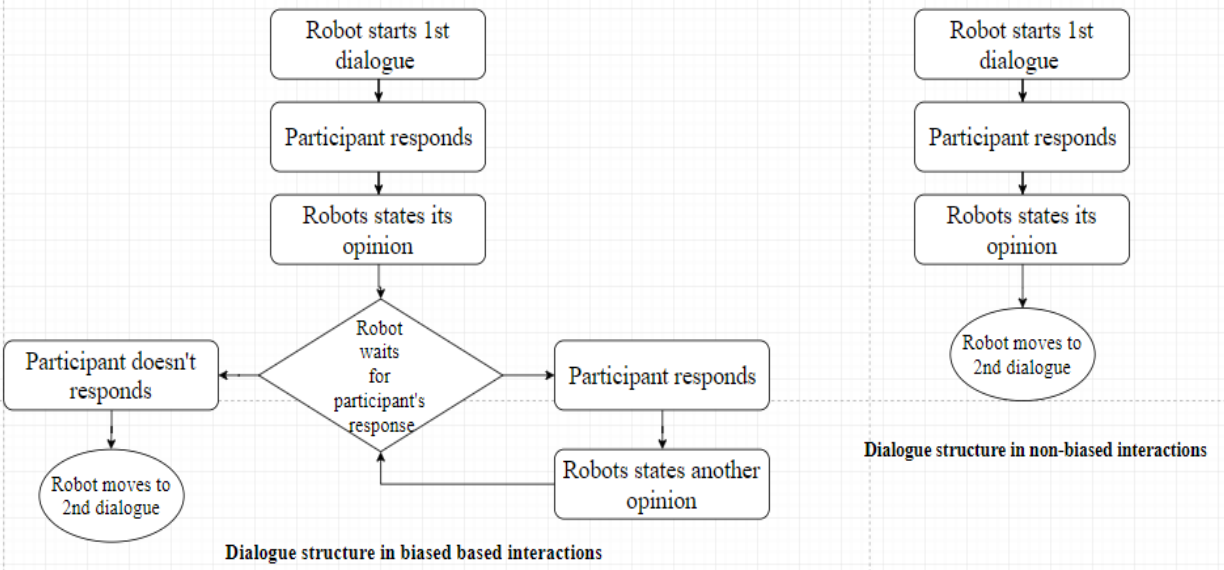

Robot starts 1st dialogue
Participant responds
Robots states its opinion
Robot waits for participant's response
Participant doesn't responds
Participant responds
Robots states another opinion
Robot moves to 2nd dialogue
Dialogue structure in biased based interactions
Robot starts 1st dialogue
Participant responds
Robots states its opinion
Robot moves to 2nd dialogue
Dialogue structure in non-biased interactions

Figure 4.10: Conversation design: Left side diagram for biased conversation and right for the non-biased

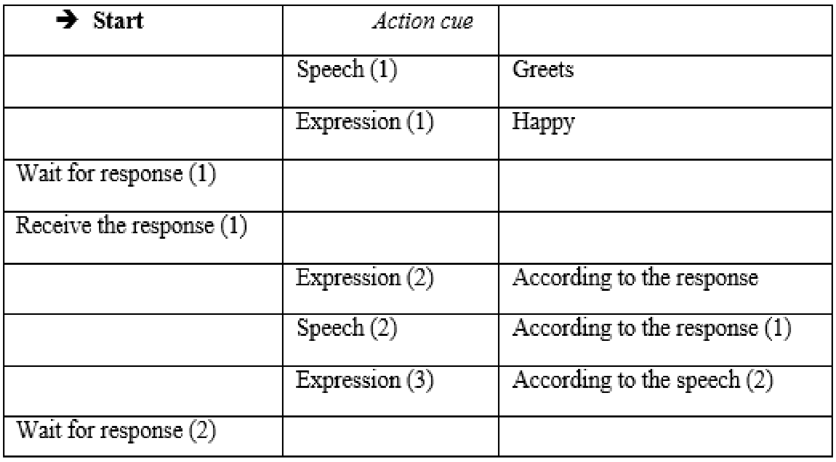
An example can be given to show the differences between non-biased and biased dialogues: in the meet-and-greet stage, there could be only three lines of speech from the robot for the baseline algorithm: (1) ‘Hello,’ (2) ‘My name is MARC, what is your name?’ and (3) ‘Nice to meet you.’ In the case of the biased algorithm, however, the robot’s dialogues changed based on the bias; with the empathy gap, for example, the robot may be overly joyous or overly sad to show the bias effects (hot-cold empathy). The lines of speech can therefore be, (1) ‘Hello my friend! I am very happy to see you today. It’s such a beautiful day. I hope you are feeling great today’, or (2) ‘Hi. Today I am not feeling very good.’ In the next section, the development of each of the algorithms is discussed.

4.2.3 Algorithms

The algorithms were designed based on the main components of the selected cognitive biases; for example, to develop the robot’s lack of knowledge or ideocracy (similar to the Dunning-Kruger effects bias) three main behaviours were selected for the robot: unable to understand own lack of knowledge, unable to understand other’s true knowledge, and understand and acknowledge own lack of knowledge. Lack of knowledge or ideocracy effects were developed in MARC to test them in topics-based conversational interaction. Therefore, MARC asks various questions to participant on different topics and makes arguments based on false facts, at the end acknowledging its stupidity.

The algorithms consist of lines of dialogue and actions performed by the robots. The dialogue was created based on the main components and effects of the selected cognitive biases, and the actions were created based on the robot’s functions and features, as shown at the beginning and at end of the conversation. For example, the misattribution algorithm was developed in ERWIN, and the algorithm included dialogue and emotional expressions. In general, the dialogue and actions cues were as presented in Table 4.2.

Table 4.2: Speech and Action cues of ERWIN used in experiments



➔ **Start**	*Action cue*	
	Speech (1)	Greets
	Expression (1)	Happy
Wait for response (1)		
Receive the response (1)		
	Expression (2)	According to the response
	Speech (2)	According to the response (1)
	Expression (3)	According to the speech (2)
Wait for response (2)		

From the above Table (4.2), it can be seen that the robot performs an action while speaking. Also, the robot shows an appropriate action after each line of dialogue finishes. For example, to ask a question to the participant, the process involves action (i.e. expression), dialogue (i.e. question) and action (i.e. expression); these expressions were chosen to be relevant to the conversation, to indicate that the robot wants to know the answer from the participants.

4.2.3.1 Dialogue Design

In both conversational and game-playing interactions, the robot interacts verbally based on the different biases or baseline, so dialogue had major impacts in each of the interactions. In the conversational interactions, dialogue was created based on the topics and with the effects of the chosen biases in our experiments (e.g. misattribution, self-serving, humours effects etc.). The conversational interactions were question-answer based, where the robot asks questions based on pre-selected topics. In this case, lines of speech came in two types:

a. questions, and

b. statements.

In most cases, questions were made direct and to the point, expecting a direct answer from the participants (e.g. Do you like football?), so that the robot could express the statements containing the selected bias's effects. The bias effects were designed in the statement, usually in response to the participant's answer (e.g. 'A Canadian invented football on ice, and it's called Rugby', showcasing the lack of knowledge or ideocracy effect). However, such design also depends on the selected bias, which sometimes changes the form of questions and statement; for instance, in the misattribution-biased algorithm, robot asks question misattributing previous information (e.g. ‘Last time you said you love to study medical journals, am I correct?’). However, in the case of non-biased algorithms, statements were made very simple so as not to show any effects of the selected biases (e.g. ‘Okay, I understand’).

In the game-playing algorithms, dialogues were based on the effects of the selected biases. In these interactions, the questions enquire about the different conditions of the game playing (e.g. ‘Do you understand the rule?’ and ‘Are you ready?’) In such algorithms, the robot expresses the biased effects through the statements delivered after the game and depending on the results of the games (e.g. ‘Sorry, I was not ready, I did not lose’, demonstrating the self-serving bias).

Now will I discuss each of the algorithms developed for the experiments.

4.2.3.2 Misattribution algorithm

The misattribution-biased algorithm was designed based on the main characteristics of the bias, misattributing and forgetting previously collected information. To reflect the effects of such biased characteristics in conversation, the algorithm was designed based on three main components of the bias, including false memory, source confusion and total forgetfulness (Figure 4.11). In the conversation, the robot expresses its misattribution bias effects in dialogue based on these three components.

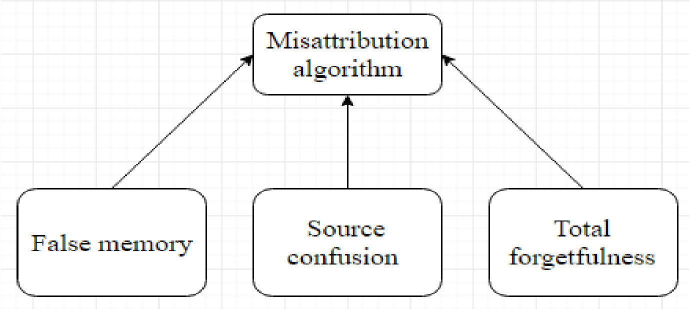


Figure 4.11: Misattribution algorithm was created based on three components of the bias.

The effects of these components were expressed in all three stages of an interaction. For example, in the meet-and-greet stage, ERWIN forgets the participant’s name, becomes confused about the participant’s address, and erroneously states the participant’s previous wellbeing. These statements were made in dialogue (see Table 4.3).

Table 4.3: Examples of dialogues with misattribution bias.

Examples of dialogue		Misattribution components used	Actions
1	‘Hello Dave. It is nice to see you again.’	Forgetfulness	Wait for response
2	‘I remember that last time you said your name is Dave.’	False memory	Wait for response
3	‘It must be someone else who looks like you.’	Source of confusion	End of topic ‘name’, on to the next topic

The robot forgets the participant’s name and misattributes someone else’s in its first dialogue. Then, the second dialogue shows the robot’s false memory, where it misattributes what participant said in a previous interaction. In the third dialogue, the robot gets confused about the source of the name ‘Dave’. All these three dialogues show the robot’s misattribution behaviours in conversation. In the following sections, the three stages of misattribution interactions are shown. The Robot waits for participant’s responses after each line of dialogue, and at the end of the third statement it moves to the next topic. However, if the participant does not respond after the robot delivers its second line of dialogue, then the robot moves to the next topic.

Meet and greet

In the misattributed interaction, the robot starts misattributing previously collected information such as names, address. In this stage, the robot misattributes information such as the participant’s name, address and previous state of wellbeing. Figure 4.12 outlines this stage.

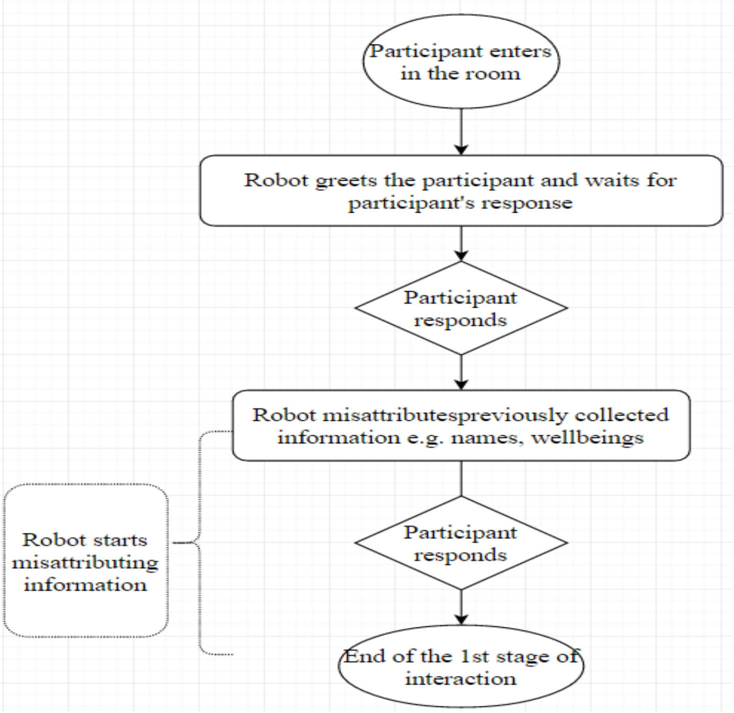

Participant enters in the room
Robot greets the participant and waits for participant's response
Participant responds
Robot misattributespreviously collected information e.g. names, wellbeings
Robot starts misattributing information
Participant responds
End of the 1st stage of interaction

Figure 4.12: Meet and greet stage for misattribution interactions

Topic-based conversation

In the misattribution interaction, the robot continues misattributing information following the misattribution algorithm. The robot starts the conversation by enquiring about previous topics and misattributes the participant's information; for example, 'Last time you said you love Beethoven, am I correct?', but the participant did not say that he or she loved Beethoven in the first place. In this case, the robot expresses sad emotions for incorrectly remembering information. Figure 4.13 shows the outline of this stage for the misattribution algorithm.

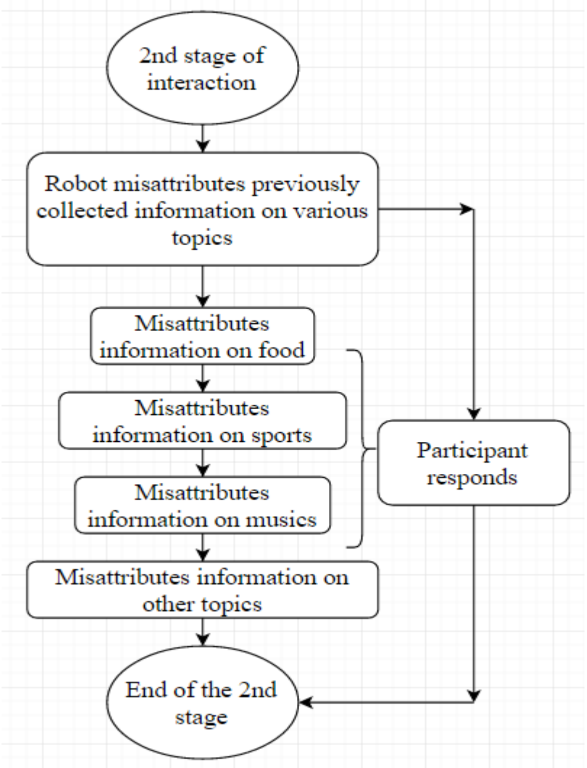

2nd stage of interaction
Robot misattributes previously collected information on various topics
Misattributes information on food
Misattributes information on sports
Misattributes information on musics
Misattributes information on other topics
Participant responds
End of the 2nd stage

Figure 4.13: Topic-based conversation stage for misattribution algorithm

Farewell

At this stage of the misattribution interaction, the robot thanks the participants for coming and chatting. The robot also requests that participants to fill out the questionnaires, similar to the introductory and non-misattributed interactions. Between each of the dialogues the robot waits for responses from the participant. Figure 4.14 shows the outline of the farewell stage of the misattribution interaction.

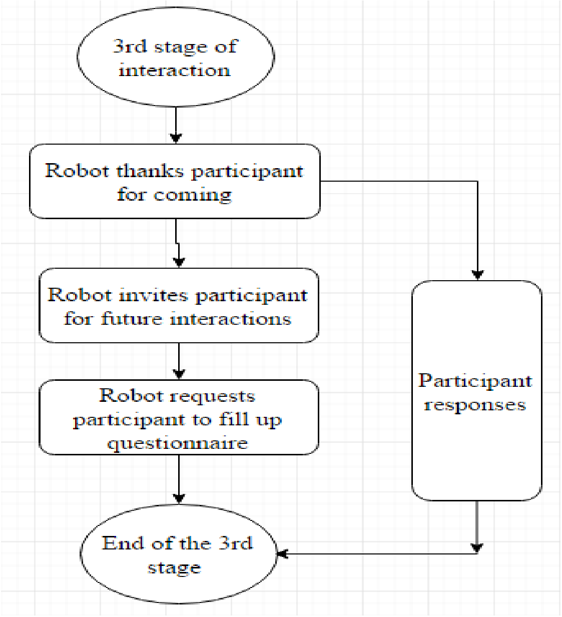

3rd stage of interaction
Robot thanks participant for coming
Robot invites participant for future interactions
Robot requests participant to fill up questionnaire
End of the 3rd stage
Participant responses

Figure 4.14: Farewell stage for misattribution algorithm

4.2.3.3 Developing the empathy gap algorithm

The empathy gap opens when someone is unable to understand another’s emotional state. In the experiments, I followed the principle of the empathy gap bias to develop a game-playing and a conversational algorithm for the robots. The game-playing algorithm was created for non-conversational interaction between the robot MyKeepon and participants, and the conversational algorithm was developed for the robot MARC.

As the interactions in my experiments lasted 10–12 minutes, at the beginning it was challenging to show the such differences in the emotional states of the robots. After several test interactions between the robot and my colleagues and after discussing the matter with my supervisor, I developed the algorithms based on two types of robot behaviours: showing excess happiness (i.e. hot-sate of empathy) and excess sadness (cold-state of empathy). The behaviours shown by the robots in the experiments to express such exaggerated expressions are particular to the robots. For example, in the game-playing interaction with MyKeepon, the robot shows hot and cold states of empathy gaps as recorded in Table 4.4.

Table 4.4: Empathy gap algorithm for the MyKeepon interactions experiment

		Actions	**Noises**
MyKeepon	***Hot state of empathy***	Maximum responses by frequently jumping, dancing, body moves, gazing	Happy noises, even if the participant is quiet
	Cold state of empathy	Minimum responses by less jumps, not dancing, less body movement and not making eye contact or by looking to the sides and sometimes being totally unresponsive	Sad noises, very quiet

Similarly, the conversational algorithm was designed based on hot states and cold states of empathy gaps. In these cases, the robot makes differences of emotional understanding with participant by expressing overly joyous or overly sad behaviours, with related dialogue. In the hot state interaction, the robot responds more actively than in the cold state. Lines of dialogue were also created to express the two components of the bias (see Table 4.5).

Table 4.5: Empathy gap algorithm for MARC interactions experiment

		Actions	**Dialogue**
MARC	***Hot state of empathy***	Faster responses, and happy and much friendlier approaches	‘Hello! It is great to see you again. I am very happy that you have come to chat with me.’
	Cold state of empathy	Minimum responses and slower movements, sometime no movement at all	‘Hi.’

The conversational structure follows the overall biased algorithm’s conversation, as shown in Figure 4.10, where the participants had the option to reply to the robot’s statement, and the robot could respond to that as well.

In the experiments, for the ‘hot’ state of the empathy, MARC starts to show great excitement from the beginning of interaction, when the participant enters in the room, and it responds to participants cheerfully. In the cold state, the robot itself exhibits sadness and does not talk much. In the experiment, such stages were randomly chosen for the participants.

Hot state of empathy gap

In the ‘hot’ state of the empathy gap, MARC remains happy at all times, and its responses are very cheerful, regardless of the participant’s responses. If the participant asks the reason for this happiness, the robot doesn’t specify any reason, which indicates that the robot does not care about what participant says, and it is busy to expressing its own emotions. Taking the baseline interaction as an outline, the differences among the stages of the hot state of empathy gap bias behaviour are shown in the Figure 4.15.

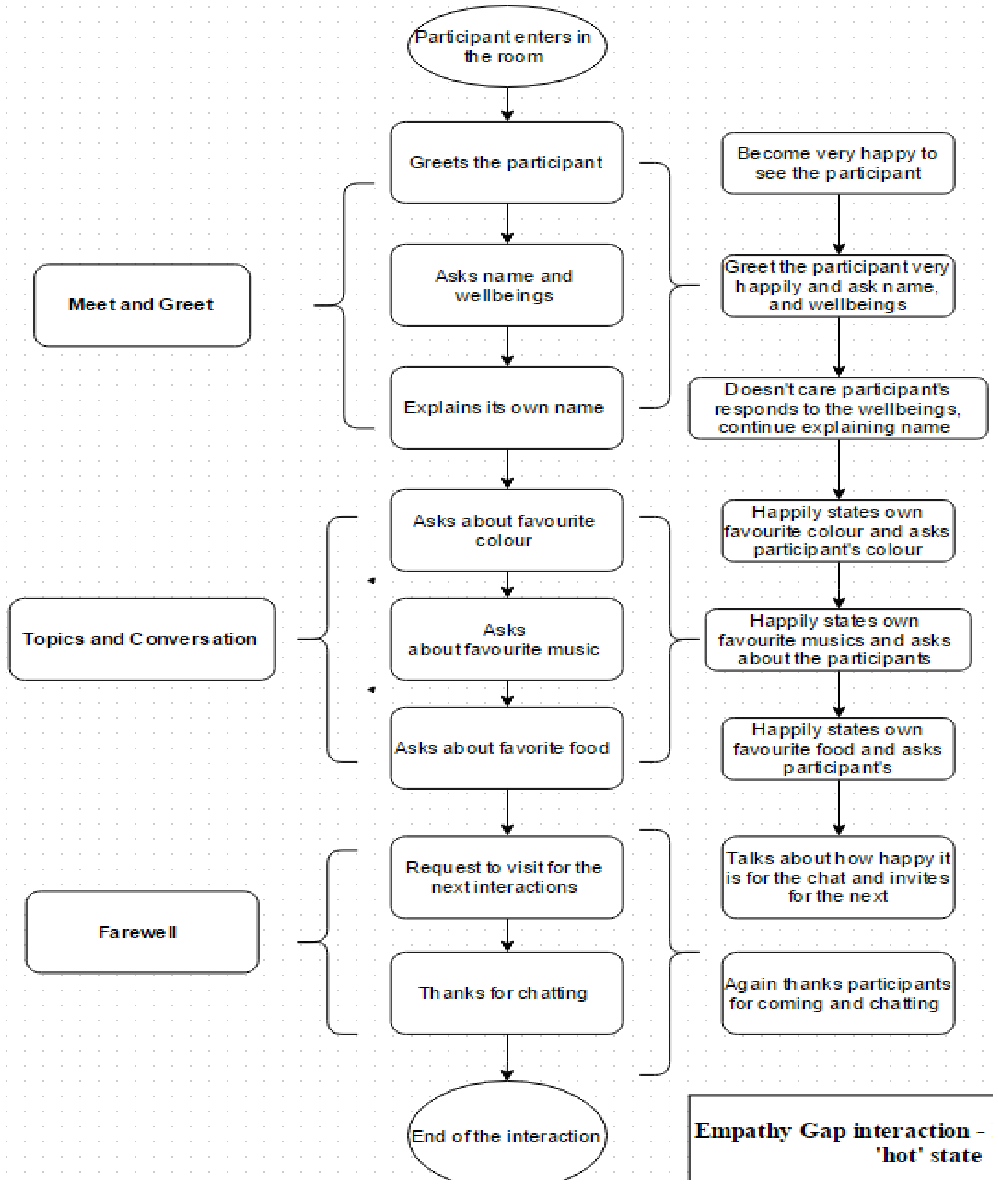

Participant enters in the room
Greets the participant
Asks name and wellbeings
Explains its own name
Asks about favourite colour
Asks about favourite music
Asks about favorite food
Request to visit for the next interactions
Thanks for chatting
End of the interaction
Meet and Greet
Topics and Conversation
Farewell
Become very happy to see the participant
Greet the participant very happily and ask name, and wellbeings
Doesn't care participant's responds to the wellbeings, continue explaining name
Happily states own favourite colour and asks participant's colour
Happily states own favourite musics and asks about the participants
Happily states own favourite food and asks participant's
Talks about how happy it is for the chat and invites for the next
Again thanks participants for coming and chatting
Empathy Gap interaction - 'hot' state

Figure 4.15. Differences in stages in empathy gap algorithm compared to baseline

Cold state of empathy gap

In the cold state of the empathy gap, MARC interacts with the opposite behaviours to exhibited in the hot state interaction. Throughout the interaction, it stays quiet, giving brief answers and remaining unwilling to talk much or express the differences in emotion to the participant.

Based on the non-biased interactions, the differences with the cold state of empathy gap bias are shown in the Figure 4.16.

The robot here just initiates the conversation and usually waits for the participants to respond, and then continues. If the participant says nothing, the robot moves to a later line of dialogue. Although the interaction structure is similar to other interactions, which means the robot talks about all topics, but it does not offer much in the way of response.

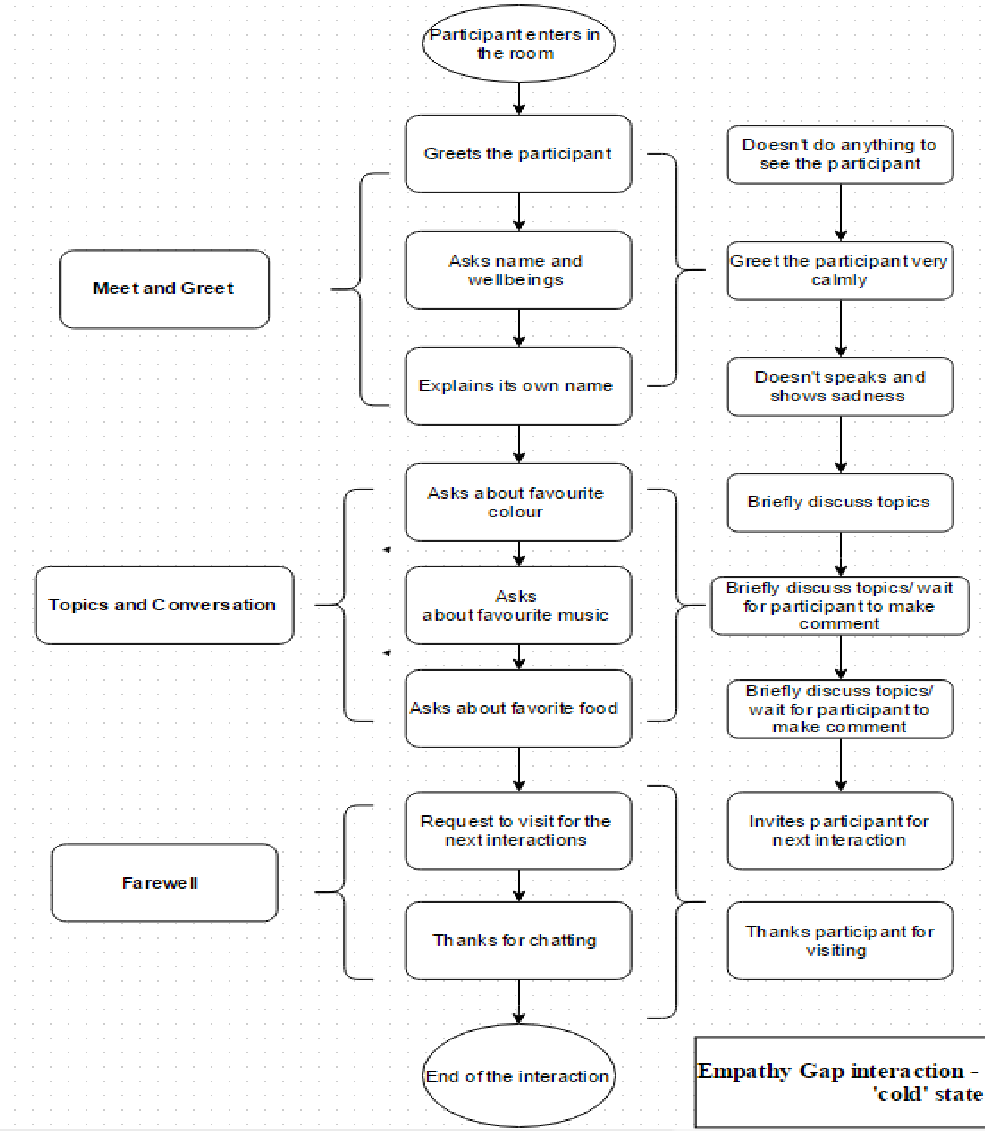

Participant enters in the room
Greets the participant
Asks name and wellbeings
Explains its own name
Meet and Greet
Doesn't do anything to see the participant
Greet the participant very calmly
Doesn't speaks and shows sadness
Asks about favourite colour
Asks about favourite music
Asks about favorite food
Topics and Conversation
Briefly discuss topics
Briefly discuss topics/ wait for participant to make comment
Briefly discuss topics/ wait for participant to make comment
Request to visit for the next interactions
Thanks for chatting
Farewell
Invites participant for next interaction
Thanks participant for visiting
End of the interaction
Empathy Gap interaction - 'cold' state

Figure 4.16. Differences of ‘cold’ state of empathy gap compared to baseline

4.2.3.4 Development of lack of knowledge (ideocracy effects) algorithm

As discussed earlier, the lack of knowledge (i.e. ideocracy) effect was developed based on the three main components of the Dunning-Kruger effect bias: the robot fails to recognise its own lack of knowledge, fails to recognise genuine skill or knowledge in others, and recognises and acknowledges its own lack of skill after being exposed. Such components were developed individually in each of the interactions, as shown in the Figure 4.15.

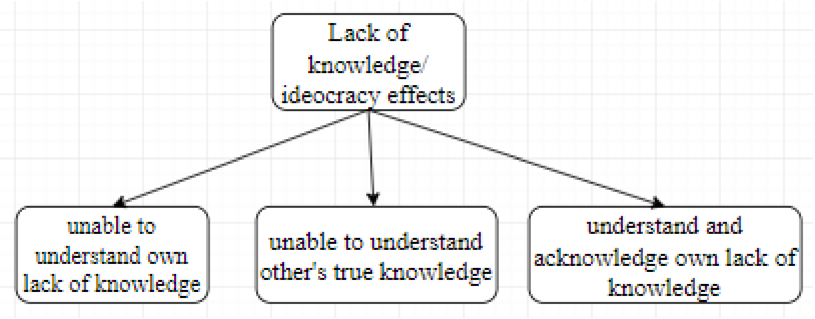

Lack of knowledge/ ideocracy effects
unable to understand own lack of knowledge
unable to understand other's true knowledge
understand and acknowledge own lack of knowledge

Figure 4.17. Three interactions of lack of knowledge or ideocracy algorithms

In these interactions, the dialogue structure is the same as the others biased dialogue (Figure 4.10), where the robot gets to respond after the participant's answers. In the interactions based on lack of knowledge or ideocracy behaviours, the robot always tries to convince the participants that whatever it says is correct and the participant is wrong. If the participant does not argue with the robot, then it moves to the next line of dialogue. An example of the dialogue in this interaction is shown in the Table 4.6.

Table 4.6: Example of dialogue from the lack of knowledge or ideocracy effects interaction

Examples of dialogue		**Lack of knowledge/ideocracy (based on Dunning-Kruger effects) components used**	**Actions**
1	'What type of music is your favourite?'		Waits for response
1	'Oh yes! I have listen to them. They are rock band from 60s.'	Unable to understand other's true knowledge	Wait for response
2	'No, you are wrong. They are actually from 40s as I remember.'	Unable to understand own lack of knowledge	Wait for response and participant reply, then move to the 3rd
3.	'Okay. May be I am wrong.'		Move to the next topic

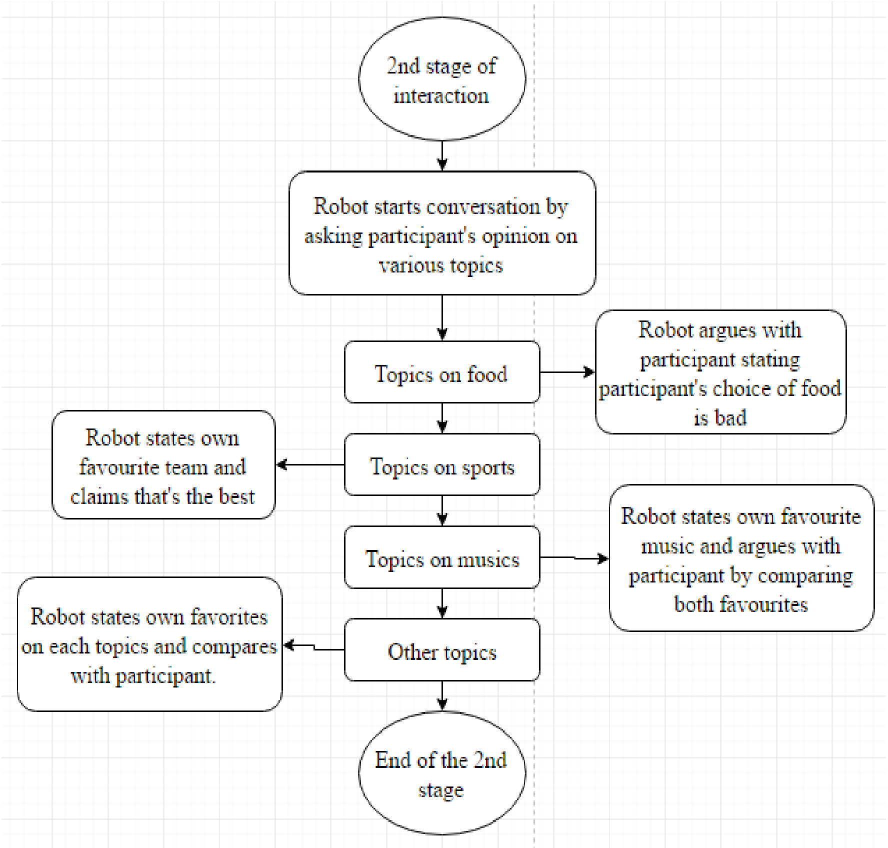

2nd stage of interaction
Robot starts conversation by asking participant's opinion on various topics
Topics on food
Robot argues with participant stating participant's choice of food is bad
Topics on sports
Robot states own favourite team and claims that's the best
Topics on musics
Robot states own favourite music and argues with participant by comparing both favourites
Other topics
Robot states own favorites on each topics and compares with participant.
End of the 2nd stage

Figure 4.18: The topic-based conversation stage of the Lack of knowledge/ideocracy effects biased interaction

The robot asks a question, waits for a response, and whatever response participants gives, the robot states an incorrect opinion and begins arguing with participants. After few arguments, the robot gives up and acknowledges its mistakes.

Meet and greet

Similar to the other algorithm's meet-and-greet stage, in this stage robot greets the participant and asks about the participant's name and wellbeing. This is the initial stage where robot and participant get acquainted. The robot initiates the biased effects in the next stage of the interaction.

Topic-based conversation

In this stage, the participant and MARC discuss various topics. The robot idiotically states wrong information about the topic, and furthermore tries to convince the participant that its information is correct. In this case, the robot does not necessarily correct the participant, but it says that the participant is wrong. If the participant tries to correct the robot, the robot first denies its mistake, but in the end understands it.

For every topic the robot states its own choice and idiotically claims it is the best.

Farewell

In the final stage of the interaction, the robot understands its mistakes and acknowledges them to the participant. The robot even apologies, if necessary. An outline of the farewell stage of the lack of knowledge or ideocracy effects biased interaction is shown in the Figure 4.19. The interaction ends with a request from robot to the participant to fill out the questionnaire and an invitation for the next interaction.

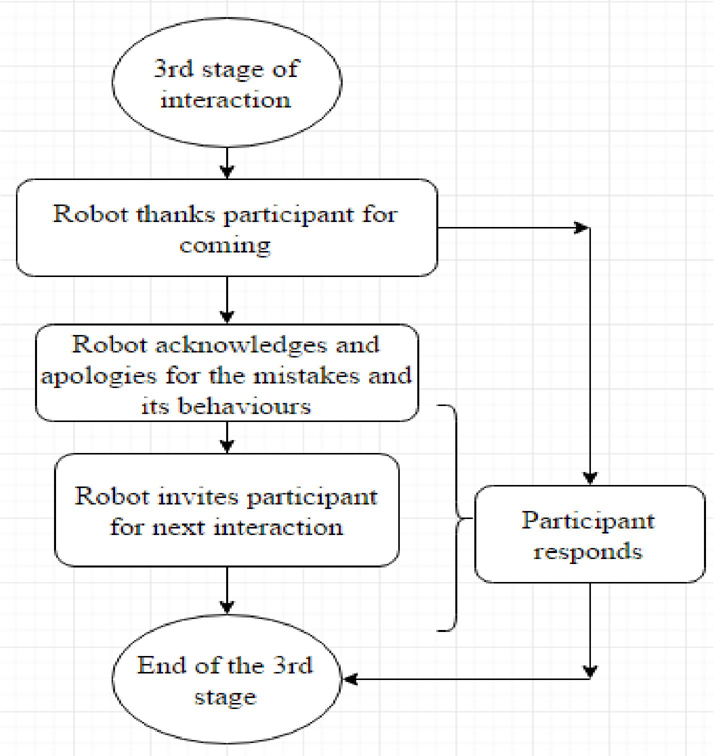

3rd stage of interaction
Robot thanks participant for coming
Robot acknowledges and apologies for the mistakes and its behaviours
Robot invites participant for next interaction
Participant responds
End of the 3rd stage

Figure 4.19. The third interaction of in the lack of knowledge or ideocracy effect algorithm, compared to the baseline

4.2.3.5 Development of self-serving bias algorithm

As discussed in the previous chapter, self-serving bias causes people to attribute positive events to themselves but negative events to other external factors. In this experiment, the key differences between the self-serving algorithm and other algorithms is that the robot credits itself for winning a round of the game against the participants but blames other factors for losing. In this case, such supposed external causes include that the robot was not ready, the robot was not looking at the game play or something got into the robot's eyes. If the participant does not agree with the robot, it tries to convince the participant to play again. In such cases, the robot sometimes blames the participants for cheating in games.

As this bias was tested in game-playing interaction, the interaction was divided into five stages. In the meet-and-greet stage, the robot welcomes the participant, asks about the participant's wellbeing and explains the meaning of its name.

Meet and greet

In this stage, the robot meets the participant, enquires about the participant's name and wellbeing, and explains its own name, as usual. No biased effect occurs in this stage.

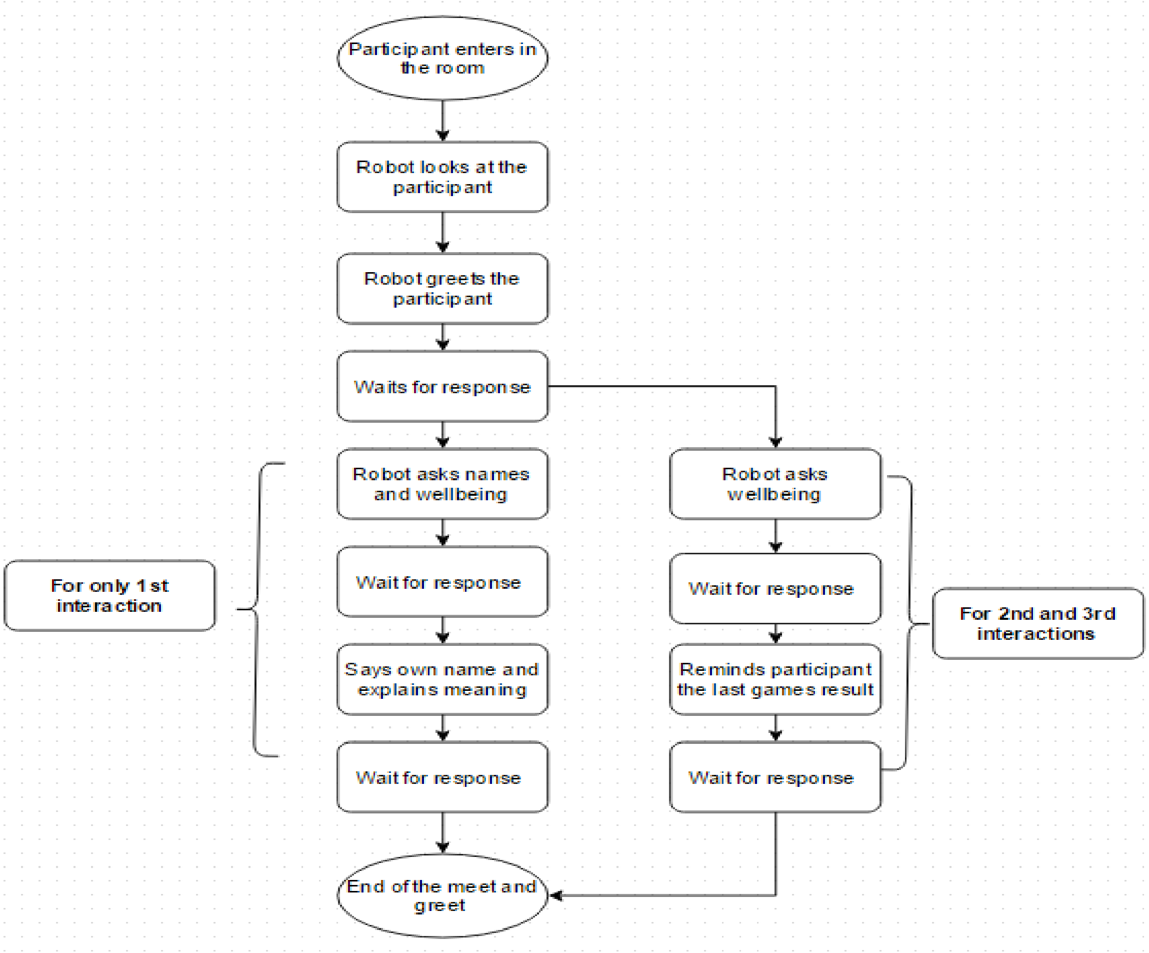

Participant enters in the room
Robot looks at the participant
Robot greets the participant
Waits for response
Robot asks names and wellbeing
Wait for response
Says own name and explains meaning
Wait for response
For only 1st interaction
Robot asks wellbeing
Wait for response
Reminds participant the last games result
Wait for response
For 2nd and 3rd interactions
End of the meet and greet

Figure: 4.20. Structural diagram of the meet and greet stage

The robot takes the initiative in the communication, so that the participant can be comfortable for the rest of the interaction. Based on the interaction number, the conversation proceeds to the next stage, where the robot explains the game's rules.

Rules explanation

In case of self-serving bias, the robot explains the rule, but if the participant asks the robot to repeat the rules, it then triggers its biased algorithm, which results in repeating the rules in such way that encourages the participant to think the robot is claiming to be more intelligent than the participant (Figure 4.21). In this case, the steps are as follows:

a. Tell the name of the game and ask participant if he or she is familiar with it.

b. Explain the rules, and ask if the participant understands the rules.

c. If the participants id not understand the first time, repeat the rules.

   I. Asks the participant why he or she did not understand the first time (stating that it was loud and clear enough).

   II. States that it can repeat the statement endless times and in endless ways, whatever the participant prefers.

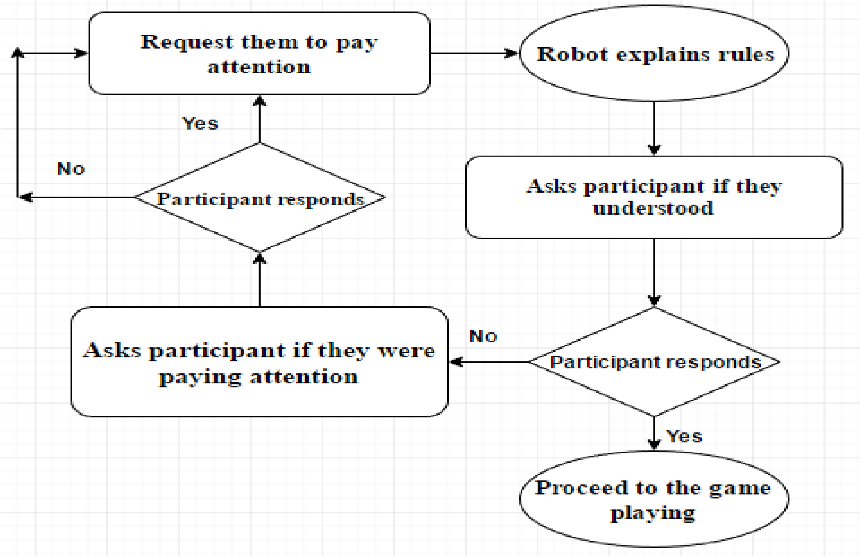

Request them to pay attention
Robot explains rules
Yes
No
Participant responds
Asks participant if they understood
No
Asks participant if they were paying attention
Participant responds
Yes
Proceed to the game playing

Figure 4.21. Structural diagram of the rule-explanation stage for biased algorithms

Game playing

In this stage, the robot responds differently for different outcomes in games. The outline of this stage is shown in Figure 4.22.

Robot wins

For the self-serving bias, if the robot wins a hand, it expresses its joy in such a way that it won the hand due to its own skills, so that it knew that the participant was going to play that particular option, and it is not a matter of chance, but the robot solved it through its own intelligence. As it is a friendly robot, it gives few tips to the participant for winning the next hand. Giving the tips and related actions expresses its over-confidence and self-serving intelligence, and as such, it acts as if it is going to win all the remaining hands against participants.

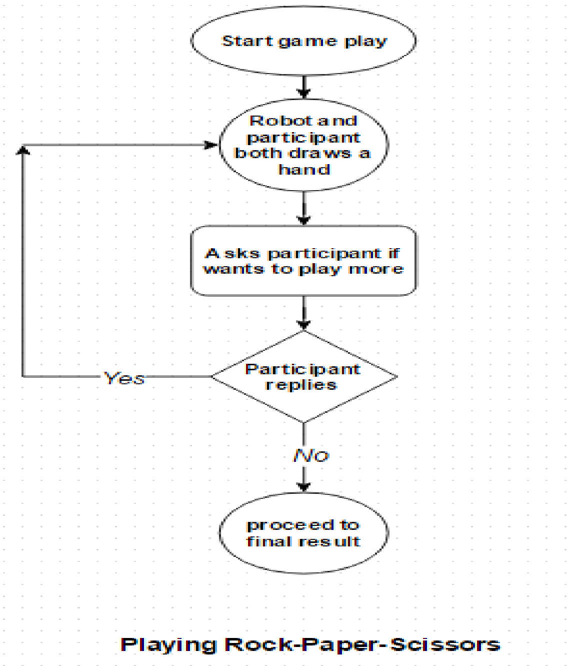

Start game play
Robot and participant both draws a hand
Asks participant if wants to play more
Participant replies
Yes
No
proceed to final result
Playing Rock-Paper-Scissors

Figure 4.22. Structural diagram of the game-playing stage

In this case, the steps are as follows:

i. Ask participant if he or she is ready to play.

ii. Draw a hand.

iii. Get excited about winning.

   a. Brags about winning.

   b. Gives tips for winning to the participant.

iv. Request participant play more.

<u>*Robot loses*</u>

For the self-serving bias, the robot does not easily accept losing the games. It tries to find reasons to blame the other for losing, such as the surroundings and even the participant. Rather than simply admitting the fact it lost the round, the robot makes excuses and lays false blame:, ‘I was not ready’ or ‘You must have cheated’. As it is a friendly robot, such arguments are limited and mostly end up with a challenge to win the next hand. Such interaction steps confirm that despite the robot being the victim of self-serving bias, it still wants to keep playing with the participant. For this biased interaction, the steps are as follows:

i. Ask participant if he or she is ready.

ii. Draw a hand.

iii. Show sad expressions after losing.

  a. Give excuses.

  b. Blame various factors.

  c. Ultimately blame the participant.

iv. Challenge the participant to play more.

v. If losing continuously, give up the game by blaming others and showing various excuses

<u>*Draws*</u>

The robot tries to deny draws as well, asking the participant to play again. An outline of the game-playing structure is shown in Figure 4.23.

Final results

The self-serving robot expresses its over-confidence in winning and blames others for losing, and does not accept its defeat. In this case, the modified steps would be as follows:

i. Calculate the results.

ii. Congratulate participant if they won.

  a. States that it let the participant win, or

b. blames others, give excuses, or

c. ultimately, blames participant for cheating.

d. Challenge for the next time

iii. Congratulates itself if the robot won.

a. States its over confidence and superiority.

b. Challenges for next time.

.

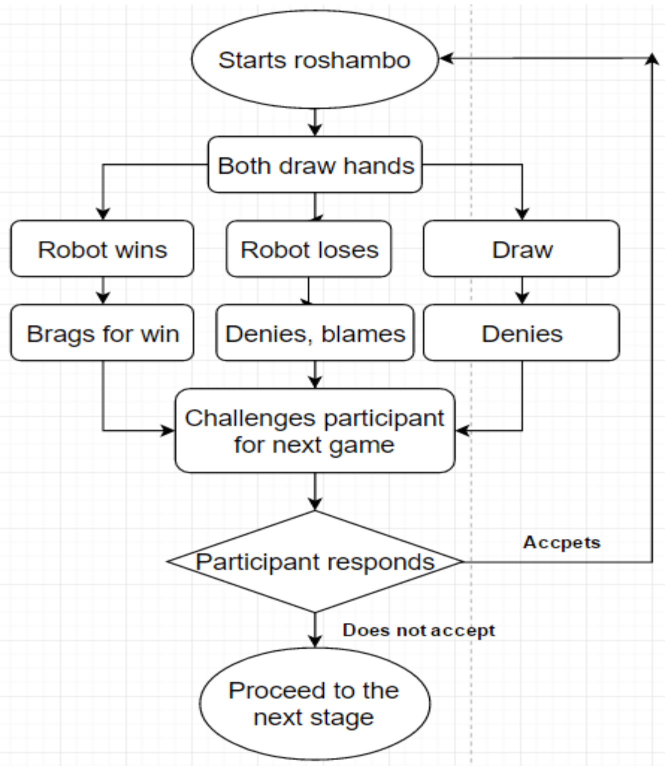


Figure 4.23. Outline of the game playing stage with outcomes for the self-serving algorithm

Farewell

The self-serving biased robot thanks participants for coming and invites them for future interactions, if the robot wins the games. In case the robot lost the game, it poses a friendly challenge for the next game. In both cases however, it thanks the participants for playing the game. In this case, the steps would be as follows:

i. Thank participant for coming.

a. State that playing was a nice experience (if robot wins the game).

b. State that playing was not a very good experience (if it loses the game).

ii. Shake hands.

iii. Invite participant for another interaction.

a. Happily invites participant for this interaction (if won).

b. Challenge participant to play again in another interaction (if loses).

The Figure 4.24 outlines the entire self-serving-bias interaction, step by step.

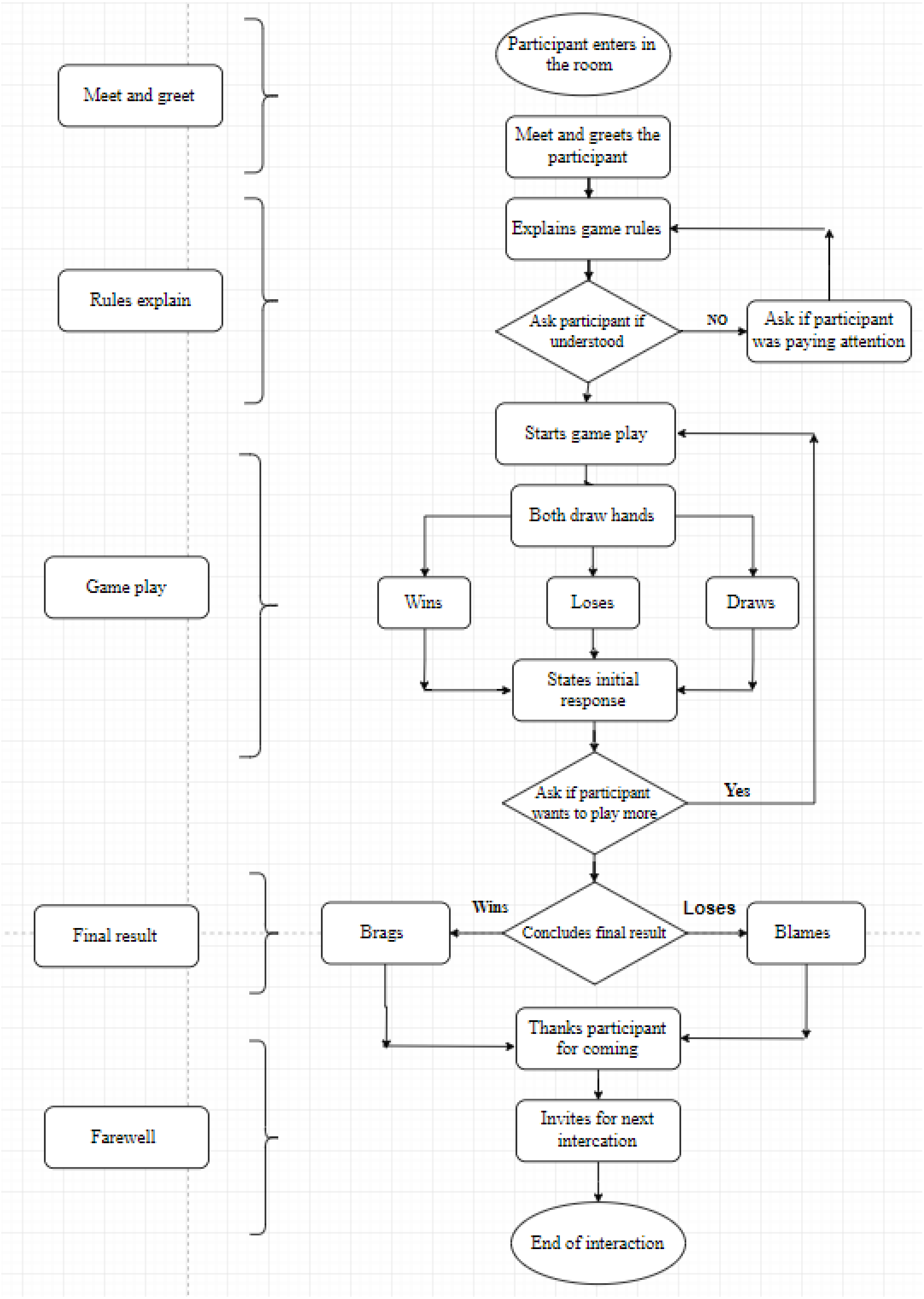

Meet and greet
Rules explain
Game play
Final result
Farewell
Participant enters in the room
Meet and greets the participant
Explains game rules
Ask participant if understood
NO
Ask if participant was paying attention
Starts game play
Both draw hands
Wins
Loses
Draws
States initial response
Ask if participant wants to play more
Yes
Wins
Loses
Brags
Concludes final result
Blames
Thanks participant for coming
Invites for next intercation
End of interaction

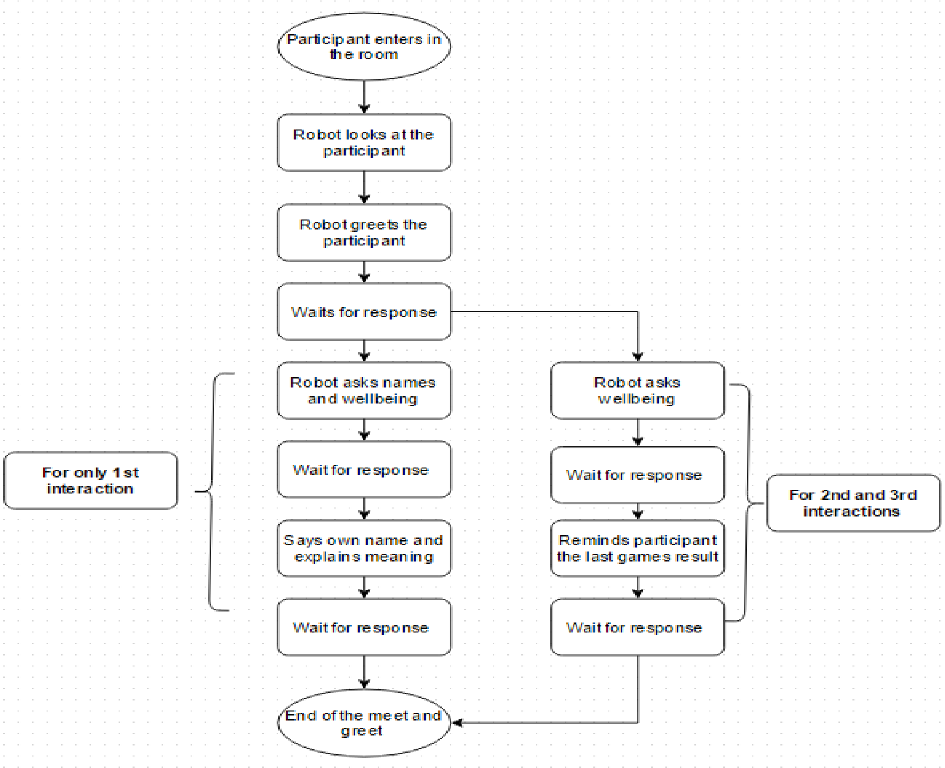
Figure 4.24. Outline diagram of all stages in self-serving biased algorithm

4.2.3.6 Development of humour effect bias algorithm

In the case of the humour effect, the robot's main goal is to humorously interact with participants, so that the memory of the interactions remains with the participants for a long period of time. There are, however, several types of humour, and understanding of humour can depend on an individual's perception. In our case, the robot delivered subjective jokes on losing and winning games. Some of the dialogue was taken from online funny speech bots (e.g. A.L.I.C.E (Shawar B A, Atwell E, 2015); Elbot, n.d.).

Meet and greet

The first stage of the interaction is similar, with all biased and baseline algorithms. In case of humour effect bias, the robot follows actions similar to those of previous interactions. For the first interaction, the robot asks for the name and wellbeing of the participant, but in the following interactions, the robot skips the name-asking part. Figure 4.25 shows a structural diagram of the meet-and-greet stage.



Participant enters in the room
Robot looks at the participant
Robot greets the participant
Waits for response
Robot asks names and wellbeing
Wait for response
Says own name and explains meaning
Wait for response
Robot asks wellbeing
Wait for response
Reminds participant the last games result
Wait for response
For only 1st interaction
For 2nd and 3rd interactions
End of the meet and greet

Figure 4.25. Structural diagram of the meet-and-greet stage

Game explanations

Game explanations to the participants can be done in three steps:

a. Tell the name of the game and ask whether the participant is familiar with it.

b. Explain the rules and ask whether the participant has understood the rules.

c. If the participants did not understand the first time, repeat the rules.

In case of the humour effect, the robot completes the first step according to the baseline interaction, but for the second step, it adds funny elements in speech processing, so that the participant can remember the rules. If it needs to go for the third step, it repeats the rule, but adds more funny elements. Figure 4.26 shows an outline diagram of the rule explanation stage.

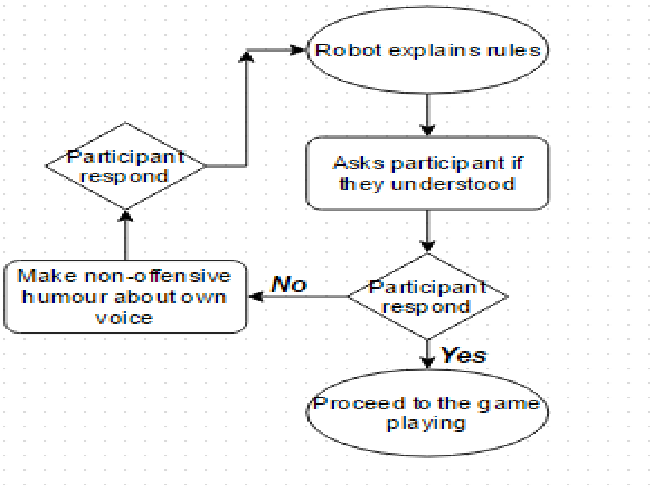

Robot explains rules
Asks participant if they understood
Participant respond
No
Make non-offensive humour about own voice
Participant respond
Yes
Proceed to the game playing

Figure 4.26. Structural diagram of the rule explanation stage

Game playing

The outline of this stage is shown in Figure 4.27. As with the other two game-playing algorithms, in this case the robot could win, lose and or draw in each round.

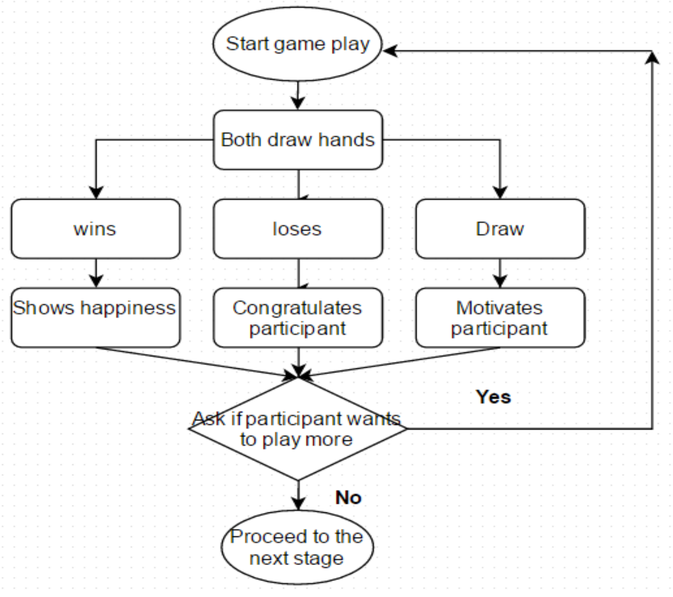

Start game play
Both draw hands
wins
loses
Draw
Shows happiness
Congratulates participant
Motivates participant
Yes
Ask if participant wants to play more
No
Proceed to the next stage

Figure 4.27. Structural diagram of game-playing stage in the humour effects algorithm.

Robot wins a hand

There are four steps in this phase:

i. Ask participant if he or she is ready.

ii. Draw a hand.

iii. Express excitement about winning and make a joke.

iv. Request that participant to play more.

In this scenario, winning is fun for the robot, and it expresses its joy in a friendly way. It also gives encouragement to the participant for the next hand. If the participant continuously loses, however, and does not want to play anymore, the robot tells funny stories of encouragement so that the participant keeps playing.

In this case, the steps are as follows:

i. Ask the participant whether he or she is ready to play.

ii. Draw a hand.

iii. Express happiness about winning.

  a. Gives tips for winning to the participant in funny ways.

iv. Requests that the participant play more.

Robot loses a game hand

The steps for this phase are supposed to be as before:

i. Ask a participant whether he or she is ready.

ii. Draw a hand.

iii. Show sadness and state a joke.

iv. Ask the participant to play more.

For the humour effect, losing is not very bad for the robot, and it expresses its defeat in a friendly way. The robot is trying to make the participants remember various moments of the interaction with the help of humour, so its winning or losing does not matter in this interaction. Also, it keeps encouraging the participant to play the next hand, but if the participant continuously wins and does not want to play anymore, the robot tells funny stories of encouragement so that the participant keeps playing. For the humour effect bias, the steps are as follows:

i. Ask participant whether he or she is ready.

ii. Draw a hand.

iii. Do not show sadness for losing, but

    a. States that it is happy that the participant won.

    b. States its determination to win the next hand.

iv. Ask participant to play more.

v. If losing continues, state its happiness about participant's wins, and

vi. keep expressing determination to the win next hand (keep playing).

If the robot and participant end up in a draw, then the robot tries to motivate the participant to play. In all the cases, the robot's responses are laced with humour.

Conclusion of the game

The games can conclude with the robot determining the overall winner. In this case, despite the result, the robot would try to continue its friendship with the participant. In this stage, the steps are as follows:

i. Calculate the results.

ii. Congratulate participants if they won, or

iii. congratulate itself if it won.

An outline with the three outcomes in the game-playing stage is shown in the Figure 4.28.

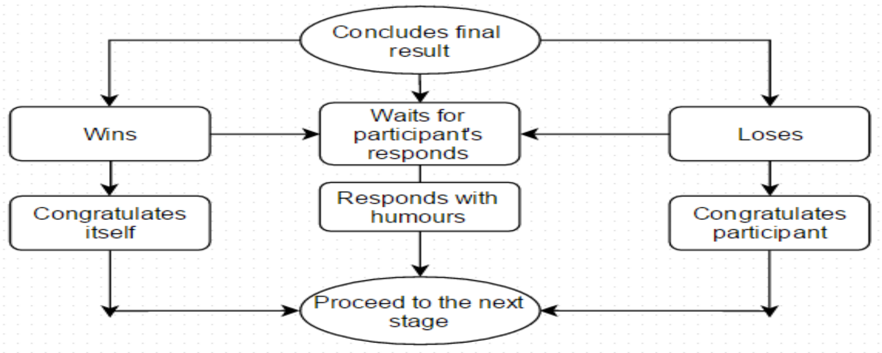

Concludes final result
Wins
Waits for participant's responds
Loses
Congratulates itself
Responds with humours
Congratulates participant
Proceed to the next stage

Figure 4.28. Structural diagram of the game playing stage with three outcomes in the humour effect algorithm

The humorous robot makes fun of both winning and losing, but greets and encourages participants for their winning and losing. In this case, the steps are as follows:

i. Calculate the results.

ii. Congratulate participants if they won.

  a. Say it let them win (in a funny way).

  b. Make joke of it.

  c. Claim to give the participant more of a challenge next time.

iii. Congratulate itself if the robot won.

  a. Make joke of it.

  b. Say that the participant let it win.

  c. Ask the participant to play again in another interaction.

In practice, the situation could go in any direction, but the robot is prepared for each possible moment with its humour.

End of the interactions

The end of the interaction is usually performed by the robot ending the dialogue, shaking hands and offering an invitation to subsequent experiments. In this case, the steps are as follows:

i. Thank participant for coming.

ii. Shake hands.

iii. Invite the participant for another interaction.

The humorous robot thanks the participants for coming and for playing, and it invites them for another interactions in a humorous way. It also promises the participants to say more funny things next time. The steps are the same as the baseline steps, only with the addition of funny speech.

The humour effect biased interaction is straightforward and similar to the baseline interaction, but with humour added to the conversation. The goal is discover whether the participant likes and prefers such friendly and humorous interaction with a robot.

The diagram of the humour effects algorithm is shown in Figure 4.29.

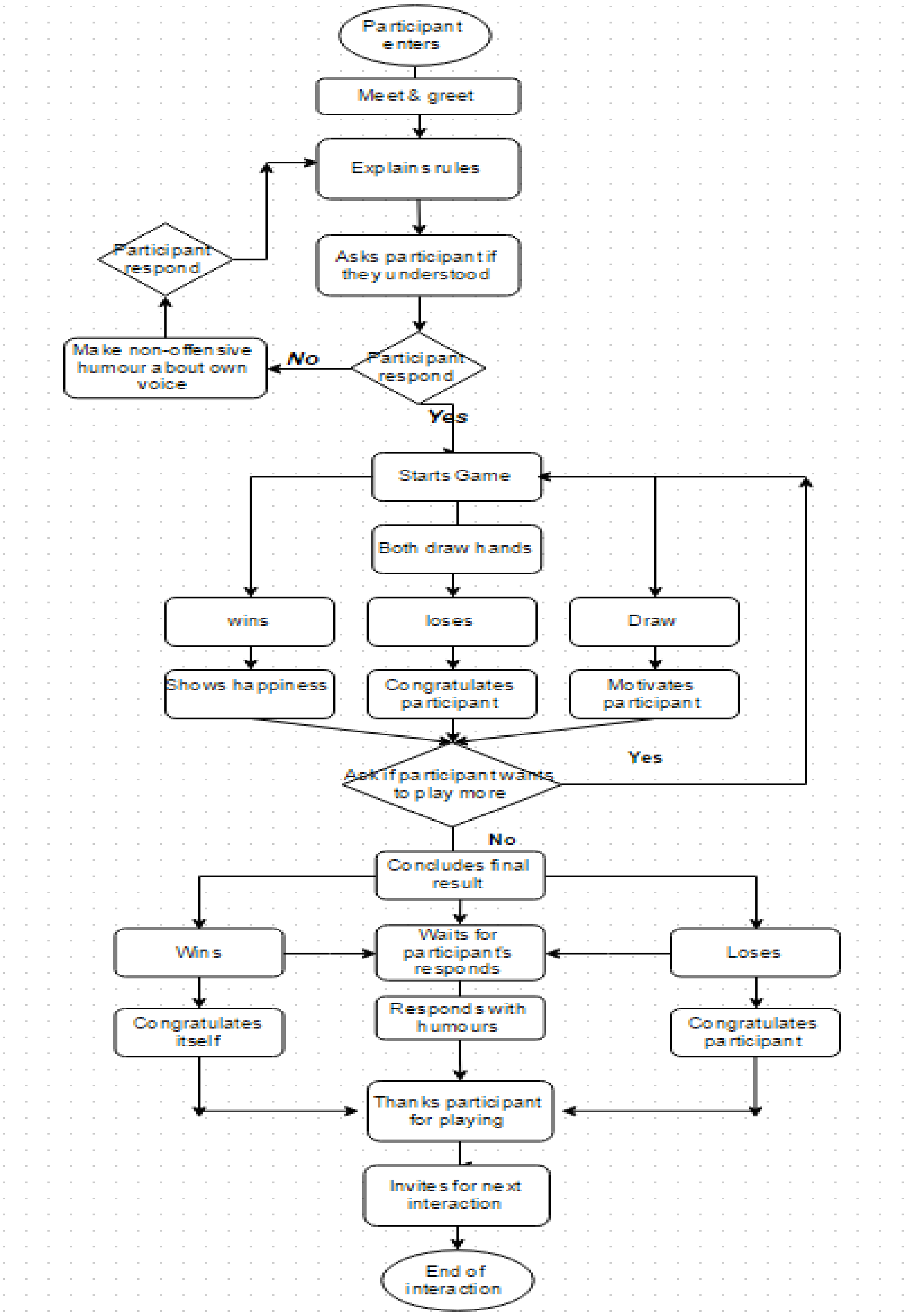


Figure 4.29. Structural diagram of all stages of humours effects algorithm.

4.2.3.7 Baseline algorithms

All the biased interactions were compared against a neutral interaction developed without the effects of the selected biases in the consecutive experiments. For example, to compare misattribution bias, a baseline algorithm was created by which the robot does not forget, and to compare empathy gap bias another baseline was created by which the robot shows appropriate expressions. Such baseline algorithms were developed for each experiment, followed by similar patterns of conversational and game-playing interactions.

Table 4.7: Examples of the non-misattribution dialogues

Examples of dialogues		Changes
1	'Hello **. It is nice to see you again.'	** states the name correctly
2	'I correctly remember your name.'	
3		*No need of the 3rd dialogue on name; robot moves to the 2nd topic*

Table 4.7 demonstrates that the robot states the name correctly in the 1st dialogue. If the robot states the name correctly, then there is no need for the second or third dialogue on the 'name' topic, so the robot moves to the next topic. Below I show the baselines for each category.

Conversational baseline algorithm

As stated earlier, a single interaction has three stages, which are meet and greet, topic-based conversation and farewell. Below, I discuss each of the stages of the non-misattribution algorithm.

Meet and greet

In the baseline (i.e. non-biased) interaction, the robot states all information (e.g. name, address) correctly from the collected information from the first introductory interaction. Figure 4.30 shows the outline of this stage.

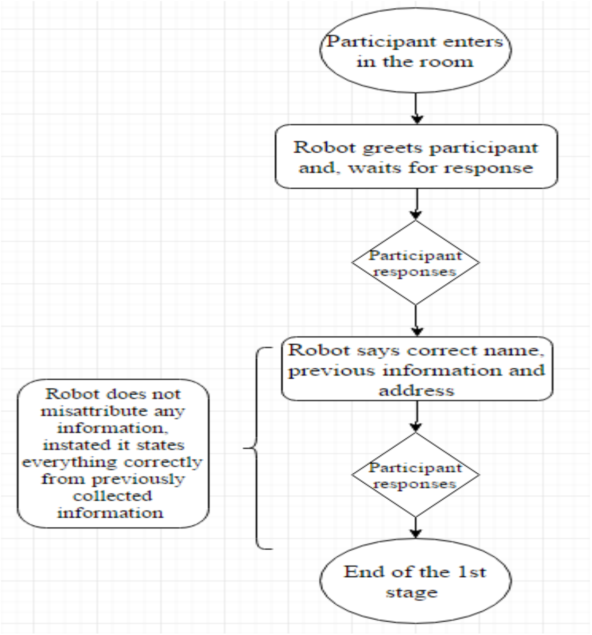

Participant enters in the room
Robot greets participant and, waits for response
Participant responses
Robot says correct name, previous information and address
Robot does not misattribute any information, instated it states everything correctly from previously collected information
Participant responses
End of the 1st stage

Figure 4.30: Meet and greet stage for non-biased interactions

Topic-based conversation

In the non-biased topic-based conversation, the robot states all information (e.g. information about music choice, food choices, etc.) in a very straight forward manner. The dialogue design is based on Figure 4.9 where the robot ends the topic by stating ‘Okay’ or ‘I understand’, and moves on to the next topic.

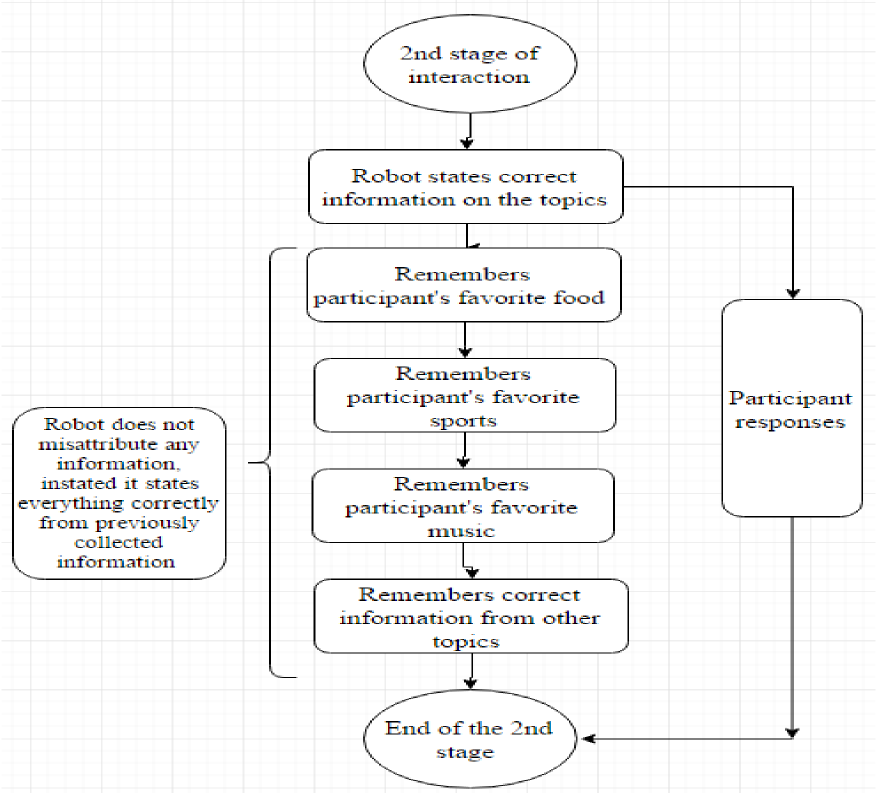

2nd stage of interaction
Robot states correct information on the topics
Remembers participant's favorite food
Remembers participant's favorite sports
Remembers participant's favorite music
Remembers correct information from other topics
End of the 2nd stage
Robot does not misattribute any information, instated it states everything correctly from previously collected information
Participant responses

Figure 4.31: Topic-based conversation algorithm outline for the non-misattribution interaction.

Farewell

In this stage, the robot thanks the participant for joining in the conversation extends an invitation to the next interaction. Figure 4.32 shows the outline of the farewell stage for the non-biased interaction. At the end of the interaction, all participants were asked to answer a set of questionnaires.

The conversational baseline (i.e. non-biased) interactions were made to be straightforward and mostly question-answer structured. The dialogue was developed carefully so that the corresponding bias would not be in effect. Such lines of dialogue were made by carrying out several mock interactions between colleagues, friends, students and the robot, and they were finalised by consulting with my supervisors so that the dialogues would be fit for the interactions.

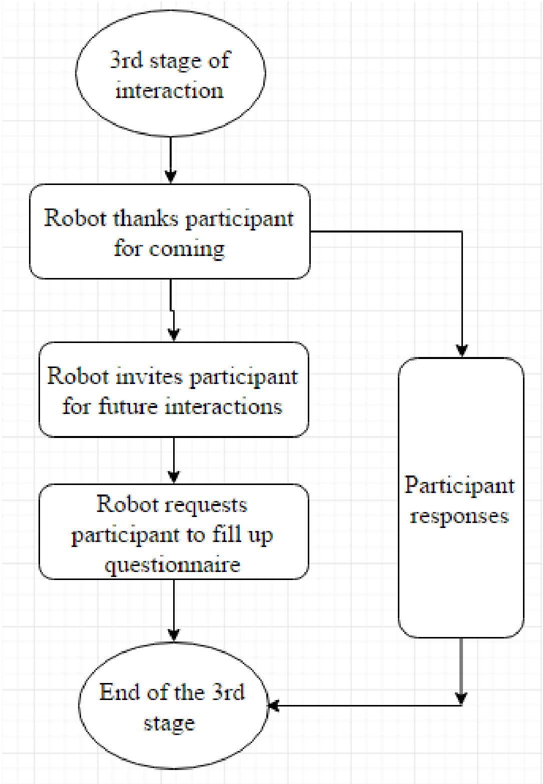

3rd stage of interaction
Robot thanks participant for coming
Robot invites participant for future interactions
Robot requests participant to fill up questionnaire
Participant responses
End of the 3rd stage

Figure 4.32: Interaction outline for the farewell stage of the non-misattribution interaction.

Game-playing baseline algorithm

The structural similarities between conversational and game-playing baseline algorithms were in the design of the dialogue, based on Figure 4.9. In both cases, the robot usually ends the current topic by answering ‘Okay’, ‘I understand’, ‘That’s good’ or something similar, and starts the next topic. In the game-playing baseline algorithms, there were five stages, including rule explanation and game playing, which were carefully designed so that the corresponding biases could not affect the baseline interactions. Similar to the other dialogue design, game-playing baseline dialogue was also designed vis a vis several mock interactions between colleagues, friends, students and the robot, and was finalised by consulting with my supervisors so that the dialogue would be fit for the interactions.

The meet and greet stage was similar to the previous conversational experiment. In the rule-explanation stage, the robot first explains the game’s rules (Figure 4.33), then asks the participant he or she has understood; if so, the robot proceeds to the game, and otherwise it repeats the rules.

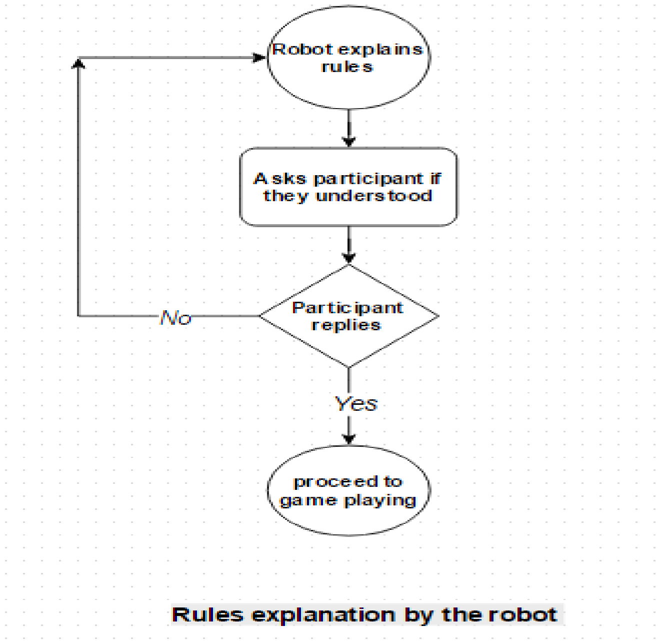


Figure 4.33. General structural diagram of game-rules explanation stage.

In the game playing, the robot maintains same conversational structure (Figure 4.34). The robot acts differently for each of the three possible outcomes of the game (Figure 4.34).

As seen in Figure 4.34, the robot and participant both draw a game hand, and if the robot wins, it says that it won; if the robot loses, it says that it lost and then asks the participant if he or she wants to play more. If the participant wants to continue, the robot starts the game again; otherwise, it proceeds to calculate the final results. The dialogue was made without the effects of the selected biases in this baseline interaction, despite the outcomes of the game. For example, the diagram below (Figure 4.35) describes MARC’s dialogue in the different outcomes.

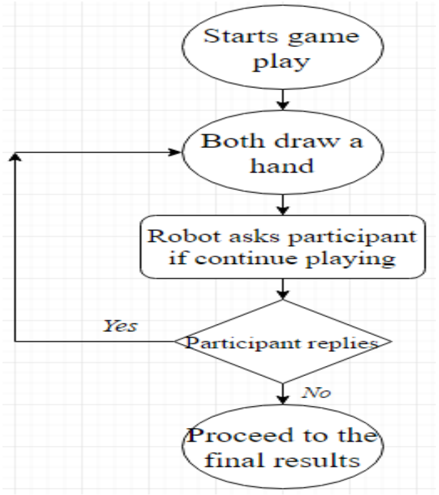

Starts game play
Both draw a hand
Robot asks participant if continue playing
Participant replies
Yes
No
Proceed to the final results

Figure 4.34. General structural diagram for game-playing stage.

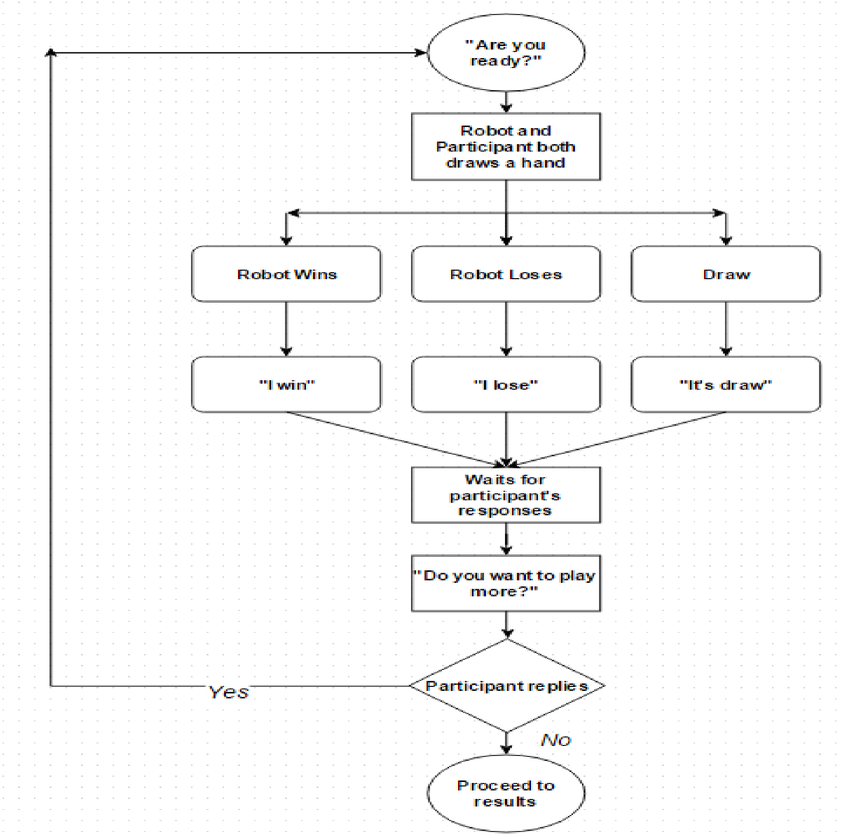


Figure 4.35. The structural diagram of game-playing with different outcomes.

The game ends with the robot saying who won the most rounds played. In the baseline interaction, the robot simply calculates who won the most and says it. As it is the Wizard of Oz experiment and as the examiner, I controlled the robot and the experiment, therefore it was easy to calculate the winner, and it was possible not to draw all hands. In this brief stage, the robot declares the winner and proceeds to the farewell.

At the end of the interaction, the robot thanks the participant for coming and playing games. It invites the participant for the next interaction as well.

4.2.4 Participants and grouping

Participants were recruited several times for the experiments. As the dialogues and actions for each algorithm are one of the most important parts of this experiment, I needed some participants to test and run the mock interactions, where it could be possible to test the initial dialogue and actions of the robot for appropriately biased and non-biased interactions and to finalise the experiments. For such mock experiments, mainly graduate students, colleagues, and friends from different working backgrounds were selected, and the final version of the dialogues and actions were set after running such mock interactions and consulting with my supervisor. For this mock interactions samples were called using contacts, mails and phones.

In all of the main experiments, participants were invited by advertisement and selected at a 50:50 ratio of men to women. As the goal of the experiments was to determine the preferences of the general population, the race, ethnicity, age ranges and backgrounds of participants were kept as heterogeneous as possible.

The interaction language was English, so it was necessary to make sure that all participants understood spoken English. However, one experiment did not require any verbal engagement (i.e. the MyKeepon experiment, Section 5.2), but the feedback form was designed in English.

Most of the experiments required multiple engagements between the robot and participants, followed by one to two weeks of time between. Therefore, participant availability for the specific dates and times were required. In some cases, a few of the participants had to be removed from the results due to their unavailability for all interactions.

To test the different biases and compare the results between biased and non-biased interactions, I had to divide the participants into several groups. This grouping was done after the participants filled in the consent forms. After ascertaining the participant’s gender, age and race, I assigned them to different groups, maintaining similar ratios of gender, age and race. In most cases, the data were not analysed based on gender, ages and race; rather the data that was analysed in comparison to the overall trends among participants, so that a clear view of all participants could be secured. The grouping of the participants was based on the methodology of that particular experiment and, the details of the grouping in each case are described in the following experiment sections, in the next chapter.

Awareness of the participants

At the beginning of the experiments, the information given to the participants was limited to a description of the process of the experiment. An introduction was given in which I discussed the tasks to be performed between the participants and robot (i.e. conversational and game playing interactions). I also discussed how many times the participants needed to interact with the robot and the data collection procedure, which was that the paper questionnaire would be supplied after the experiments. I discussed the game rules and how to play, but any information such as the developed biases or how the robot would behave in the experiments was not discussed, as those idea could affect the judgement of the participants.

4.2.5 Metrics and data collection

The current hypothesis refers to a comparison between two interactions – one interaction with the robot which shows the behaviours based on selected cognitive biases and participants, and the other one is the robot without such biased behaviours with the same participants. Therefore, two sets of behaviours were developed in the robot: biased behaviours based on selected cognitive bias and non-biased behaviours. The hypothesis also indicates a comparison of the participants' fondness towards these two interactions to determine which one the participant liked the most. As per the hypothesis, it was necessary to compare the biased and non-biased robot-participant interactions. Therefore, comparing between two sets of responses from participants after the interactions with robots is reasonable, so that the effects of the biases can be distinguished and the differences in participants' responses are assuredly the results of the changes in the robot's behaviours due to the presence and absence of the cognitive biases.

Consequently, to measure the responses of the participants, I considered four conditions, which are as follows:

- How much the participants liked the interaction with the robot;

- Whether the participants were comfortable during the biased and non-biased interactions;
- Whether the participants were pleased by interacting with the robot; and
- The developed rapport between each participant and the robot.

To measure these four conditions, I chose four metrics, namely likeability, comfort, pleasure and rapport:

Likeability. This category was to measure the likeability of the overall interactions for the participants, determining to what degree the participants liked the overall interaction and how much they liked it. In similar studies, researchers have used categories such as 'likeability' as a common variable in experiments to find out participant's fondness of interactions (e.g. Kim & Mutlu, 2014).

Comfort. This category was to measure the feeling of relaxation among the participants when they were interacting with the robot. In our experiments, participants interacted with different algorithms, by which the robot could brag, lie, call the participant a cheater, forget things and exhibit many biased behaviours. Such behaviours are common in humans, but could come as surprise in robots. Therefore, I wanted to find out how comfortable participants felt experiencing such biased and imperfect behaviours in robots during the interactions.

Pleasure. This category was to measure the enjoyment or happiness—in other words, the overall satisfaction—the participant felt in the interactions with the robot. The interactions were based on cognitive biases which caused the robot showed various imperfect behaviours, usually uncommon in human–robot interactions; therefore, it is important to know whether the participants were pleased or satisfied in experiencing such behaviours from the robots during their interactions. Therefore, I wanted to find out whether the participants were pleased or displeased by experiencing the selected biased and imperfect behaviours from the robot (e.g. idiocy or forgetfulness).

Rapport. This category was supposed to measure the level of understanding between the two parties (i.e. participant and robot) and their ability to communicate well with to another. It was chosen to find out how close participants felt to the robot, how easily participants thought they could understood the responses from the robot, and whether they saw the interaction as meaningful. It is important to understand the level of such mutual closeness between the robot and participants, as the robot in the experiments showed biased and imperfect behaviours common in people but uncommon in the social robots. Therefore, I chose this ‘rapport’ variable to measure such closeness between the robot and participants and to find out whether participants built such rapport with the robot in the interactions after seeing such human-like biased and imperfect behaviours from the robot.

Each of these categories was analysed to measure one or both conditions of the experiments:

1. the participant’s overall preferences to interact with the robot with the biased or non-biased algorithm, and
2. changes in the participant’s preferences over a period of time.

For each metric, a set of Likert-style questions were asked to the participants after the interactions so that participants could rate their experiences. The questions were crafted based on similar human-robot interaction experiments in which researchers used similar metrics for their experiments, so that the questionnaire for the current experiments was already validated for such metrics. In my earlier experiments, the questionnaires were created to measure how much participants enjoyed the interactions with biased and non-biased versions of the robots (i.e. to assess the likeability metric); as the experiments went further and I tested new biases with different methods, I needed more complex measures to get accurate analysis. The different measurements of the specific experiments are discussed in the corresponding experiments sections.

I mentioned that data collection was carried out using different point Likert scale questionnaires. There are different scales used in the human-robot interactions studies. Nomura and Kanda (2003) developed and used a scale called Negative Attitudes towards Robots Scale (NARS) which measures negative attitudes towards robot based on Likert style questions. An overview of existing scales used in human-robot interactions studies was carried out by Bartneck et al (2009) which describes the scales, their acceptability and reliability. In the current study, different scales were used based on the type of biases chosen to develop and the robot used in experiments.

Justification of using different point Likert scale

As I mentioned in the first experiment that a cognitive bias called misattribution was used in robot for the first time and it was needed to investigate the participant’s responds thoroughly. Therefore a 10-scale response option was given to the participants so that they can express their responses in detailed manner, and more measurement precision can be possible. A 10-scale response can provide better opportunity detect changes and more power to explain the point of view of the participants (Wittink DR & Bayer LR, 2003). In this experiment I used 10-scale Likert questions (Cronbach's Alpha 0.927) to get participants feedbacks.

For the experiment to investigate the effects of empathy gap bias developed in MyKeepon robot I used a 5-scale Likert questionnaire. Based on the literature a five-point scale appears to be less confusing and to increase response rate (Babakus and Mangold, 1992; Devlin et al., 1993; Hayes, 1992). Also, MyKeepon had limited movements abilities to express the differences between two set of behaviors, therefore I offered participants more specific choices to response using this 5-point Likert scale questionnaire (Cronbach's Alpha 0.930).

For investigating the effects of misattribution, empathy gap, ideocracy and lack of knowledge effects (influenced by Dunning-Kruger effect), self-serving and humors effects cognitive biases developed in the MARC robot I used seven-point scale Likert questionnaire. Studies suggested that seven-point scale might improve reliability and validity for the items (Churchill and Peter, 1984) and the response rate (Malhotra, 1993). Also, for these experiments, I used humanoid MARC robot which expressed different behaviors clearly by using speech and movements and so I didn't need to limit participant's response choices or to offer them excessive choices, and therefore a seven-point Likert scale questionnaire (Cronbach's Alpha 0.996 and 0.993) was a perfect these experiments.

The questions for these four factors were created based on various sources, for example, to understand the participant's fondness or likability of the experiment, eight questions were selected from Hone and Graham (2000); the 'comfort' questions were drawn from Hassenzahl (2003); and the 'rapport' questions taken from Mutlu et al. (2012). Some of the questions were formulated based on the preliminary findings and methods of my experiments. All the questionnaires appear in the appendix B, page 244.

Justification of using statistical methods

In the first and second experiments, the same participants were interacting with two versions of the robots (i.e., biased and non-biased); I therefore needed to compare the means of two different samples coming from same groups participants interacting with two different versions of the robots. The purpose of the analysis is to determine whether there is statistical evidence that the mean difference between paired observations on such interactions is significantly different from zero, and thus the paired sample t-test was used to analyse the data collected from the first and the second experiments.

For the third and fourth experiments, the means were calculated using SPSS and followed different procedures for the two groups: a one-way repeated measure ANOVA to analyse the first group’s data and a mixed ANOVA to analyse the second group’s data. In all measurements, the p value was <0.05 (i.e., a very small probability of this result occurring by chance). The Cronbach's alpha is 0.916, which indicates a high level of internal consistency for our scale.

I chose the one-way repeated measures ANOVA for the first group (i.e., all participants interacted with all versions of the robot for three times) because it is an extension of the paired-samples t-test and is used to determine whether there are any statistically significant differences between the means of three or more levels of a within-subject’s factors (i.e., independent variables). For the first group, participants interacted with all biased and non-bias conditions; that is, they were exposed to all conditions of the experiment, with each condition presented an equal number of times for the participants; a repeated measure ANOVA is therefore justified.

I chose the mixed ANOVA for the second group (i.e., a group of participants interacted with only one version of the robot for three times) because it compares the mean differences between groups over two or more times, and all subjects have undergone two or more conditions, but the subjects were also assigned to two or more separate groups (i.e., individual algorithm interactions).

4.2.6 Room setup

All interactions were setup in a room which was approximately 16 m × 18 m.

The diagram of the room (Figure 4.36) shows the setup for the interactions.

a. Tables with pre-experiment and post-experiment questionnaires.

b. The place where participants sit in front of the robot for the interaction.

c. Place for someone to control the robot.

The participant enters into the room through the door (indicated by an arrow). At first, the participant sits at Table 1 and are given the participant information form and consent form to sign. The participant then goes to section 'b' and sits in front of the robot. The wizard (i.e. myself) starts controlling the robot from section 'c' and continues throughout the interaction. After the interaction, the participant is requested to go back to section 'a' and answer the final set of questionnaires. After completing the questionnaires, the participant leaves the room through the door.

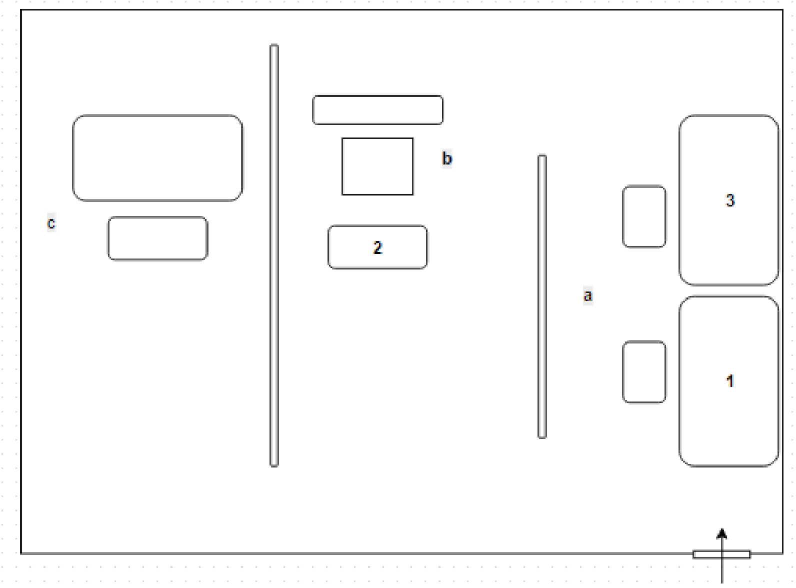


Figure 4.36. The experiment room setup

4.3 Summary

This chapter describes the important part of the study, where I planned and prepared for experiments using cognitive biases in social robots. The main outcomes and procedures are summarized as follows:

a. Stages:
   i. conversational, 3 stages,
   ii. game playing, 5 stages;
b. Dialogue design:
   i. baseline dialogue, robot ends with 'Okay', 'I understand', 'That's great' etc.,
   ii. biased dialogues, robot continues with selected biased effects;
c. Participants:

i. mock experiment—graduate students, friends and colleagues,

ii. main experiment—mixed in gender, race and age;

d. Measurement:

i. participant’s overall preferences towards the various algorithms, in terms of

1. likeability
2. pleasure
3. comfort, and
4. rapport with the robot; and

e. Room:

i. a room with 3 partitions,

ii. tables with pre-experiment and post-experiment questionnaires,

iii. the place where participant sits in front of the robot for the interaction, and

iv. the place for the wizard to control the robot.

I expect that the reader can follow and understand my experiments, how I prepared the robots and developed five biases on robots, and how participants were selected.

Chapter 5:

Analysis of the Experiments and Results

5.1 Conversational interactions with ERWIN based on the misattribution bias

5.1.1 Hypothesis and experiment goals

The aim of this experiment was to uncover how the misattribution bias developed in the robot ERWIN affects human–robot interactions. A local hypothesis for this experiment was created based on the main hypothesis:

> *If the misattribution bias is applied in social-robot interaction, then a participant's preference towards the robot will increase due to the human-like behaviours the robot exhibits as a result.*

To determine the effects of misattribution bias, participants interacted with the robot with misattribution bias and scored its 'likeability' (μ1) and they interacted with the robot without misattribution bias and scored its 'likeability' (μ2).

The null hypothesis is that the mean score of the participant's responses in (μ1) is the same as the participant's responses in (μ2):

i.e. Hypothesis 0: $\mu1 = \mu2$

where,

Hypothesis 0 = null hypothesis

(μ1) = the means of the participant's responses from biased robot interactions

(μ2) = the means of the participant's responses from non-biased robot interactions

The alternate hypothesis is

Hypothesis 2: $\mu1 > \mu2$

where,

Hypothesis 2 = participant's 'likeability' scores are greater in biased robot interactions.

5.1.2 Experiment methodology

As discussed earlier (chapter 3.5.1), misattribution is a memory bias which supposed to show that the robot forgot and misattributed information previously collected from users. Therefore, in this experiment, participants interacted with the robot ERWIN several times. As can be seen in Figure 5.1, in the first interaction, the robot collected information through the introductory conversation, and in the second interaction robot forgot and misattributed the collected information. In the non-misattributed interaction, the robot interacts with the participant without forgetting any previously collected information so that the biased and non-biased interactions can be compared.

This was the first experiment in which I developed the misattribution-bias algorithm in ERWIN, and it was tested with 30 participants. Each of the participants had three interactions with the robot, followed by at least a week of time between the two interactions. Participants were given a set of ERWIN's expression pictures at the beginning of the introductory interactions, which corresponded with the names of various emotions (see the appendix). Participants were asked to choose the correct emotion names from the list. At that point, participants had never seen ERWIN before, so such recognition could tell us about their skills in recognising ERWIN's emotions in the interactions.

The first interaction was common for all participants, however, after this first interaction, two differing interactions for each of the participants was given; one used the misattribution algorithm, and the other used the non-misattributed algorithm. In general, half of the participants went through the misattributed algorithm as their second interaction, and the other half went through the non-misattributed as their second interaction. The reason for this distribution was to make the comparison fair between biased and non-biased interactions by not determining any particular order for the interactions.

The three interactions were done by maintaining a time interval of at least two weeks to allow long-term affectivity in the participants. Figure 5.1 shows three interactions for a participant.

The first introductory experiment was to allow participants to familiarise themselves with the experimental environment and robot. The experiment was carried out in three steps:

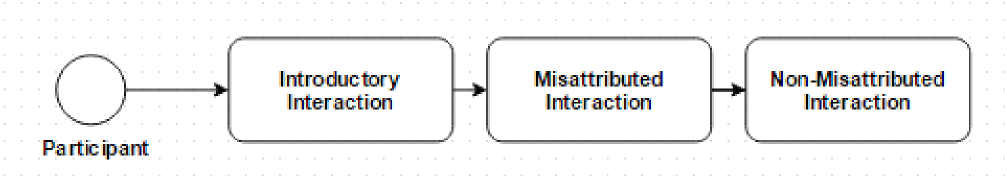


Figure 5.1. Each of the participants interacted with ERWIN three times

The first step was identification, and the participants were asked to identify the different facial expressions of ERWIN from pictures to see whether they could disambiguate the different expressions without meeting with the robot; the second step was the conversational session with ERWIN, where the robot started friendly conversation, greeting the participant, asking various questions and asking some general questions on various subjects such as hobbies, favourite colours, sports and others. The conversation's purpose was to allow the collection of basic information on the participants that would be used in the second and third experiments for ERWIN to misattribute. This initial conversation ended with a request from ERWIN to evaluate its performance. The participants were given a brief questionnaire on their experience with ERWIN.

In the second experiment, the participants were divided into two groups, with ERWIN remembering and making general conversation with the first group and misremembering the collected information for the other group's participants. In both cases, the participants were asked to answer questionnaires at the end of the experiment to find out which group of the participants was happier and had more satisfactory interactions with ERWIN.

In the third experiment, the participants from the previous experiment's 'non-misattributed group' experienced misattribution bias in conversations, and vice versa. Figure 4.5 shows the grouping process. At the end, all participants answered the same questionnaire as that given in the second experiment to determine what characteristics in ERWIN participants liked the most. All experiments were Wizard-of-Oz experiments, where the robot was controlled remotely, and participants were watched through ERWIN's eye cameras.

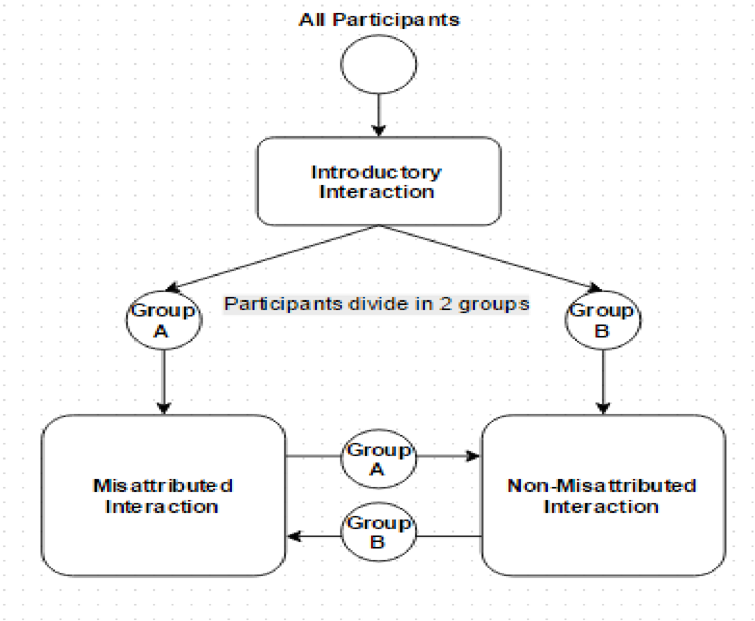

All Participants
Introductory Interaction
Group A
Participants divide in 2 groups
Group B
Misattributed Interaction
Group A
Group B
Non-Misattributed Interaction

Figure 5.2. The group division in the experiments

In all of the interactions, participants interacted one-on-one with ERWIN. The steps of the interactions were as follows:

1. Participant enters the room.
2. Participant sits in front of the robot (ERWIN).
3. The robot greets the participant and waits for response.
4. Participant replies.
5. Robot asks name, address and wellbeing, and waits for response.
    a. In the introductory interaction, robot asks name, address and wellbeing.
    b. In the unbiased interaction robot states correct name, address and wellbeing.
    c. In the biased (misattribution) interaction robot misattributes, forgetting information about name, address and wellbeing.
6. Participant replies.
7. Robot then moves to the topic-based conversation and discusses four topics (e.g. sports, music, food and book) and waits for response.

8. Participant replies.
    a. In the introductory interaction, robot replies with ‘okay’ and ‘thank you’ and moves to the next topic.
    b. In the unbiased interaction, robot states correct information in all topics.
    c. In the biased (misattribution) interaction, robot forgets and misattributes information about all topics.
9. Robot then moves to the farewell conversation, thanks participant for coming and waits for response.
10. Participant replies.
11. Robot invites participant for the next interaction and waits for response.
12. Participant replies.
13. Robot requests participants to complete questionnaires and waits for response.
14. End of the interaction.

5.1.3 Participants and grouping

Participants were recruited via advertising efforts. A total of 30 participants were selected. Participants were mixed in age, gender and background. As stated earlier, participants were divided into two groups to interact with the robot.

A demographic of the participants is as follow:

Table 5.1: Details of the participants

Gender	**Age**	**Occupation**
	18 -20 => 0	
Male 15	21 – 35 => 15	Student => 10
Female 12	36 – 60 => 15	Professionals => 15
No answer 3		No answer 5
30	30	30

5.1.4 Data collection and measurements

Data were collected in the form of questionnaires. After each of the interactions, participants were given a set of questionnaires to answer. The questionnaires structured offered possible answers on a Likert scale, so that participants could rate their experiences. This was the first experiment where human-like cognitive bias was used in a robot, and I wanted participants to rate their experiences from a range of options. Therefore, the rating options ranged from '1' and '10', where 1 represented 'least agreement' and 10 represented 'most agreement'. Questionnaires were the same for all of the interactions.

The experiment was to determine the participant's preferences for the algorithms. To find the participant's preferences, I used likeability as the measurement factor. There was a total number of 11 Likert-scale questions and four 'yes or no' based questions. To analyse data in the next section, responses from the Likert questions were used.

5.1.5 Experiment scenario

An interaction scenario can be described for the further clarification. 'Participant 1' was a young women and a student from the university. She found it amusing when ERWIN forgot previous discussions with her. ERWIN said, 'I remember that last time you said your favourite sport is wrestling. So please tell me what's going on in there?' 'Participant 1' never said that she like wrestling, but she laughed and replied, 'No, I don't like wrestling. I sometime watch football but that's rare.' ERWIN then apologized for the mistakes and started the next topic with another misattribution.

In case of the non-biased interactions, however, participant's responses were comparably limited, it could be because of their expectations of an interaction with a robot matched the reality, and they did not experience anything that could surprise them or made them think the robot's behaviours unusual. For example, when 'Participant 2', a fulltime employed at the property agency, took part in our experiment. In the non-misattribution interaction, ERWIN asked her, 'I remember that last time you said your favourite sport is cricket. So please tell me what's going on in there?' As the robot was correct by recalling her favourite sport, 'Participant 2' replied, 'I could not find any spare time to watch since last time.' Then ERWIN jumps to the next topic, asking, 'Last time you said your favourite food is curry.

Am I correct?' 'Participant 1' again generally answered, 'Yes, you are correct.'; and the conversation continued with ERWIN jumping to the next topic.

Figure 5.3. Different moments of the participant interacting with ERWIN

Figure 5.3 shows different moments of participants interacting with ERWIN. The pictures were captured using the robot's eye camera, so the robot could not be seen in the frame.

5.1.6 Experimental results and statistical analysis

A Shapiro-Wilk test was performed to check the normality (Yap & Sim, 2011) where the Sig. values came out as p>0.05, and therefore parametric test (t-test) was carried out to analyse data.

As the participants interacted with all algorithms once, they each had the chance to compare the two interaction behaviours. Since the participants and questionnaires were same for both biased and non-biased interactions, I ran a descriptive analysis and a paired sample *t* tests to analyse and compare the data. In both cases, the Sig (*p* value) was <0.05 (i.e. a very small probability of this result occurring by chance, under the null hypothesis of no difference).

The reliability scores from the 11 questionnaires were 0.94 and 0.756 (Cronbach's alpha), which are very high. In this case, the high Cronbach's alpha supports the ratings compared between the biased interaction and the non-biased interaction. I added the scores of all questions for each participant and ran a paired-sample *t* test to compare.

In the introductory interaction, most of the participants recognised all emotions of ERWIN correctly; however, a few participants mixed up one or two emotions (e.g. joy with surprise).

5.1.6.1 The justification of the crossover groups of participants

This section describes the ratings analysis between the cross-over groups. As in this experiment participants were divided into two groups (e.g. A and B) to interact with the biased and non-biased versions of the robot ERWIN, and at the end a crossover group's interactions were performed between those A and B groups (Figure 5.2), it is reasonable to explain how summing up such collected data from crossover interactions are justifiable.

I performed a *t* test between the biased condition group from the second interaction with the biased condition group from the third interaction to verify whether it is possible to combine the all biased groups for the final analysis, and the results are shown In Tables 5.2 and 5.3.

Table 5.2: Descriptive statistics from t test between first two cross-over groups.

Algorithms	***N***	**Min.**	**Max.**	**Mean**	**Std. Deviation**
Total Biased (Group 1)	15	61.00	107.00	85.26	13.76
Total Biased (Group 2)	15	77.00	100.00	90.6	6.68

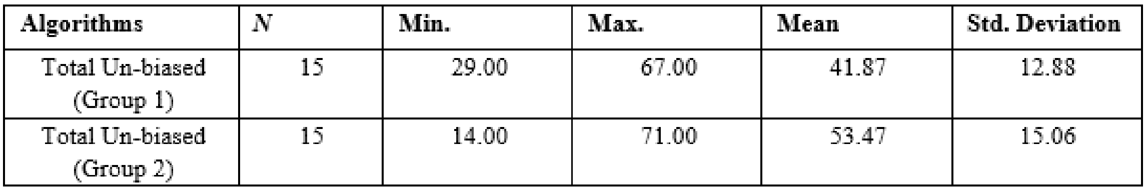
Table 5.3: Results from t test between first two cross-over groups.

Algorithms	**Paired Differences**					***t***	**df**	**Sig.**
	Mean	**Std. Deviation**	**Std. Error Mean**	**95% Confidence Interval for Difference**				
				Lower	**Upper**			
Biased (G1) vs Biased (G2)	-5.33	16.50	4.26	-14.47	3.8	-1.25	14	0.23

Table 5.2 and 5.3 show that there was no difference in the mean score between the biased G1 and G2. Therefore, it is possible that the crossover of the biased groups did not affect the ratings of the participants and combining the two biased groups is justified.

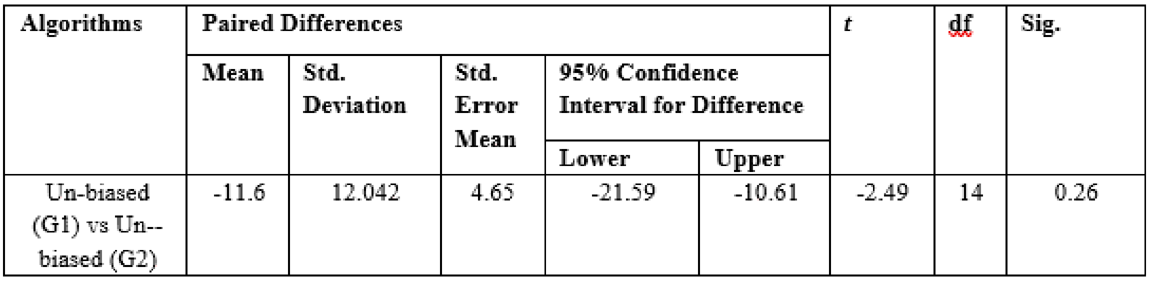
I also performed a *t* test for the non-biased condition group from the second interaction with the non-biased condition group from the third interaction to verify whether it was possible to combine the all unbiased groups for the final analysis, and the results are shown in Tables 5.4 and 5.5.

Table: 5.4: Descriptive statistics t test from between second two crossover groups.



Algorithms	***N***	**Min.**	**Max.**	**Mean**	**Std. Deviation**
Total Un-biased (Group 1)	15	29.00	67.00	41.87	12.88
Total Un-biased (Group 2)	15	14.00	71.00	53.47	15.06

Table 5.5: Results of t test from between second two crossover groups



Algorithms	**Paired Differences**					***t***	**df**	**Sig.**
	Mean	**Std. Deviation**	**Std. Error Mean**	**95% Confidence Interval for Difference**				
				Lower	**Upper**			
Un-biased (G1) vs Un--biased (G2)	-11.6	12.042	4.65	-21.59	-10.61	-2.49	14	0.26

From the table 5.4 and 5.5, there is no difference in in the mean score between the unbiased G1 and G2. Therefore, the crossover of the biased groups did not affect the ratings of the participants and combining the two biased groups is justified.

From the above tests, it can be said that combining the two biased and two non-biased groups for analysing the differences between biased and non-biased interactions is possible, as there are no significance differences in ratings.

5.1.6.2 Statistics of non-biased interactions

Table 5.6: The descriptive statistics and mean of the unbiased interactions.

Algorithm	***N***	**Min.**	**Max.**	**Mean**	**Std. Deviation**
Unbiased	30	14.00	71.00	47.67	14.98

In Table 5.6, *N* represents the number of participants, 30, the minimum average of rating is 14 and maximum average of rating is 71 with the mean value of 47.67. The standard deviation tells us that how much value differs from the mean value for the group, and for unbiased interactions it is 14.98.

5.1.6.3 Statistics of misattribution biased interactions

Table 5.7: The biased interactions and the descriptive chart.

Algorithm	***N***	**Min.**	**Max.**	**Mean**	**Std. Deviation**
Misattribution Biased	30	63.7333	107.00	87.93	25.562

In Table 5.7, *N* represents the number of participants, 30, the minimum average of rating is 63.7333 and maximum average of rating is 107 with the mean value of 87.93. From the standard deviation it can be said that data were spread both higher and lower from the mean by the value of 25.562.

As mentioned earlier, I added the ratings of all question for each participant, so it can be said that the minimum total rating for biased algorithms was 61, and the maximum was 107. The mean here was calculated over the summed ratings. The standard deviations that data were distributed more closely for non-biased interaction than biased interaction.

5.1.6.4 Pairwise statistics of the biased algorithms compared to baseline

The pairwise statistics in Table 5.8 shows a comparison of misattributed and non-misattributed algorithms.

Table 5.8: Comparison chart of biased and unbiased interactions

Algorithms	**Mean Differences**	**Std. Error**	**Sig.**	**95% Confidence Interval for Difference**	
				Lower Bound	**Upper Bound**
Misattributed vs Non-Misattributed	40.27	2.51	0.000	35.13	45.41

As Table 5.8 shows, the biased algorithms scored at least 40 points higher in each question for each of the participants. Figure 5.4 was plotted using the mean differences to show the preferences of the participants by bias.

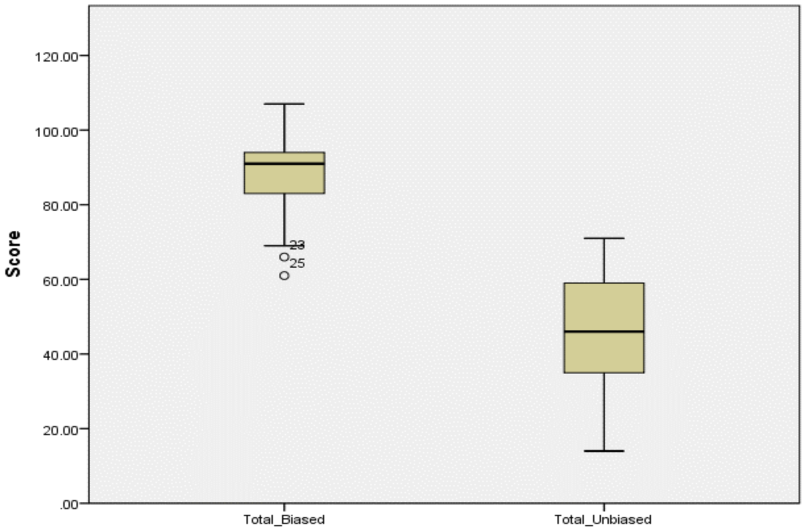


Figure 5.4. Rating distributions of the biased and non-biased interactions.

Figure 5.4 shows the distributions of the ratings from participants in biased and unbiased conversation. The graph shows that in the biased interactions, the participant's ratings were distributed mainly between 70–110, but in the non-biased interactions, such distributions were 15–70. However, the biased-interaction ratings are congested in the area between 85–95, whereas the non-biased ratings are between approximately 35–60.

The mean is 40.27, which show that there is an average 40-point difference in ratings, expressing a preference for the biased interactions. In the questionnaires, the rating options were from 1–0, so in this case, the average of 40 points actually suggests that in each question, the participants rated biased interactions an average of 4 points higher.

The means of biased and unbiased responses are shown in Tables 5.5 and 5.6. The mean for biased-interaction data is 87.93, approximately 40 points more than the mean for unbiased interactions (47.67), suggesting that for each question, the participant's responses were an average of 40 points (in this rating, 4) higher for biased than unbiased. It can also be seen that in Figures 5.1 and 5.2, the ratings of biased responses lay between 60 and 110, whereas the unbiased responses lay between 20 and 70. The graph in the Figure 5.3 shows the participant's preferences for the biased conversation over unbiased. Tables 5.1 and 5.2 show the descriptive analysis for the biased and non-biased interactions as well.

From the Table 5.3, the 2-tailed sig (*p* value) from the paired sample *t* test came to <0.05, which indicates the significance of the collected data over a large population. From the above *t* value, $t = 16.024$ and $p < 0.001$ (i.e. indicating a very small probability of this result occurring by chance). As p < 0.001, there are significance differences between our null hypothesis and alternative hypothesises. The null hypothesis (H0) can be rejected, since $p < 0.05$.

Therefore, the other alternative hypothesis H2 can be accepted, because µ1((87.93)) > µ2(47.67), where H2 = participant's 'likeability' scores, which were greater for the biased robot interactions.

In this experiment, from a total of 30 participants, 15 were male, 12 were female, and three did not specify their gender. As there is very little gender deviation, I did not show the results based on the gender. In this experiment, there was no participant younger than 21 and no participant older than 55. Although I divided age groups into two (21–35 and 36–55), the data are analysed based on all the participants, each considered an 'adult'. Based on occupation, participants reported being either students (n = 10) or working full time jobs (n = 15), although five did not indicate their occupations. However, I did not analyse results based on occupations, since only two occupations were available; rather I analysed data on overall participants and based on their responses on the questionnaire.

Therefore, it can be said that, overall, the participants liked the biased interactions more than the non-biased. From the above statistical analysis, it can be concluded that misattribution bias affected the interaction, and overall, the participants liked the interactions using misattribution bias with ERWIN more than the non-biased versions of the robot. The experimental results support our global hypothesis, which states that using misattribution cognitive bias in social human–robot interactions can increase the participant's likeability. In the next experiments, we will see the use of other cognitive biases in different robots and how participants reacted to such interactions.

5.2 Game-playing interactions with MyKeepon based on the empathy gap bias

In this experiment, I use the empathy gap bias in the robot MyKeepon to examine the effects of this bias in interactions between the robot and participants in game-playing interactions. The robot shows empathy-gap-biased behaviours created based on the hot-state and cold-state of the bias, and these states were randomly assigned to the participants. The robot plays 'tap-on-head' with the participants, where the participants tap on the robot's head, and the robot must jump as many times as it has been tapped. However, in this experiment, winning or losing the game was not a result of interest; rather, it was important to ascertain the participant's reactions to the robot's behaviours based on the biased algorithms. Therefore, the robot can win or lose all the game terms, but how much participants liked the robot's behaviours during the interactions were measured based on Likert questionnaires.

Similar to the previous experiment, a local hypothesis was made for this experiment based on the main hypothesis:

> *If the empathy gap bias is applied in social-robot game-playing interaction, then a participant's fondness or likability of the robot can be increased due to the personally relatable behaviours the robot exhibits as a result.*

To check the effects of empathy gap bias, participants interacted with the robot with the empathy gap bias and scored their ‘preference’ (μ1), and participants interacted with the robot without the empathy gap bias and scored their ‘preferenc’ score (μ2).

The null hypothesis is that the mean score of the participant’s responses in (μ1) is same as the participant’s responses in (μ2):

i.e. Hypothesis0: μ1 = μ2

where,

Hypothesis 0 = Null hypothesis

(μ1) = the means of the participant’s responses from biased robot interactions

(μ2) = the means of the participant’s responses from non-biased robot interactions

The alternate hypothesises are:

Hypothesis 2: μ1 > μ2

where,

Hypothesis 2 = participant’s ‘likeness’ scores are greater in biased robot interactions.

5.2.1 Experiment methodology

I tested the empathy gap effects in both game-playing and conversational interactions. For this experiment with the MyKeepon robot, I used tap on head and clapping games to play with participants where the robot responded to the number of taps and claps. To find out the effects of the empathy gap bias, a comparison was needed only with non-biased interactions, and therefore only two interactions were made for each participant with the robot in this experiment. To express the hot state and cold state of empathy, the robot showed different movements and made different noises. In this experiment, the robot was set to randomly show a state of empathy gap for participants, and throughout the interaction, the robot maintained this such state. Therefore, winning or losing the game was not a factor of the interaction, but rather how the robot showed the state of the empathy gap was the important point of the interaction. The empathic state of the participants was not measured because the expressions of the state of robot and the participants’ reactions to that were the targets of this experiment. The experimental procedure is shown in the Figure 4.20. Participants were assigned to the non-biased interaction, after which they experienced the other algorithm in the second interaction, with at least a week in between. The participants were randomly assigned to the hot or cold states of the biased algorithms. The robot was controlled manually (Wizard of Oz) by me and therefore, so real time responses were produced.

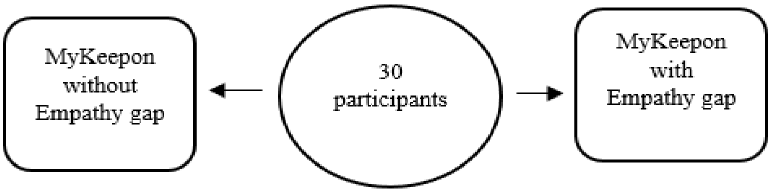


Figure 5.5. Each participant interacted both versions of MyKeepon.

As stated earlier, ViKeepon supports a different range of sounds, including wake up, yawn, sleep, chimp, sneeze and others. With custom moves and sounds, different interaction moments were created, such as different greetings, gestures, dance movements, and emotions such as happiness, sadness, joy and others. Each participant was allowed up to 10 minutes to interact with the MyKeepon. The interaction process included greetings, trying to establish eye contact with participants, responding to participant inputs by showing making movements, biased or non-biased, and different expressions. Participants could tap MyKeepon's head, and they could clap to make MyKeepon dance. The game was, in general, that the robot would have to jump the same number of times the participant tapped its head. Based on the principal of the hot-state empathy gap bias, the robot stays excited and shows excess happiness even if it loses, and in the cold state robot stays quiet and unresponsive even if it wins in games. In the non-biased interactions, the robot shows no overly excited or unresponsive behaviours; rather, it maintains a balance between over excitation and unresponsiveness by limiting its jumps, dances and noises.

All participants were given a verbal and a written set of instructions about interacting with the robot at the beginning of the experiment. In all the interactions, participants interacted with the robot on a one-on-one basis.

5.2.2 Participants and grouping

A total of 28 participants were selected from the responses to the advertisement. Genders, background, races and ages were mixed in the participants. In this experiment, the advertisements were placed near bus stations, so most of the respondents were of the general public rather than being students.

Table 5.9: Details of the participants

Gender	Age	Occupation
	18 -20 => 0	
Male 13	21 – 35 => 16	Student => 13
Female 12	36 – 55 => 12	Professionals => 13
No answer 3		No answer 2
28	28	28

Each of the participants interacted two times with the robot (Figure 5.5), where in one interaction, participants interacted with MyKeepon without empathy gap behaviours, and in the next interaction interacted with a biased version of the robot.

5.2.3 Data collection and measurements

Participants were requested to complete a set of questions at the end of each interaction. The questionnaires used a Likert scale and were the same for both experiments, so that the differences of the participant's preference could be compared in the two interactions. There was a total of 14 questions in the questionnaire. In addition, participants could leave their own comments about the interaction.

Similar to the previous ERWIN experiment, in this case I wanted to find out the fondness of the participants to interaction with the biased or non-biased algorithms, so data were collected based to determine overall likeability. In this experiment, I used a 5-point Likert scale, because most of the invited participants were from the general public. Studies (e.g. Weijters, Cabooter & Schillewaert, 2010) show that a 5-point scale is useful when respondents are from the general public. The questions were the same for all the interactions, so that a comparison could be possible. For example, participants were asked whether they found the different behaviours of the robot meaningful, and they were asked to rate these behaviours.

5.2.4 Statistics of game-playing interactions based on empathy-gap bias in MyKeepon, the toy-like robot

The test for Normality, examining standardized skewness and the Shapiro-Wilks test (Yap & Sim, 2011), indicated the data were statistically normal (p>0.05), and therefore parametric test (t-test) was carried out to analyse data. In this experiment, 28 participants interacted with MyKeepon with empathy-gap-biased and non-biased algorithms.

Participants were asked 14 questions. As in the previous ERWIN experiment, the measure of this experiment was preference for the algorithms. The questionnaires and participants were same for the both interactions; and paired sample *t* tests were used to analyse the data. In this statistical analysis, the Sig (*p* value) came out as <0.05 (i.e. a very small probability of this result occurring by chance) under the null hypothesis of no difference, so the results are statistically significant.

The reliability scores from the 14 questionnaires for the biased and unbiased interactions are 0.987 and 0.83. Because in this experiment data were collected in ways similar to the ERWIN experiment, I used a similar method of adding the ratings from questionnaire answers for each of the participants, and used a paired sample *t* test to compare.

5.2.4.1 Statistics of unbiased interactions

Table 5.10 shows the descriptive statistics and the means of the unbiased interactions.

Table 5.10: Statistics of unbiased MyKeepon interactions.

Algorithm	***N***	**Min.**	**Max.**	**Mean**	**Std. Deviation**
Unbiased	28	43	64	52.39	5.91

In Table 5.10, *N* represents the number of participants, 28, minimum average of ratings is 43 and maximum is 64. The mean value is 52.39 and from the standard deviation value it can be said that the ratings were spread from mean value by 5.91.

5.2.4.2 Statistics of empathy gap biased interactions

Table 5.11 presents the descriptive statistics of the empathy-gap-biased interactions.

Table 5.11: Statistics of biased MyKeepon interactions.

Algorithm	***N***	**Min.**	**Max.**	**Mean**	**Std. Deviation**
Empathy Gap Biased	28	47	64	55.35	4.99

In Table 5.11, *N* represents the number of participants, 28, minimum average of ratings is 47 and maximum is 64. The mean value is 55.35 and from the standard deviation value it can be said that the ratings were spread from mean value by 4.91.

5.2.4.3 Pairwise statistics of the biased algorithms compared to baseline

The pairwise statistics in Table 5.12 show a comparison of misattributed and non-misattributed algorithms.

Table 5.12: Pairwise comparison between empathy–gap-biased and non-biased interactions.

Algorithms	**Paired Differences**					***t***	**df**	**Sig.**
	Mean	**Std. Deviation**	**Std. Error Mean**	**95% Confidence Interval for Difference**				
				Lower	**Upper**			
Empathy Gap vs Non-Empathy Gap	2.97	1.95	0.37	2.20	3.72	8.03	27	0.000

As Table 5.12 shows, the biased algorithms scored at least 2.97 points higher in each question. The difference in means between the empathy gap and non-empathy gap ratings is 2.97, which indicates that there is an approximate 3-point difference in ratings in preference to empathy gap interactions. There are other statistical measures in Table 5.10 show standard deviation (1.95), error means (0.37) and the boundaries of ratings distributions (2.2–3.72).

A whisker plot (Figure 5.6) has been made using the mean differences between non-biased and biased ratings. The figure shows the distribution of the ratings in average differences between the responses from participants in biased and unbiased conversation. The graph shows the number of participants who preferred the biased interaction over the unbiased interaction.

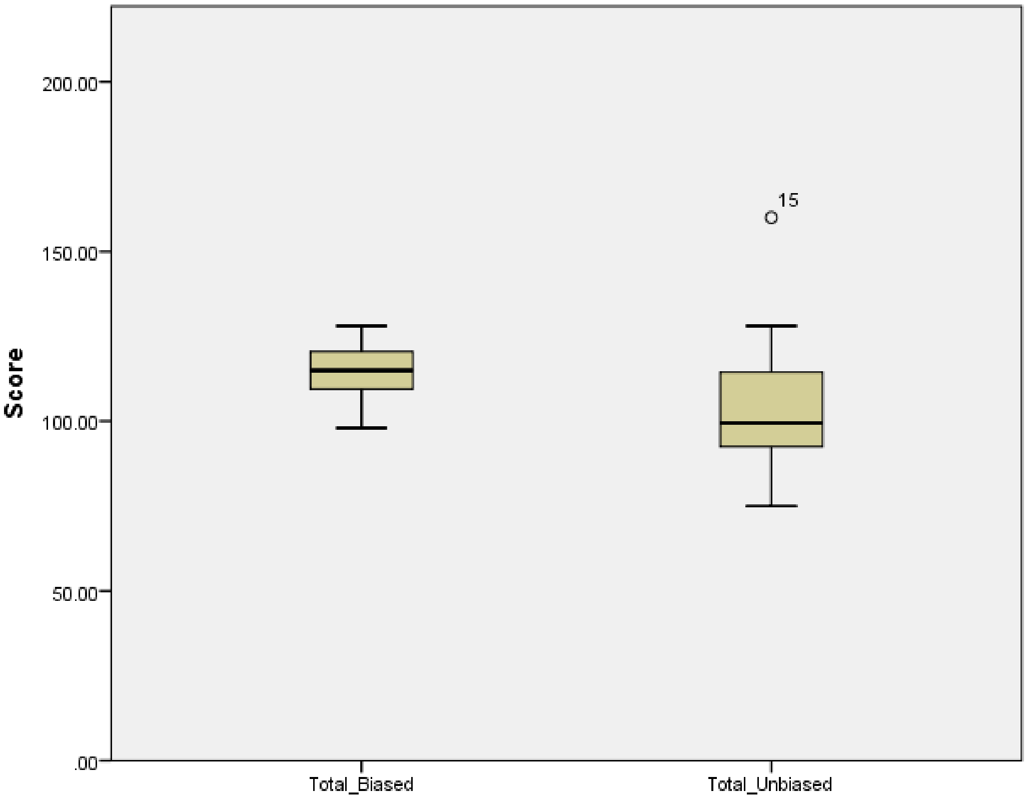


Figure 5.6: Rating distributions of the biased and non-biased interactions.

5.2.5 Statistical result discussion

Tables 5.8 and 5.9 show the mean ratings non-biased and biased algorithms, which are 52.39 and 55.35. After comparing the two means, the biased mean is approximately 3.0 (2.96) points more than the unbiased mean, which indicates that for each question, the participant's responses were on average rated 2 points higher for biased than for unbiased interactions. This result can be seen in Table 5.10, where the mean differences are calculated as 2.97. In this experiment, the questionnaire was designed based on a 5-point scale, so the means and differences are smaller in numbers.

The correlation between the two sets of scores was calculated as 0.95. The pattern of change is consistent for each participant for each question. From the *t* value, $t = 8.032$ and $p < 0.001$, which is a very small probability of this result occurring by chance, under null hypothesis of no difference. So, the null hypothesis is rejected, since $p < 0.05$. According to the above measurements, there is evidence ($t = 8.032$, $p < 0.0001$) of preference for the biased robot interactions over non-biased interactions. In these interactions, participants preferred the biased version of MyKeepon, on average, by approximately 2.97 points. With a 95% confident interval, the lower and upper limits are 2.20706 and 3.72151, respectively, which means a larger population would prefer the biased interaction by a mean of 2.97 and with a range between 2 and 3 points for each question. Therefore, from the above statistical analysis, it can be concluded that the development of cognitive biases does affect interactions between the robot and the participants, and overall the participant's liked the interactions using empathy gap biases in the MyKeepon experiments.

From Table 5.10, the 2-tailed sig (p value) from the paired-sample t test was <0.05, which indicates the significance of the collected data over a large population. From the above value, $t = 8.03$ and $p < 0.001$ (i.e. a very small probability of this result occurring by chance). As $p < 0.001$, there are significance differences between our null hypothesis and alternative hypothesises. So, the null hypothesis (H0) can be rejected, since $p < 0.05$.

As such, the alternative hypothesis H2 can be accepted, because $\mu 1$ ((55.35)) > $\mu 2$ (52.39), where H2 = participant's 'preference' scores being greater in biased robot interactions.

In this experiment from total of 28 participants 13 were male and 13 were female, and two did not specify their gender. As there are very few numbers of gender deviation, I did not show the results based on the gender. In this experiment, there was no younger participant than 21 and no older participant than 55. Although I separated age groups into two (21-35 and 36-55) but data are analysed based on all the participants are in general 'adult' group. Based on the occupation, participants were only either students or doing full time jobs and two did not provide any answers. However I did not analyse results based on gender and occupations due to only 13 on each category were available, rather I analysed data on overall participants and based on their responses towards questionnaire.

Therefore, it can be said that the participants overall liked the biased interactions more than non-biased. From the above statistical analysis, it can be concluded that the hot and cold empathy gap biases affected the interaction, and overall, the participants significantly liked and preferred the interactions using empathy gap bias with MyKeepon.

5.3 Conversational interactions with MARC based on the misattribution, empathy gap and lack of knowledge or ideocracy effects

This was the third experiment of the series investigating cognitive biases in the robot for human–robot interaction. In this experiment, I used previously investigated misattribution and empathy gap biases and added some new behaviours, namely a lack of knowledge, or ideocracy (similar to the Dunning-Kruger effects), in a humanoid robot MARC. The reason for using previously investigated biases was to find out whether the positive responses received from the participants in the ERWIN and MyKeepon experiments were for the algorithm or the robot itself and to investigate the effects of such biases in a humanoid robot-participant interaction. In this experiment, I developed similar misattribution and similar empathy gap conversational algorithms in MARC, and I developed a conversational algorithm for the lack of knowledge or ideocracy effects bias as well. Similar to the previous experiments, in the end I compared the robot's biased behaviours against its baseline behaviours in conversational interactions.

To investigate the effects of the three biases and the imperfectness in the humanoid robot MARC, three local hypotheses were created, based on the main hypothesis of the research.

For the interactions using misattribution cognitive bias, the first was as follows:

If the misattribution bias is applied in social-robot interaction, then a participant's fondness or likability of the robot can be increased due to the human-like behaviours the robot exhibits as a result.

For the interaction using empathy gap bias, the second was,

If the empathy gap cognitive bias is applied in social-robot interaction, then a participant's fondness or likability of the robot can be increased due to the human understanding the behaviours the robot exhibits as a result.

For the interactions using the lack of knowledge or the ideocracy bias,

If the lack of knowledge, or the ideocracy, bias is applied in social-robot interaction, then a participant's fondness or likability of the robot can be increased due to the human-like behaviours the robot exhibits as a result.

For all the cases,

To ascertain the effects of the biases, participants interacted with the robot with a given bias and scored the robot's 'likeability' (μ1), and the participants interacted with the robot without misattribution bias and scored its 'likeability' (μ2).

The null hypothesis is that the mean score of the participant's responses in (μ1) is same as the participant's responses in (μ2):

i.e. Hypothesis0: μ1 = μ2

where,

Hypothesis 0 = Null hypothesis

(μ1) = the means of the participant's responses from biased robot interactions

(μ2) = the means of the participant's responses from non-biased robot interactions

The alternate hypothesis is

Hypothesis 2: μ1 > μ2

where,

Hypothesis 2 = participant's 'preference' scores are greater in biased robot interactions.

Therefore, for the individual biased versus non-biased experiments, if the null hypothesis gets rejected, it can be proven that the participants liked and preferred to interact with the biased version of the robot.

5.3.1 Experiment methodology

In this experiment, three cognitive biases—misattribution, empathy gap and lack of knowledge or ideocracy effects—were compared with the non-biased behavioural algorithm. The interactions were one-on-one between the participant and the robot MARC, and the robot was controlled manually (Wizard of Oz) throughout the interactions.

Similar to the previous misattribution experiment with ERWIN, in this experiment, the bias was expressed after collecting the basic information of the participants. The main differences with the previous ERWIN experiment were the robots (e.g. ERWIN and MARC) and the way robot expressed the bias effects.

Similar to the previous experiment with the empathy gap interaction, in this experiment, robot MARC expressed hot and cold states of the empathy gap in conversational interactions with the participants. When in the hot state, the robot expressed faster and more enthusiastic responses with a friendlier approach, but in the cold state, responses were minimal. Similar to the previous MyKeepon experiment, the participant's emotions and mood were not measured, as only the robot's behaviours were concerned in the experiment. As mentioned previously, we know that the person influenced by the empathy gap bias is not aware of his or her own effects and therefore does not care about others' emotions at all. For this reason, in the experiment with empathy gap effects in robots, I developed the bias and showed such effects in interactions without acknowledging the participant's true emotions at that time, so that the empathy gap bias could be truly justify. The differences between the MyKeepon and MARC empathy gap experiments are mainly about the way the robots expressed the effects of the bias and the interaction procedures, which were game playing and conversations.

The lack of knowledge effect was tested using conversational interactions. As this bias was mainly developed based on Dunning-Kruger effects, in the interactions, the robot expressed its lack of knowledge by connecting incorrect thoughts to conversations on different topics.

Similar to all three bias based interactions with the robot, participants interacted with the baseline or non-biased algorithm three times maintaining two-three weeks interval between two interactions. At the end, all data from the biased interactions were compared with the data from the baseline interactions (Figure 5.7).

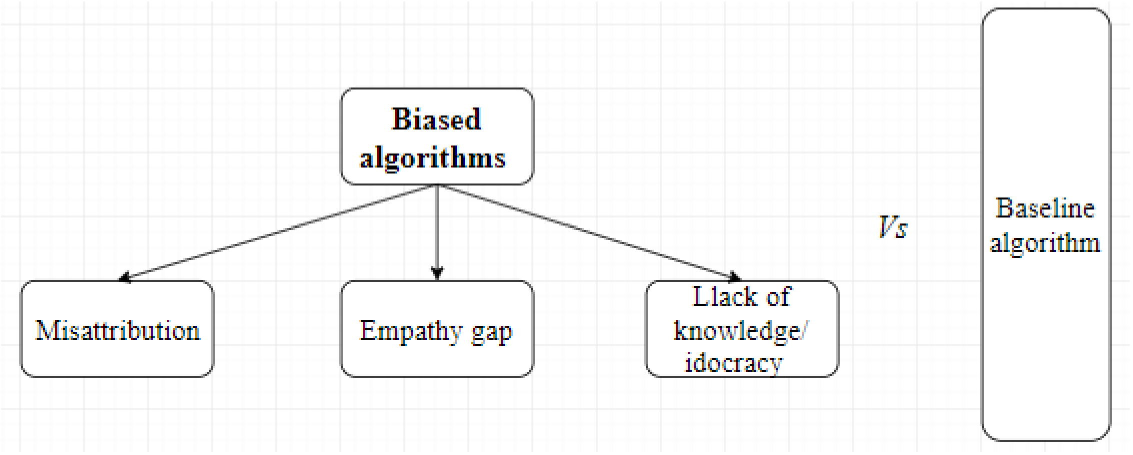


Figure 5.7. The experiment structures.

Two different setups were chosen for this experiment: One group of participants interacted with all four algorithms (Figure 5.8), and the other group interacted with only one algorithm, three times (Figure 5.9). This setup was to aid determining the participant's reactions in two different interaction scenarios over time.

Three interactions were completed between the participant and the robot, spanning a two-month time period, in order to observe long time interval effects and to allow the participant to reflect on their previous interactions. The interactions occurred on a one-on-one basis, where each participant interacted with MARC individually by making conversation for at least 8 minutes. The conversation ended with the robot requesting that the participant complete the questionnaire.

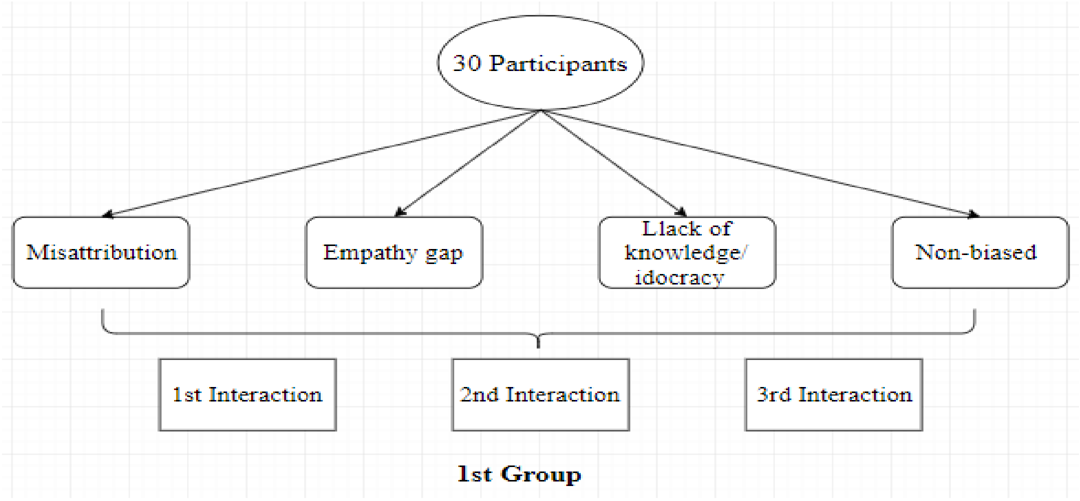


Figure 5.8: The first group: 30 participants interacting with all algorithms.

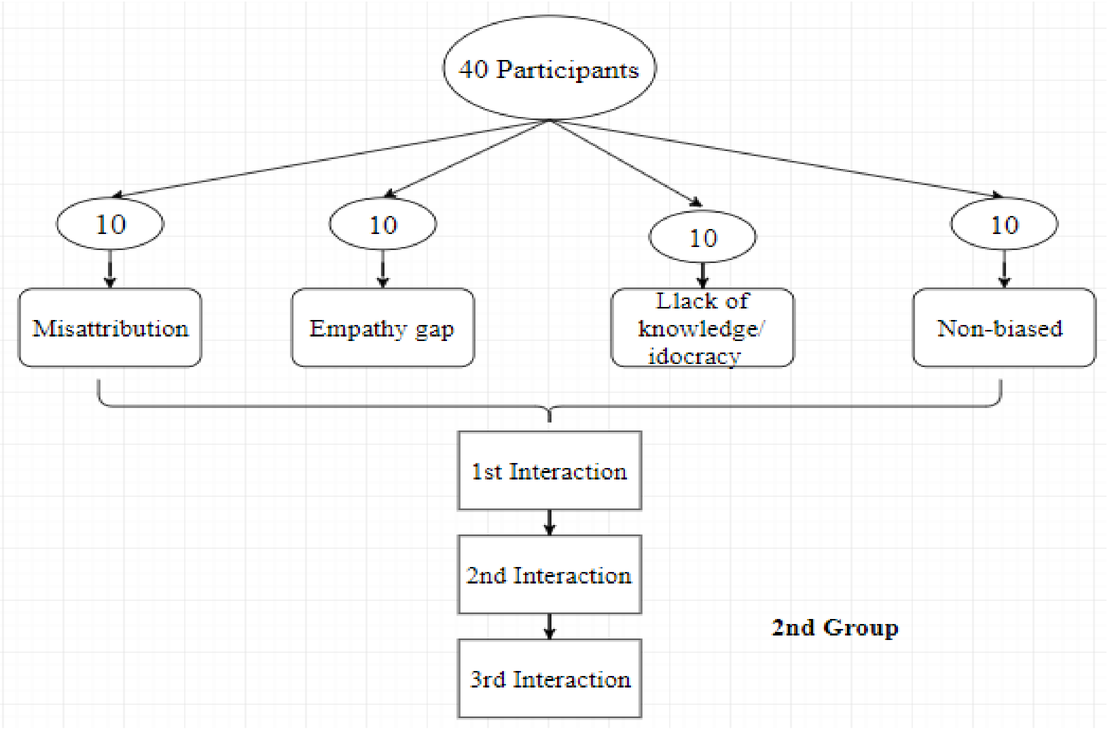


Figure 5.9: The second group: 40 participants interacting with one algorithm.

The interaction steps were are as follows:

1. Participant enters in the room.
2. Participant sits in front of the robot (MARC).
3. Robot greets the participant and waits for response.
4. Participant replies.
5. Robot asks name, address and wellbeing and waits for response.
   a. In the introductory interaction, robot asks names, address and wellbeing.
   b. In the unbiased interaction, robot states correct names, address and wellbeing.
   c. In the misattribution biased interaction, robot misattributes, forgets information about name, address and wellbeing.
   d. In the hot state of empathy, robot asks name, address and wellbeing very enthusiastically.

  e. In the cold state of empathy, robot asks name, address and wellbeing in very minimalistic way.

  f. In the lack of knowledge or ideocracy effects biased interaction, robot states name, address and wellbeing.

6. Participant replies.

7. Robot then moves to topic-based conversation and discusses four topics (e.g. sports, music, food and literature) and waits for a response.

8. Participant replies.

   a. In the introductory interaction, robot replies with ‘okay’, ‘thank you’ and moves to the next topic.

   b. In the unbiased interaction, robot states correct information on all topics.

   c. In the biased (misattribution) interaction, robot forgets and misattributes information on all the topics.

   d. In the hot state of empathy, robot discuss each of the topics very enthusiastically by showing its excessive happiness.

   e. In the cold state of empathy, robot discusses each of the topics very calmly, almost unresponsively, showing its excessive sadness and unwillingness to interact.

   f. In the lack of knowledge or ideocracy effects condition, robot express its own statements for each of the topics, and if participant disagree, then robot understands its mistakes.

9. Robot then moves to the farewell conversation and thanks participant for coming and waits for response.

10. Participant replies.

11. Robot invites participant for the next interaction and waits for response.

12. Participant replies.

13. Robot requests participants complete questionnaires and waits for response.

14. End of the interaction.

5.3.2 Participants and grouping

A total of 70 participants were randomly chosen, from responses to advertisements. The number of different genders, races and age groups were maintained as equal for both groups.

For the first group (shown in Figure 5.8), 30 participants were selected, and each participant interacted with all four behavioural algorithms (three biased and one unbiased). In the second group (shown in Figure 5.9), 10 participants from total 40 were selected to interact with each of the individual algorithms (individual biased or unbiased) three times. As with the first group, all 30 participants interacted with all biased and unbiased algorithms, so their responses would be based on the comparisons between the biased and unbiased interactions. Such responses would reflect a comparable outcome between those interactions involving the cognitive bias of the robot and those without, in developing human-robot interactions. In the second group, each of the 10 participants interacted with their selected individual algorithm three times. Such interactions could tell us the effects of each individual algorithm in developing human-robot interactions with maintain time intervals.

A demographic of the participants are as follow:

Table 5.13: Details of the participants

Gender	Age	Occupation
	18 -20 => 8	
Male 30	21 – 35 => 33	Student => 33
Female 34	36 – 60 => 19	Professionals => 28
No answer 6	No answer 10	No answer 9
70	70	70

The entire experiment took a long time to complete, because of variations in the participants' schedules.

5.3.3 Data collection and measurement

As stated earlier, in this experiment, data were collected based on three factors: likeability, comfort and rapport. Participants were given a questionnaire after each of the interactions, to collect the data for later analysis. The questionnaires were made using a 7-point Likert scale, where ‘1’ represented ‘least agreeableness’ and ‘7’ represented ‘most agreeableness’. The questionnaires were the same for both groups, and in all three experiments.

These three categories were used to measure two conditions:

- Participant’s preference for the algorithms (i.e. how much they liked the interaction).
- Changes in preference over an extended period (i.e. to find out whether such preferences for the interactions changed over time).

The three factors were chosen based on the findings of the experiment. The biased interactions were based on specific component of the selected bias. From such interactions between the participant and robot, I wanted to answer the following questions:

- Are the participants feeling *comfortable* to interact with the robot with and without the effects of the selected biases and components?
- Does the participant *like* interacting with the robot with and without such biases?
- How do such interactions between participants and the robot with or without such biases affect the *rapport* between the participant and robot?

Likeability, comfort and rapport are three of the most important measures of participant response in the current experiment, because to evaluate an interaction it is important to know the interlocutors’ mutual fondness, how comfortable they were and how strong the rapport between them becomes. Based on these three factors, I calculate the participants’ overall preferences and how they change over time period. The questionnaires were created from various sources, such as Hone and Graham (2000), Hassenzahl (2003) and Mutlu (2012), and are attached in the appendix section.

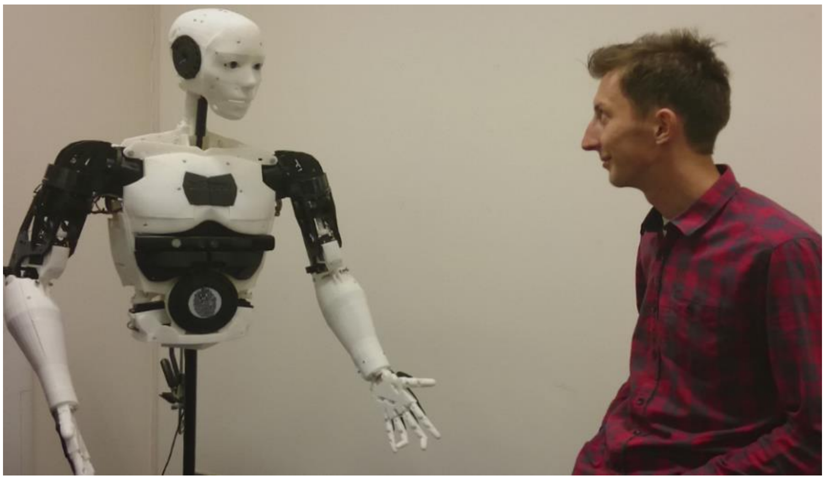


Figure 5.10. A participant is interacting with MARC.

5.3.4 Statistics for conversational interactions based on misattribution, empathy gap and lack of knowledge or ideocracy effect biases in MARC the humanoid robot

5.3.4.1 Introduction

As discussed earlier, data were collected on seven-point Likert questions to measure following three scales:

1. comfort,
2. likeability, and
3. rapport.

In this section, for each algorithm, I present the means of such factors from the statistical analysis and later compare these with the baseline.

In all the graphs, the *x*-axis represents the factors and algorithms, and the *y*-axis represents the marginal means.

5.3.4.2 Statistics of baseline interactions

First group

Table 5.14 shows the calculated statistical means based on each of the factors in three interactions:

Table 5.14: Means of the factors for first group in baseline interactions.

Interactions	**Measurements**	**Means**
1st Interaction	*Comfort*	4.87
	Likeability	4.92
	Rapport	4.72
2nd Interaction	*Comfort*	3.96
	Likeability	4.18
	Rapport	3.97
3rd Interaction	*Comfort*	3.96
	Likeability	3.41
	Rapport	3.29

Table 5.14 demonstrates that the highest mean for comfort is 4.87, in the 1st interaction; for likeability, it is 4.92, and for rapport it is 4.72, each for the first interaction. In the second and third interactions, such means values dropped for all three factors.

Second group

Due to the different style of analysis of the second group's data, the overall means of the three factors from three interactions are calculated (Table 5.15).

Table 5.15: Means of the factors for the second group in baseline interactions.

Measurements	**Means**	**Std. Error**	**95% Confidence Interval**	
			Lower Bound	**Upper Bound**
Comfort	3.41	0.08	3.23	3.6
Likeability	4.10	0.07	3.94	4.26
Rapport	3.75	0.05	3.63	3.9

In the second group, the average means for comfort in three interactions is 3.41, for likeability 4.10, and for rapport 3.75.

5.3.4.3 Statistics of misattribution biased interactions

First group

Table 5.16 shows the statistical means based on each of the factors in three interactions.

Table 5.16: Means of the factors for first group in misattribution interactions.

Interactions	**Measurements**	**Means**
1st Interaction	*Comfort*	5.59
	Likeability	5.78
	Rapport	5.00
2nd Interaction	*Comfort*	5.37
	Likeability	5.59
	Rapport	4.94
3rd Interaction	*Comfort*	5.11
	Likeability	5.22
	Rapport	4.82

Participants gave higher ratings for misattribution interactions than for the baseline (Table 5.16).

Second group

Due to the different style of analysis of the second group's data, the overall means of the three factors from three interactions are calculated Table 5.17.

Table 5.17: Means of the factors for the second group in misattribution interactions

Measurements	**Means**	**Std. Error**	**95% Confidence Interval**	
			Lower Bound	**Upper Bound**
Comfort	5.42	0.08	5.25	5.6
Likeability	5.17	0.07	5.01	5.32
Rapport	5.34	0.05	5.22	5.46

From this table, it can be said that the second group's participants rated higher for the misattribution interactions than the baseline (Table 5.10).

5.3.4.4 Statistics of empathy gap biased algorithm

First group

Table 5.18 shows the calculated statistical means based on each of the factors over three interactions.

Table 5.18: Means of the factors for first group in empathy gap interactions.

Interactions	Measurements	Means
1st Interaction	*Comfort*	5.59
	Likeability	5.78
	Rapport	5.00
2nd Interaction	*Comfort*	5.37
	Likeability	5.59
	Rapport	4.94
3rd Interaction	*Comfort*	5.11
	Likeability	5.22
	Rapport	4.82

Table 5.18 demonstrates that the first group's participants rated empathy gap interactions higher than the baseline (Table 5.14) for each of the factors.

Second group

Due to the different style of analysis of the second group's data, the overall means of the three factors from three interactions are calculated (Table 5.19).

able 5.19: Means of the factors for the second group in empathy gap interactions

Measurements	Means	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Comfort	5.23	0.08	5.06	5.41
Likeability	5.12	0.07	4.97	5.27
Rapport	4.57	0.05	4.45	4.69

Participants from the second group, overall, liked the empathy gap biased interactions more than the baseline (Table 5.15).

5.3.4.5 Statistics for lack of knowledge and ideocracy effects biased algorithm

First group

Table 5.20 shows the calculated statistical means based of each of the factors in three interactions. In most of the figures, this 'ideocracy' is denoted as 'DK'.

Table 5.20: Means of the factors for first group in lack of knowledge or ideocracy effects interactions

Interactions	Measurements	Means
1st Interaction	*Comfort*	5.59
	Likeability	5.78
	Rapport	5.00
2nd Interaction	*Comfort*	5.37
	Likeability	5.59
	Rapport	4.94
3rd Interaction	*Comfort*	5.11
	Likeability	5.22
	Rapport	4.82

The means calculated in Table (5.20) suggests that the first group's participants overall liked the lack of knowledge or ideocracy effects biases more than the baseline (Table 5.14).

Second group

The means from the second group's data are shown in Table 5.21:

Table 5.21: Means of the factors for the second group in lack of knowledge or ideocracy effects interactions

Measurements	**Means**	**Std. Error**	**95% Confidence Interval**	
			Lower Bound	**Upper Bound**
Comfort	5.58	0.08	5.4	5.76
Likeability	5.14	0.07	4.99	5.29
Rapport	5.00	0.05	4.88	5.12

Similar to previous biased interactions, the second group's participants liked the lack of knowledge or ideocracy effects biased interactions more than the baseline (Table 5.15).

The means for the individual algorithm suggested that all biased interactions scored higher than the baseline interactions in the both groups. In the next sections, I offer a comparison based on overall ratings of the factors for the group with three interactions.

5.3.4.6 Pairwise comparison of means based on overall ratings of three factors

This section presents comparisons of the algorithms based on the three factors (i.e. comfort, likeability and rapport).

First group

Table 5.22 shows pairwise comparisons of the means between baseline and biased interactions.

Table 5.22: Pairwise comparisons between biased and non-biased algorithms in the first group.

Algorithm Pairs		**Mean Differences**	**Std. Error**	**Sig.**	**95% Confidence Interval for Difference**	
					Lower Bound	**Upper Bound**
Baseline	Misattribution	-10.13	0.62	0.000	-11.88	-8.38
Baseline	Empathy gap	-5.06	0.6	0.000	-6.77	-3.36
Baseline	Lack of knowledge/ideocracy	-9.7	0.52	0.000	-11.17	-8.22

The mean differences between the baseline and the biased algorithms was negative, indicating (as shown earlier) that the means of the biased algorithms were higher in all interactions compared to the baseline. The Sig. value was 0.000 (<0.05), which indicates the statistical significance of the data.

A whisker plot (Figure 5.11) has been made using the mean differences between non-biased and biased ratings. The whisker plot shows (5.11) the distribution of the ratings based on the overall means for each of the algorithms. In other words, it can be called the influence graph, as the differences in the overall means of the algorithms can be easily spotted, and therefore, the influences of the algorithms can be seen.

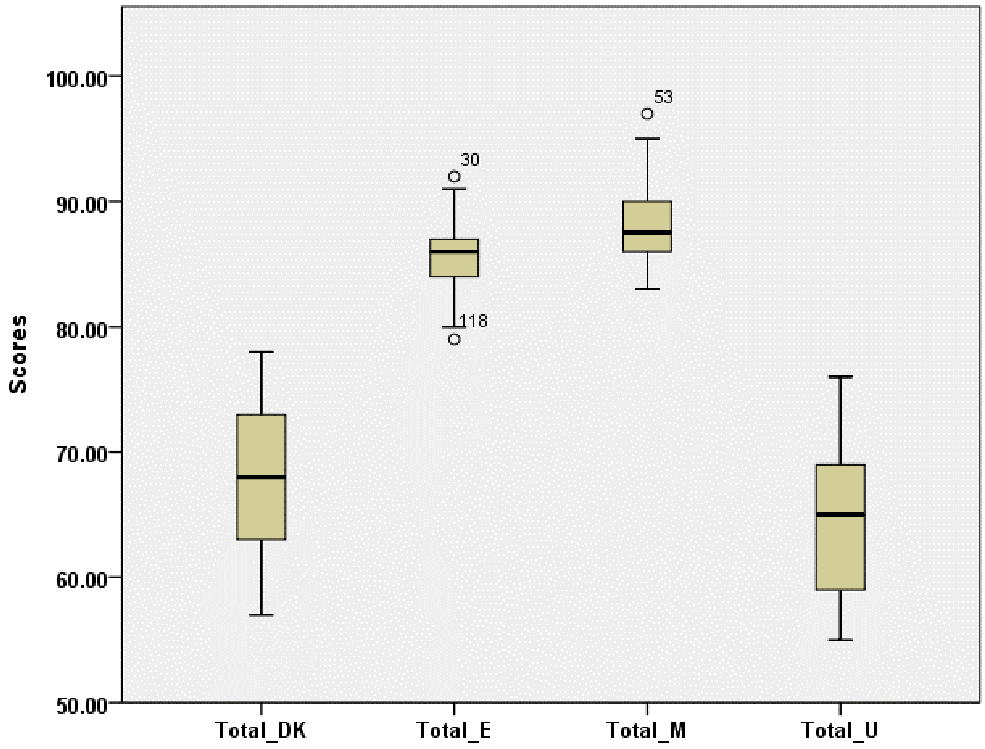


Figure 5.11: Rating distributions of the biased and non-biased interactions from the first group of participants

Total_DK represents lack of knowledge, Total_E represents empathy gap, Total_M represents misattribution and Total_U represents non-biased algorithms

Second group

Table 5.23: Pairwise comparisons between biased and non-biased algorithms in the second group.

Algorithms		Mean Differences	Std. Error	Sig.	95% Confidence Interval for Difference	
					Lower Bound	Upper Bound
Baseline	Misattribution	-1.56	0.05	0.000	-1.7	-1.4
Baseline	Empathy gap	-1.23	0.05	0.000	-1.38	-1.07
Baseline	Lack of knowledge/ideocracy	-1.49	0.54	0.000	-1.63	-1.33

In the Table 5.23, the mean differences of baseline compared to biased algorithms came as negative, which means and as shown earlier, the means of the biased algorithms were higher in all interactions compared to the baseline. The Sig. value came as 0.000 (<0.05) which indicate the statistical significance of data.

A whisker plot (5.12) has made using the mean differences between non-biased and biased ratings. The whisker plot shows (5.12) the distribution of the ratings based on overall means for each of the algorithms. The Table 5.23 shows the overall means of three factors for each of the algorithms. In other words, it can be called the influence graph as the differences of overall means of algorithms can be easily spotted, and therefore, the influences of algorithms can be seen.

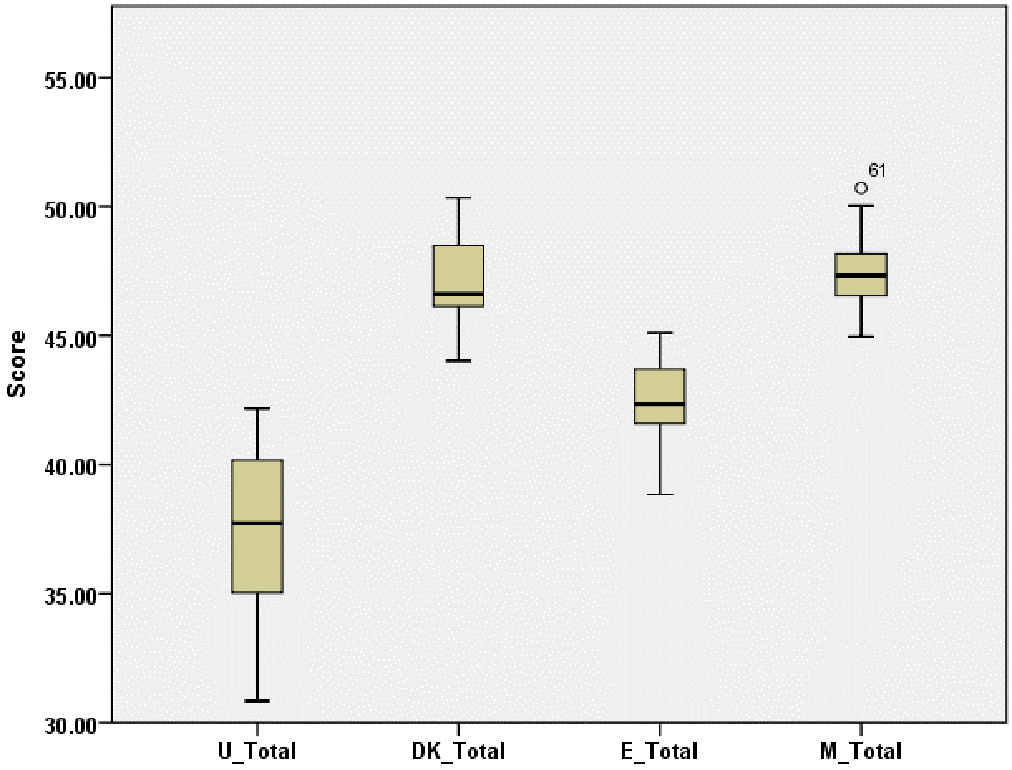


Figure 5.12: Rating distributions of the biased and non-biased interactions from the 2nd group participants.

DK_Total represents lack of knowledge, E_Total represents empathy gap, M_Total represents misattribution and U_Total represents non-biased algorithms

5.3.4.7 Pairwise comparison of means *based on three interactions*

This section compares algorithms based on the three interactions. The purpose of this section is to understand the influence of the algorithms over a long period of time.

First group

Table 5.21 shows the means for each of the interactions. These means are from descriptive statistics, which are based on the means of each of the factors in corresponding interactions.

Table 5.24: Average means of algorithms in each of the interactions in the first group.

Algorithms	Interactions	Means
Baseline	1st	14.51
	2nd	12.11
	3rd	10.67
Misattribution	1st	16.51
	2nd	15.89
	3rd	15.16
Empathy Gap	1st	15.37
	2nd	12.53
	3rd	14.47
Lack of Knowledge/Ideocracy	1st	16.04
	2nd	14.8
	3rd	16.16

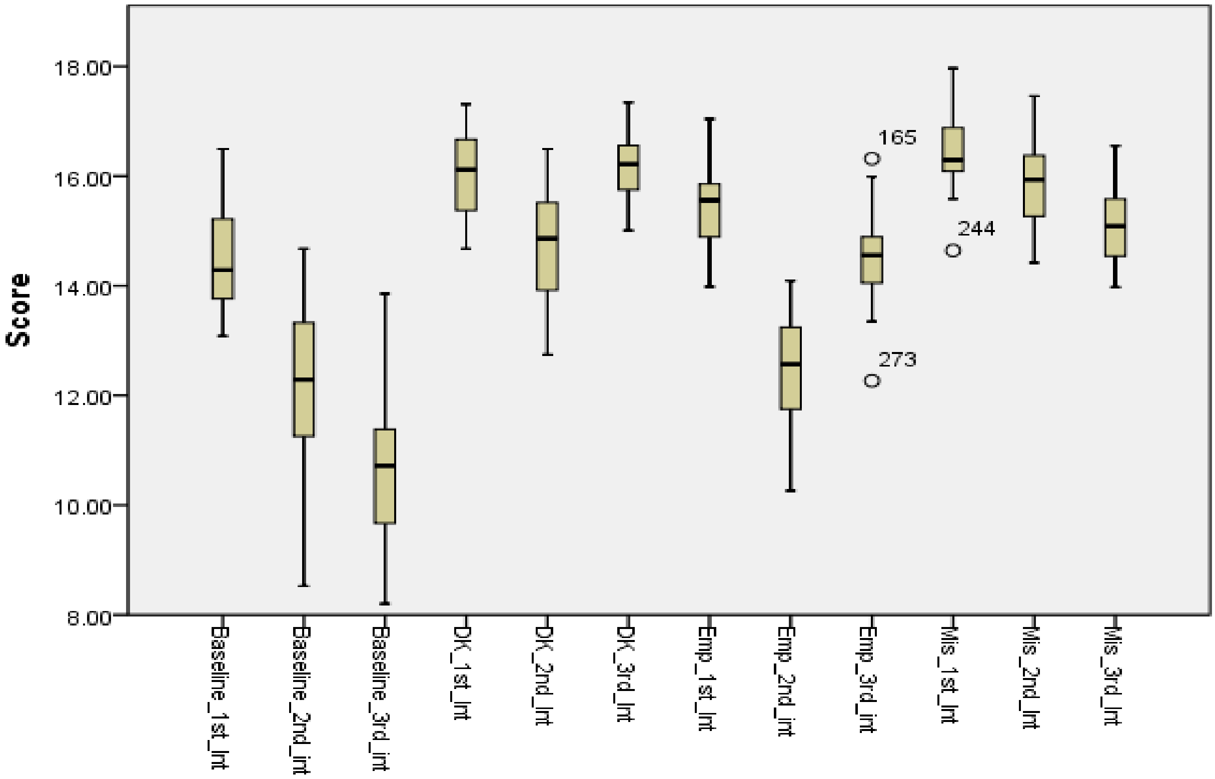


Figure 5.13: Overall means of all algorithms in each interaction in the first group.

DK represents lack of knowledge, Emp represents empathy gap, Mis represents misattribution and Baseline represents non-biased algorithms

In Table 5.24, the means were calculated in the 4 (algorithms) × 3 (factors) ANOVA. The table shows the increase and decrease of ratings of the algorithms in each interaction.

A whisker plot (Figure 5.13) was made using the mean differences between non-biased and biased ratings. The figure shows the distribution of the ratings based on overall means for each of the algorithms, along with each algorithm's means for first, second and third interactions. It shows the preferences of the participants for the algorithms in three interactions maintaining two-three weeks intervals between two interactions. From Table 5.20 and Figure 5.13, it can be said that biased algorithms get more positive and higher ratings from the participants who interacted with the robot maintaining two-three weeks intervals.

Second group

Similarly, Table 5.25 shows the means in each of the interactions. These means are drawn from descriptive statistics, based on the means of each factor in corresponding interactions.

Table 5.25: Average means of algorithms in each of the interactions in the second group

Algorithms	Interactions	Means
Baseline	1st	13.53
	2nd	10.39
	3rd	9.87
Misattribution	1st	16.23
	2nd	14.86
	3rd	16.73
Empathy Gap	1st	15.92
	2nd	14.25
	3rd	14.63
Lack of Knowledge/Ideocracy	1st	17.44
	2nd	14.83
	3rd	14.9

Figure 5.14 depicts the mean differences between non-biased and biased ratings. It shows the distribution of the ratings based on the overall mean ratings for each of the algorithms, along with the means of the three interactions which were held maintaining two-three weeks intervals between two interactions. From Table 5.21 and Figure 5.14, it can be said that participants preferred biased interactions more over non-biased.

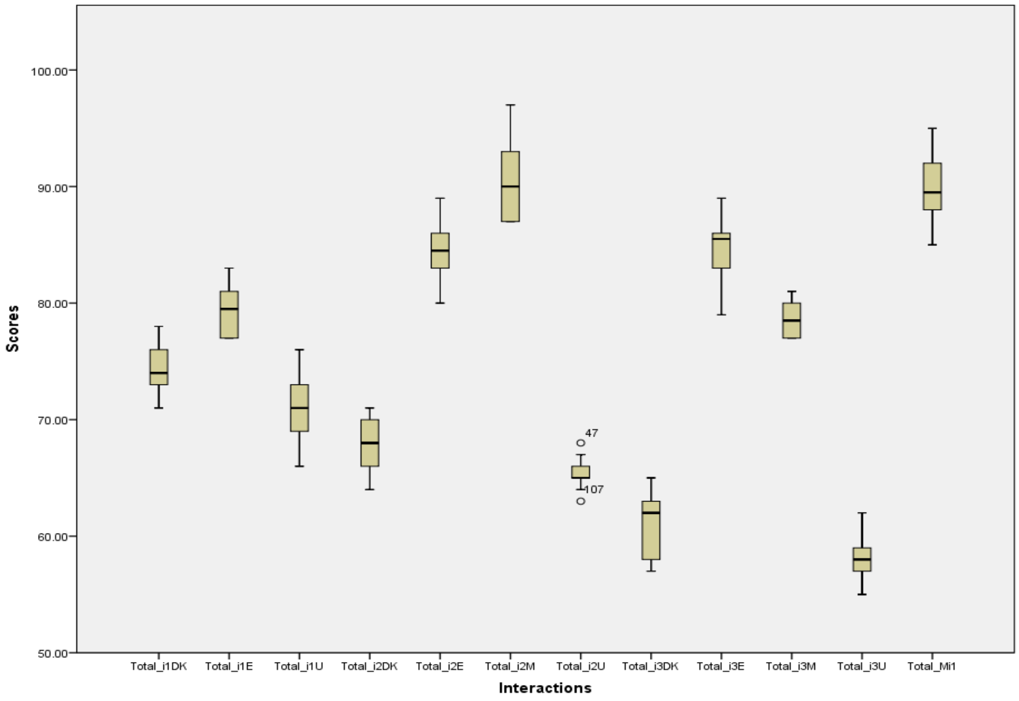


Figure 5.14: Overall means of all algorithms in each interaction in the second group.

Total_DK represents lack of knowledge, Total_E represents empathy gap, Total_M represents misattribution and Total_U represents non-biased algorithms

5.3.5 Statistical result discussions

In this experiment, among 70 participants, 30 were male, 34 female, and six did not specify. Based on age groups, eight were 18–20, while 33 were 21–35, only 19 were 36–60, and 10 participants did not indicate age. Based on the occupations, 33 were students, 28 were working professionals, and nine did not specify. As can be seen, participants are mixed in age, gender and occupation, and limited numbers of participants fell into a specific category. I therefore analysed the data as a whole rather than based on gender, age or occupation.

For better understanding the comparisons of algorithms I summarise all the means of all conditions from the first group into one table, such as:

Table 5.26: Summary of the results from the first group

Interactions	Measurements	Baseline Means	Misattribution Means	Empathy gap Means	Lack of knowledge and ideocracy effects Means
1st Interaction	***Comfort***	4.87	5.59	5.59	5.59
	Likeability	4.92	5.78	5.78	5.78
	Rapport	4.72	5.00	5.00	5.00
2nd Interaction	***Comfort***	3.96	5.37	5.37	5.37
	Likeability	4.18	5.59	5.59	5.59
	Rapport	3.97	4.94	4.94	4.94
3rd Interaction	***Comfort***	3.96	5.11	5.11	5.11
	Likeability	3.41	5.22	5.22	5.22
	Rapport	3.29	4.82	4.82	4.82

In Section 5.3, I have described the analysis results for individual algorithms and for pairwise comparisons. The pairwise comparisons were done comparing the baseline algorithm with the biased algorithms. Table 5.23 shows the average ratings for each factor in three different experiments. Similarly, Table 5.16 shows the ratings from the first group of participants for the misattributed algorithm. Table 5.18 shows the ratings from the participants of the first group for the empathy gap algorithm, and Table 5.120 shows the average ratings of three interactions from the lack of knowledge or ideocracy effects algorithm interactions of the first group of participants. Tables 5.17, 5.19, 5.21 and 5.24 show the average ratings for the second group of participants for the baseline, the misattribution, the empathy gap and the lack of knowledge or ideocracy effects algorithms, respectively. These tables show that the mean ratings of the variables are generally higher for the biased algorithms than for the baseline.

The overall means for each of the algorithms were shown in the Table 5.22, and the graph was plotted for Figure 5.11. The differences were calculated based on the baseline and are negative, which indicates that the participants rated biased algorithms higher than baseline algorithms. There are, however, not many differences in ratings between the factors in the biased algorithms. As stated earlier, for the second group's data, a mixed (3×4) ANOVA was carried out on the dimensions (3) and algorithms (4). The means from three dimensions came out as shown in the Table 5.22. In Figure 5.11, it can be seen that with each of the dimensions, the participant's ratings were varied, but compared to the baseline, the participants rated the biased algorithms much higher.

For better understanding the comparisons of algorithms I summarise all the means of all conditions from the second group into one table, such as:

Table 5.27: Summary of the results from the second group

Measurements	Baseline Means	Misattribution means	Empathy gap Means	Lack of knowledge and ideocracy effects
Comfort	3.41	5.42	5.23	5.58
Likeability	4.10	5.17	5.12	5.14
Rapport	3.75	5.34	4.57	5.00

Tables from 5.14 to 5.21 show a descriptive analysis of each dimension for individual algorithms. From those tables, it can be seen that the mean of the factors or dimensions is higher for the biased algorithms than for the baseline. There were stable positive increments in the ratings for each of the dimensions in all biased algorithms over the baseline algorithm. For example, the comfort dimension mean ratings increased from 3.42 (baseline) to 5.58 (lack of knowledge or ideocracy). For likeability, the mean ratings increased from 4.1 (baseline) to 5.17 (misattribution), and for the rapport dimension, the mean ratings increased from 3.75 (baseline) to 5.34 (misattribution), which are statically significant increases of 2.17, 1.07 and 1.59 (95% confidence interval, $p < 0.0005$). There was a statistically significant difference between means, so the null hypothesis can be rejected and the alternate hypothesis accepted. The graph in Figure 5.13 and 5.14 plots the distributions of the three factors or dimensions for all algorithms. As can be clearly seen in the graph, the participants rated the biased interactions much higher than the baseline interactions.

Figure 5.14 shows the distributions of the means plots from each algorithm based on the three interactions. In fact, all biased algorithms have means that are higher than the baseline. The graph was generated in the repeated measure test in SPSS using post-hoc analysis. From this graph, it can be said that the interactions based on biased algorithms were more popular than the baseline across all interactions.

In all the pairwise comparisons, the Sig (*p* value) was <0.05 (i.e. a very small probability of this result occurring by chance), under the null hypothesis of no difference. Therefore, the null hypothesis was rejected, since $p < 0.05$. There is strong evidence of participants preferring biased algorithm interactions over baseline interactions. It can therefore be said that the first group of participants overall preferred the biased algorithms interactions over the baseline interactions. Statistical analysis from both group's data suggest that the biased algorithms were able to influence the participants to like and prefer the biased interactions more than the baseline single interactions. Furthermore, the second group's results (Tables 5.17, 5.19 and 5.21, and Figures 5.12 and 5.14) suggest that such preferences were quite stable, since drops in ratings were much less for the biased interactions than for the baseline interactions.

For the both groups, and for all three biases, the 2-tailed sig (*p* value) from the paired sample *t* test was <0.05 which indicates the significance of the collected data over a large population (Table 5.19 and 5.20), with the sigma $p < 0.001$ (i.e. a very small probability of this result occurring by chance). Also, with $p < 0.001$, there are significance differences between our null hypotheses and alternative hypothesises. Therefore, for the all three biases, the null hypothesis (H0) can be rejected, since $p < 0.05$.

Therefore for all three biased algorithms, the alternative hypothesis H2 is accepted, because for all the cases $\mu1 > \mu2$, where H2 = participant's rating of 'likeability' scores are greater in all the biased robot interactions.

Overall, the participants liked the biased interactions more than the non-biased. From the above statistical analysis, it can be concluded that all the interactions using biased algorithms affected the interaction, and in general the participants significantly liked and preferred the interactions using misattribution, empathy gap and lack of knowledge or ideocracy effect biases with MARC.

5.4 Game-playing interactions with MARC based on the self-serving and humour effects

In this final experiment, I compared a robot's baseline behaviours with its two biased behaviours (i.e. self-serving and humour effects bias) through game-playing interactions. The robot was controlled remotely (Wizard of Oz).

The experiments were based on the following local hypothesises for each of the biases used:

For the self-serving cognitive bias:

If the self-serving cognitive bias is used in social-robot game-playing interactions, then a participant's preference or likability for the robot can be increased due to the human-like behaviours the robot exhibits as a result.

For the humour effects bias:

If the humours effects cognitive bias used in social-robot game-playing interactions, then a participant's preference or likability towards the robot can be increased due to the human-like behaviours the robot exhibits as a result.

To determine the effects of such biases, participants interacted with the robot with such biases and offered their 'likeability' score (μ1), and the participants interacted with the robot without misattribution bias, again providing a 'likeability' score (μ2).

The null hypothesis is that the mean score of the participant's responses in (μ1) is same as the participant's responses in (μ2):

i.e. Hypothesis0: μ1 = μ2

where,

Hypothesis 0 = Null hypothesis

(μ1) = the means of the participant's responses from biased robot interactions

(μ2) = the means of the participant's responses from non-biased robot interactions

The alternate hypothesis is

Hypothesis 2: μ1 > μ2

where,

Hypothesis 2 = participant's 'likeness' scores are greater in biased robot interactions.

Therefore, for the individually biased vs non-biased experiments, if the null hypothesis is rejected, it can be concluded that the participants liked and preferred to interact with the biased version of the robot.

5.4.1 Experiment methods

In this experiment, two bias based interactions were compared with another interaction, developed without the selected biases (Figure 5.15). Two common human biases—self-serving and humour effects bias—are compared with the baseline algorithm.

In the experiments, the robot plays the popular paper-rock-scissors game with the participants. This game was used because it is easy to understand, is played in many countries and is familiar to all ages and genders. In addition to its accessibility, several other features of this game made it appropriate for the selected cognitive bias. The timing of the game is particularly important, since if one player is slower than the other, the first can change his or her decision or adapt it to win by cheating; this possibility was an important feature for the experiments.

In the self-serving bias, people tend to ascribe success to their own abilities and failures to others. To express this bias's effects, I arranged a rock-paper-scissor game-playing interaction with participants where the robot would portray winning as its own achievements and skills while bragging about it. On the other hand, when it failed, the robot always made excuses and blamed participants. As this bias also can influence self-esteem, to show this in interactions, the robot frequently bragged about winning and accused participants of cheating when it lost.

The humour effect is a memory bias in which people tend to focus on humorous events more than non-humorous ones. In the experiment, this bias was compared with a non-biased interaction to explore these effects in the robot in human-robot interactions. To express the humour effect bias, the robot told various light-hearted jokes to the participants in a game-playing interaction regardless of the outcomes of the game so that the participants could focus on the bias effects more than the games. In the non-biased interaction, the robot did not tell any jokes in the conversational part so that participants could find the differences between the two interactions.

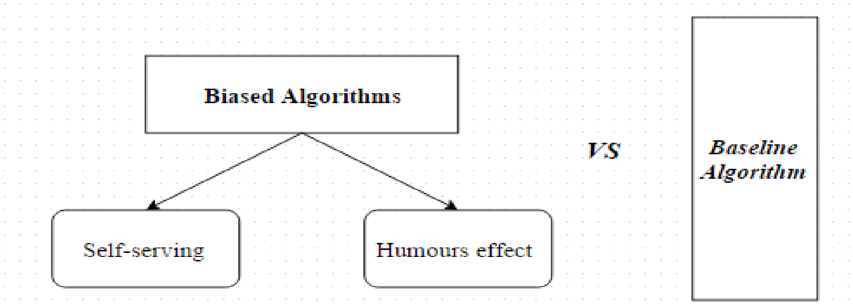


Figure 5.15. Experiment structure.

In this experiment, a group of participants interacted with an individual algorithm over three interactions.

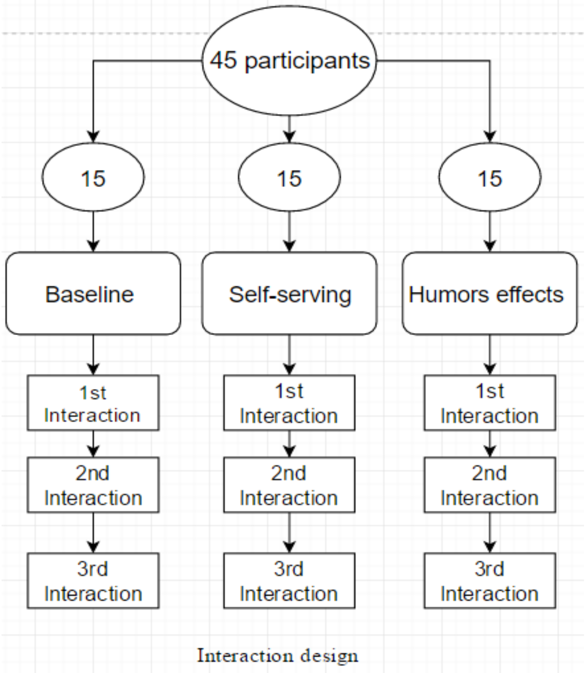


Figure 5.16. Interaction structure.

As depicted in Figure 5.16, participants interact with a selected algorithm in all interactions.

The steps of the interactions were as follows:

1. Participant enters in the room.
2. Participant sits in front of the robot (MARC).

3. Robots greets the participant and waits for response.
4. Participant replies.
5. Robot asks name, address and wellbeing and waits for response.
    a. In the non-biased interaction robot asks and states names, address and wellbeing.

    b. In the self-serving biased interaction robot asks and states names, address and wellbeing.

    c. In the humour effects interaction robot asks and states name, address and wellbeing with humour.
6. Participant replies.
7. Robot then moves to the game-explanation stage and asks participant if he or she know about rock-paper-scissors.
8. Participant replies.
    a. In case the participant knows the games and rules, robot then moves to the game-playing stage.

    b. In case the participant does not know the game,

        i. in non-biased interaction, robot explains the games and its rules and ask if the participant has understood;

        ii. in the self-serving biased interaction, robot explains the game rules and ask if the participant has understood; and

        iii. in the humour effects interaction, robot explains the game rules and ask whether the participant has understood.

    c. Participant replies with ‘yes’ or ‘no.’

        i. If yes, then in all interactions robot moves to the game playing.

        ii. If not, in non-biased interaction, robot explains the game rules again.

        iii. If not, in self-serving biased interaction, robot asks participant to pay more attention and explains the game rules again.

        iv. If not, in humour effects interaction, robot makes a joke and explains the game rules again.

    d. Robot then then moves to the game-playing stage.

9. Robot and participant both draw their hands.

    a. If robot loses,

        i. in non-biased interaction, robot congratulates participant and asks for the next round;

        ii. in self-serving interaction, robot denies losing, makes excuses, calls the participant cheater, and demands next round; and

        iii. in humours interaction, robot congratulates the participant with humour and asks for the next round.

    b. If robot wins,

        i. in non-biased interaction, robot encourages participant and asks for the next round;

        ii. in self-serving interaction, robot brags and asks for next round, and

        iii. in the humour effects interaction, robot makes a joke and asks for the next round.

    c. In the case of a draw,

        i. in non-biased interaction, robot declare the hand a draw and asks for the another round;

        ii. in self-serving interaction, robot denies draw, demands that it be recognized as the winner, and asks for next round; and

        iii. in the humour effects interaction, robot makes a joke and asks for the next round.

10. Robot states the final results:

    a. If robot loses,

        i. in non-biased interaction, robot congratulates participant and requests to play next time;
        ii. in self-serving interaction, robot denies losing, makes excuses, calls the participant cheater, and demands to play again next time; and
        iii. in humour effects interaction, robot congratulates participant with humour and asks to play again next time.
    b. If robot wins,
        i. in non-biased interaction, robot encourages participant and asks to play again next time.
        ii. in self-serving interaction, robot brags and asks to play again next time; and
        iii. in humour effects interaction, robot makes jokes and asks to play again next time.

11. Participant replies.
12. Robot then moves to the farewell stage:
    a. In non-biased interaction, robot thanks participant and invites him or her for further interactions.
    b. In self-serving interaction, robot brags about winning or makes excuses for losing and invites participant for further interactions.
    c. In humorous interaction, robot makes jokes and invites participant for further interaction.
13. Participant replies.
14. Robot asks participants to complete the questionnaire and waits for a response.
15. End of the interaction.

5.4.2 Participants and grouping

A total of 45 participants were randomly chosen from responses to advertisements. The participants were divided into three groups, where each of the groups were assigned only one algorithm to interact for three interactions (Figure 5.16). The number of different gender, race and age groups were maintained as equal for both groups. For all interactions, 15 participants were selected to interact with each of the individual algorithms (biased or unbiased) throughout the experiments. Such interactions should reveal the effects of each individual algorithm in human–robot long-term interactions.

A demographic of the participants is as follow:

Table 5.28: Details of the participants

Gender	Age	Occupation
	18 -20 => 5	
Male 22	21 – 40 => 25	Student => 24
Female 15	41 – 60 => 8	Professionals => 11
No answer 3	No answer 2	No answer 5
40	40	40

5.4.3 Data collection and measurements

Data was collected based on the measurements of the participant's preference towards the biased and unbiased versions of the robot MARC. A seven-point Likert questionnaire was used to measure four conditions such as: comfort, pleasure, likeability and rapport.

As with the previous MARC conversational experiments, the measurements required were as follows:

- Participant's preference for the algorithms (to get the participant's preferences), and
- Changes in preference over a long period of time (to find out whether such preferences for the interactions changed over time).

Three out of four factors are used in the previous MARC experiment. As this experiment was based on game-play interactions, another factor was added in the questionnaire to determine whether the participants enjoyed game playing with the robot. This factor was called ‘pleasure’, how pleased participants were in game-playing interactions. Based on these four dimensions I calculate the participant’s overall preference and how it changes over time. Similar to the previous MARC experiment, in the current experiment, the measurements of the baseline algorithm were developed to compare with other two biased algorithms. The questionnaire was created from various sources; however, for the factors used previously, I used the similar questions, and the new ‘pleasure’ factor questions were collected from Kim and Mutlu (2014) (see appendix).

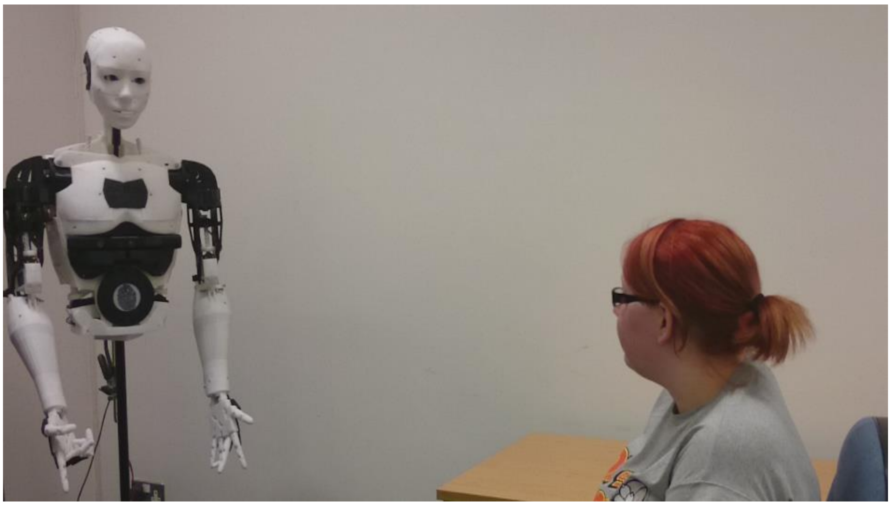


Figure 5.17a: Participants interact with self-serving biased MARC.

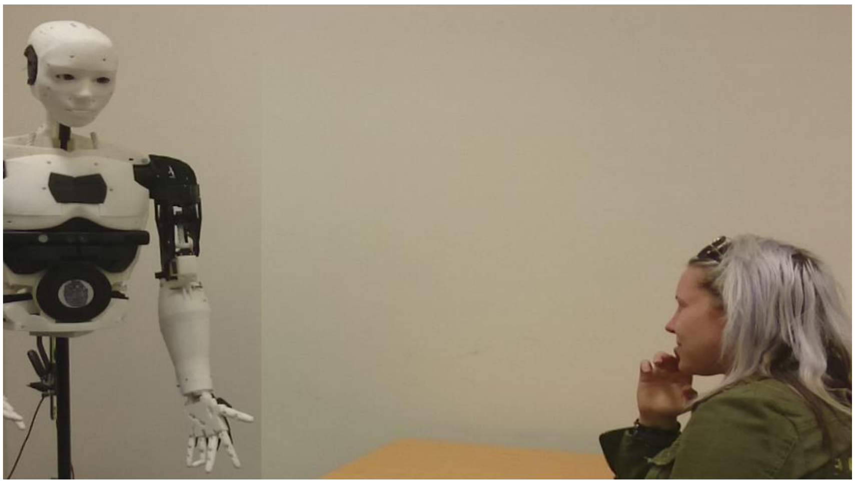


Figure 5.17b: Participants interact with humours effects biased MARC.

5.4.4 Experiment scenario

Two experimental scenarios are shown in Figure 5.17a and 5.17b. Initially, participants were very excited to play games with the robot. One situation from the experiment can be described as an example. ‘Participant 1’ interacted with the self-serving biased robot three times. In the first interaction, she was very competitive in games against MARC. In her first interaction, MARC told her to pay attention when she wanted to know the game rules for the second time; and she reacted with concern. In the first game round, MARC won and bragged about the win, but she wanted to continue the game. In second round, ‘Participant 1’won, and MARC told her, ‘I was not ready. I want to play more.’ She replied positively, ‘Okay, we both won one time.’ The next round, ‘Participant 1’ won again, and MARC said, ‘Why you are always drawing hand after me. It’s not fair. I want to play more.’ Then she replied, ‘No, I was not, we drew hand together. But I can play more.’ The next hand, MARC won and said, ‘I win. If you play fair, then I will win all times.’ ‘Participant 1’ expressed her surprised and replied, ‘No you draw hand after me this time. I want to play once more.’ MARC lost the next hand, and said, ‘You tricked again. How can I play if you trick every time?’ MARC also added, ‘Do you want to play more?’ But ‘Participant 1’ did not want to play more and replied, ‘No, I won 3 times and I win.’ The robot replied, ‘Next time I will win. This time you drew hands after seeing my hand. Next time I will be more careful,’ and jumped to the next part of the conversation. In her comment, ‘Participant 1’ wrote, ‘It was a quite unexpected experience but I’m glad that I won the games. I am looking forward for next meeting.’

In humorous algorithm, MARC makes jokes when games, which made participants chuckle and made the interaction more enjoyable. One of the participants commented after their first interaction, ‘I enjoyed a lot. Although robot managed to win but I am not at all sad the robot made the whole thing very enjoyable.’

In case of non-biased interactions, robot showed general friendly behaviours without biased effects. For example, ‘Participant 2’ who was in the baseline algorithm group came for the first-time interaction with MARC and found the robot as very ‘formal’. If MARC wins a game, it says, ‘I win this hand. Do you want to play more?’ ‘Participant 2’ replied, ‘Yeah sure. Let’s play again.’ MARC won again and said, ‘I win again. Do you want to play more?’ ‘Participant 2’ might have wanted to win at least one hand, so he replied, ‘Yeah why not? I want to play more.’ This time ‘Participant 2’ won and MARC said, ‘Congratulation you win. Do you want to play more?’ ‘Participant 2’ replied, ‘Yes! Finally, I beat you. But I will play next time not now.’ In his comment, Participant 2 wrote, ‘I like the game playing specially when I beat him. Overall, he’s very formal.’

5.4.5 Statistics of the game-playing interactions based on self-serving and humour effects biases in MARC the humanoid robot

5.4.5.1 Justification of the statistical analysis procedure

In this experiment, the participant’s grouping was similar to the second group of the previous conversational MARC experiment (Section 5.3), so a similar mixed ANOVA analysis between algorithms, factors and interactions was performed. The pairwise statistics and graphs were calculated by alternating between and within the subjects of the ANOVA. All results were statistically significant, which means that the Sig. (p value) came to 0.000 (<0.05) (i.e., a very small probability of this result occurring by chance).

In all graphs, the *x*-axis represents the factors and algorithms and the *y*-axis represents the marginal means.

5.4.5.2 Statistics of baseline interactions

Means of the four factors from three interactions are shown in Table 5.29.

Table 5.29: Means of all factors in baseline algorithms.

Measurements	**Means**	**Std. Error**	**95% Confidence Interval**	
			Lower Bound	**Upper Bound**
Comfort	4.35	0.14	4.05	4.64
Pleasure	4.97	0.14	4.68	5.27
Likeability	4.82	0.2	4.4	5.23
Rapport	3.97	0.17	3.63	4.31

The standard error is the standard deviation of the distribution of the mean. It also describes the lower and upper limits of the ratings, with a 95% confidence level.

5.4.5.3 Statistics of self-serving biased interactions

Means of the four factors from three interactions are shown in Table 5.30.

Table 5.30: Means of all factors in self-serving algorithms.

Measurements	Means	Std. Error	95% Confidence Interval	
			Lower Bound	Lower Bound
Comfort	5.73	0.14	5.43	6.02
Pleasure	5.38	0.14	5.09	5.77
Likeability	5.59	0.2	5.17	6.00
Rapport	4.89	0.17	4.53	5.22

Table 5.30 shows the means of all factors. Compared to baseline (Table 5.29), the means values are higher for the self-serving interactions.

5.4.5.4 Statistics of humours effects biased interactions

Means of the four factors from three interactions are shown in Table 5.31.

Table 5.31: Means of all factors in the humour algorithms.

Measurements	Means	Std. Error	95% Confidence Interval	
			Lower Bound	Lower Bound
Comfort	5.29	0.15	4.9	5.58
Pleasure	5.6	0.18	5.22	5.96
Likeability	5.19	0.17	4.85	5.54
Rapport	5.37	0.13	5.11	5.64

The means for the humour effects interactions are higher than the baseline (5.29). Therefore, in this game-playing experiment, the mean ratings from the biased interactions are higher than the baseline interactions for all three interactions. In the next section, I compare the means of baseline interactions with those of biased interactions.

5.4.5.5 Pairwise comparison of means based on *overall ratings of four factors*

The pairwise comparisons between baseline and biased (humour effects and self-serving) algorithms are shown in Table 5.32.

Table 5.32: Pairwise comparison between algorithms.

Algorithms		**Mean Differences**	**Std. Error**	**Sig.**	**95% Confidence Interval for Difference**	
					Lower bound	**Upper Bound**
Baseline	Self-serving	-0.87	0.17	0.000	-1.31	--0.42
Baseline	Humours	-0.83	0.17	0.000	-1.28	-0.39

The differences are calculated in negative numbers, as the means for self-serving and humour effects biases are higher than the baseline.

A whisker plot (Figure 5.18) uses the mean differences between non-biased and biased ratings, showing the distribution of the ratings based on overall means for each of the algorithms.

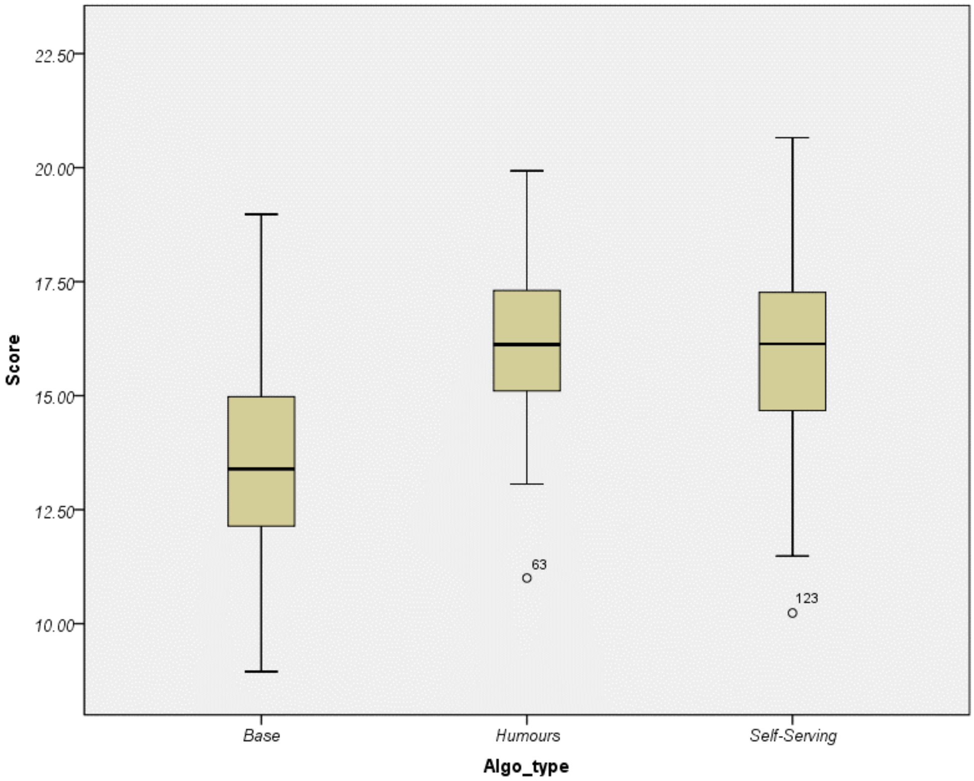


Figure 5.18: Distributions of the means of all factors in three algorithms.

Overall, in the three interactions, all the factors scored higher for self-serving and humour effects biased than for the baseline interactions.

5.4.5.6 Pairwise comparison of means *based on three interactions*

Table 5.23 shows the average rating of means from all factors in each of the interactions. In this case, the within-subject factor was chosen interactions numbers, and the between-subject factor was chosen algorithms types.

Table 5.33: Pairwise comparison between algorithms in each interaction.

Algorithms	factor1	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Baseline	*1st Interaction*	5.43	0.15	5.13	5.74
	2nd Interaction	4.44	0.13	4.17	4.71
	3rd Interaction	3.69	0.14	3.4	3.97
Humours	*1st Interaction*	5.61	0.15	5.31	5.92
	2nd Interaction	5.32	0.13	5.05	5.6
	3rd Interaction	5.14	0.14	4.85	5.43
Self-Serving	*1st Interaction*	5.83	0.15	5.53	6.14
	2nd Interaction	5.26	0.13	4.99	5.54
	3rd Interaction	5.08	0.14	4.79	5.37

The above table shows that means from baseline interactions are lower than for the biased interactions; however, in a few cases the baseline means of ratings are higher than the biased (e.g. baseline first-interaction mean is 5.43, higher than for either humour or self-serving on the second and third interactions), but if the comparisons are based on corresponding interactions (e.g. baseline first interaction mean with biased first-interaction means), then biased interactions means are always higher than the baseline.

A whisker plot (Figure 5.19) using the mean differences between non-biased and biased ratings shows the distribution of the ratings based on overall means for each of the algorithms, for each interaction. A continuous decrease is observable in the means for all interactions. However, such declines are greater for the baseline interaction. Compared to baseline drop, the biased means decreased a very small amount. Therefore, it can be said that participants who interacted with biased algorithms liked these biased based interactions, and their overall preference changed less than that of the participants who interacted with baseline behavioural algorithms.

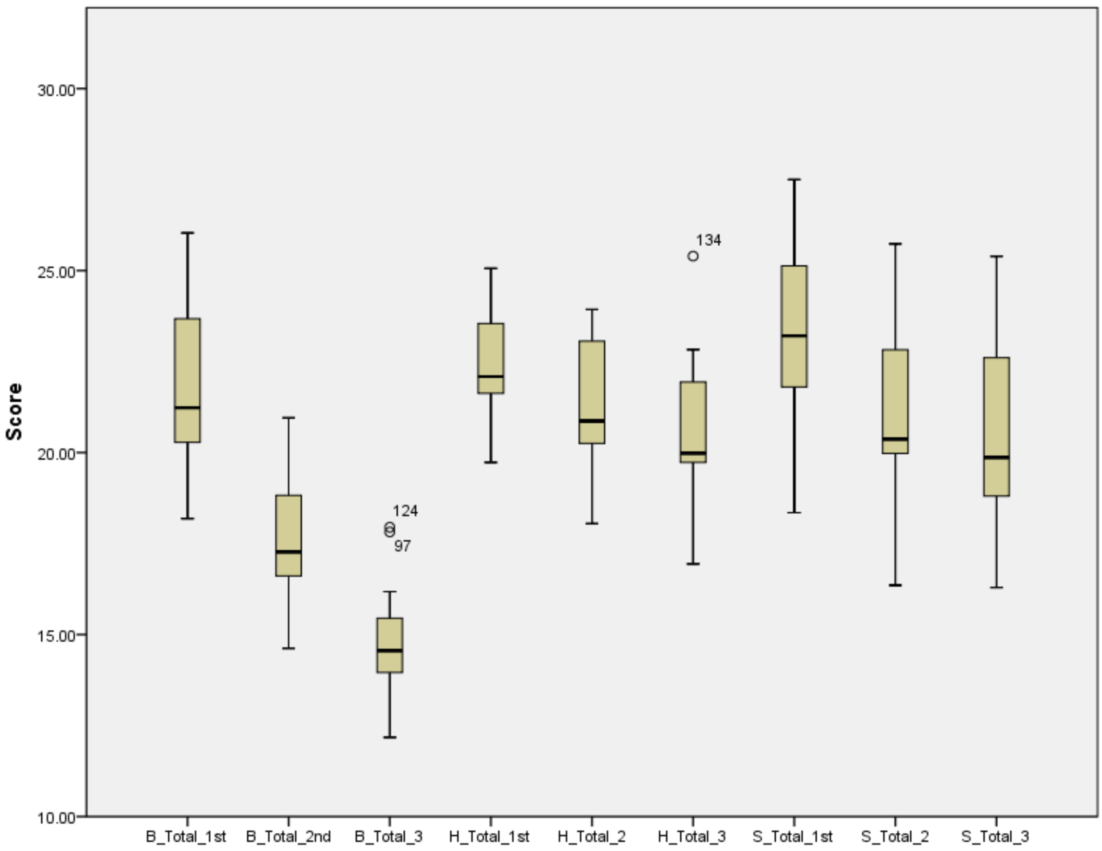


Figure 5.19: Distributions of average means of three algorithms in all the interactions.

5.4.6 Statistical results discussion

In this experiment total 40 participants were selected among them 22 were male and 15 were female and three did not specify anything. Based on age, 18-20 years were five, 25 were between 21 and 40 age group, and 8 were between 41-60 and two did not specify. Based on occupations, 24 were students, 11 were professionals and five did not specify. Similar to the previous experiments, I analyse data as whole rather than based on gender, age and occupations.

In this experiment, the overall statistical analysis shows a very positive influence of self-serving and humour effects biases. In Tables 5.32 and 5.33, it can be seen that in all the factors the biased robot scored higher than the baseline. Between the self-serving bias and the baseline, the differences of the means for the four factors (Table 5.32) are 1.38 for comfort, 0.77 for experience likeability, 0.97 for rapport and 0.41 for pleasure. Self-serving bias scored high in all factors, but as seen, in the pleasure factor, the difference from the baseline is much less than with the others.

As can be seen form the rating distributions in Figure 5.19, in the first interaction, there is very small difference in the average means of both algorithms; however, in the second and third interactions, participants' ratings dropped significantly for baseline (21.75 − 14.76 = 6.99). On the other hand, self-serving ratings dropped in the second and third interactions, but compared to baseline, the drop was relatively small (23.34 – 20.33 = 3.01). In the interview, participants from the self-serving bias group commented on the robot's excuses for losing a game, as at the beginning they thought MARC was genuinely not ready (or the Wizard) or that they drew their hand faster, but when MARC started making excuses over and over, they found it 'interesting' and 'entertaining'. In the second and third interactions, self-serving-biased MARC accused them of cheating whenever it lost a game—in one participant's opinion, this behaviour was highly entertaining and enjoyable.

From Table 5.32, the 2-tailed sig (*p* value) from the paired sample *t* test was <0.05, which indicates that the collected data is significant over a large population. From the Table 5.32, the Sigma $p < 0.001$ (i.e. a very small probability of this result occurring by chance). For both biased interactions compared to non-biased, as the $p < 0.001$, there are significance differences between our null hypothesis and alternative hypothesises. Therefore, the null hypothesis (H0) can be rejected, since $p < 0.05$.

Furthermore, for all the biased interactions compared to non-biased, the H2 is accepted because for each biased versus non-biased interaction $\mu 1 > \mu 2$, where H2 = participant's 'likeability' scores being greater in biased robot interactions.

For better understanding the comparisons of algorithms I summarise all the means of all conditions from this experiment into one table, such as:

Table 5.34: Summary of the results

Measurements	**Baseline Means**	**Self-serving Means**	**Humours effects Means**
Comfort	4.35	5.73	5.29
Pleasure	4.97	5.38	5.6
Likeability	4.82	5.59	5.19
Rapport	3.97	4.89	5.37

Therefore, it can be said that the participants overall liked the biased interactions more than non-biased. From the above statistical analysis, it can be concluded that self-serving and humour effects biases affected the interaction, and overall, the participants significantly liked the interactions using the self-serving and humours biases with the robot MARC.

5.5 Discussions of the statistical analysis

The experimental results show that in both ERWIN and MyKeepon experiments, participants favoured biased interactions. In both experiments, all participants interacted with both biased and non-biased algorithms and preferred the biased versions of the interactions. In these two experiments, participants were randomly chosen in groups to determine their first interactions to be biased or non-biased. So participants who interacted the first time with the biased behavioural algorithm were scheduled to interact with unbiased algorithm in their second interactions. Therefore, the overall preference in favour of biased algorithms in both experiments was the result of careful comparisons of both algorithms by the participants. In both experiments the biased algorithms was able to form preliminary companionship with the participants.

In the experiments using MARC, non-biased robots did not show any common human-like biased behaviour other than generic impressions, which might be expected from a robot by the participants; this could be the reason for the difference in ratings between biased and baseline algorithms. In the biased interactions, however, participants found behaviours such as forgetfulness, emotional differences, idiotic behaviours, denying, blaming and humorous comments, all very common in their lives. This commonality could be the reason for the participant accepting and liking the biased interactions more than non-biased, and giving such high ratings. When participants found these familiar behaviours in the biased versions of robots, they may have found it easier to relate to the robots, which might be why the biased algorithms received higher ratings than the non-biased.

From the statistical results described in this chapter, it can be concluded that the robots with cognitive biases were able to develop better interactions than those with non-biased interactions. In all the experiments, biased interactions were more popular and garnered more positive responses from the participants. Participants liked ERWIN’s behaviours more when it misattributed information, liked MyKeepon’s overly playful or non-respondent behaviours, and liked MARC’s forgetfulness, emotional differences and idiotic conversations more than their unbiased behaviours in similar interactions. Biased behaviours were also favoured in different situations such as winning, losing or tying a game, in which the robot brags about a win but blames the participants or the external causes for losing or drawing.

From the experiments and analysis of collected data, however, it can be concluded that cognitive biases in robots can enhance their capacity for companionship with users over a robot without biases. These results show the significance of using bias-based algorithms in the social robots. Such robots need to interact with their users many times over a long time. Such interactions may become repetitive and boring where the robot always shows nice, mechanical behaviours such as always agreeing with users, perform any task nicely, never showing any type of tiredness, or always obeying the user’s commands. Such social robots could be useful as household workers, but they may be difficult to see as friends or companion. As in reality, friends and companion are those with whom we like to spend time, who can agree or disagree with us and who can act more lively than a repetitive machine. Therefore, if the biased algorithms are used in such future household robots who could forget, brag about doing something good, show competitive behaviours in games, demonstrate different emotional states than users, and try to cheer the user if he or she is sad, then the companionship between users and their household robots would be much easier to develop. The cognitive biased algorithms suggested in this thesis may help to develop such imperfectly friendly social robot.

5.6 Summary

This chapter has described four of the experiments thoroughly. At first, for each experiment, a local hypothesis was made; I then briefly described the experimental methodology, including groups of participants, data collection, and a brief experimental scenario for each of the biases used. I discussed the statistical analysis of each of the corresponding experiments and justified the statistical methods I used. At the end, I briefly discussed what I learnt from all the experimental results and analysis. After reading this chapter the reader should know and understand how I carried my experiments and analysed the data collected. I have also explained the experimental results and what important information I found after analysing the data.

In the next chapter, I elaborate the discussion of the findings from the statistical analysis and answer and explain the hypothesis and research questions.

Chapter 6:
Discussion of the Hypothesis and Research Questions

6.1 Analysing the first research question, based on statistics

> *RQ 1. Despite the robot's appearance, functions and features, can a robot interact with humans in ways similar to human–human interaction by engaging with humans over both the long term and the short term, drawing on cognitively biased behaviours?*

The first research sub-question can be answered by summing up the current studies and experimental results. In the experiments, three different robots have been used, each of which had different types of appearances and functionalities (e.g. ERWIN, MyKeepon, and MARC) (Figure 6.1). The robot ERWIN looks has a head that stands on a metal base and can express five prototype emotions by moving its jaws and eyebrows. MyKeepon, on the other hand, can be described as a small, yellow toy which with various dance moves and jumping abilities, and MARC is a humanoid robot and able to perform various gestures as humans do.

We applied cognitive biases to these three robots and in all our experiments, and the robots were engaged in face–to-face interactions multiple times with the same participants. Previously discussed data statistics (see Chapter 5) show that participants overall preferred the biased algorithm robots more than the non-biased. The misattribution bias was developed on both ERWIN and MARC, and the empathy gap bias was developed on both MyKeepon and MARC. All interactions were compared with their non-biased version of interactions.

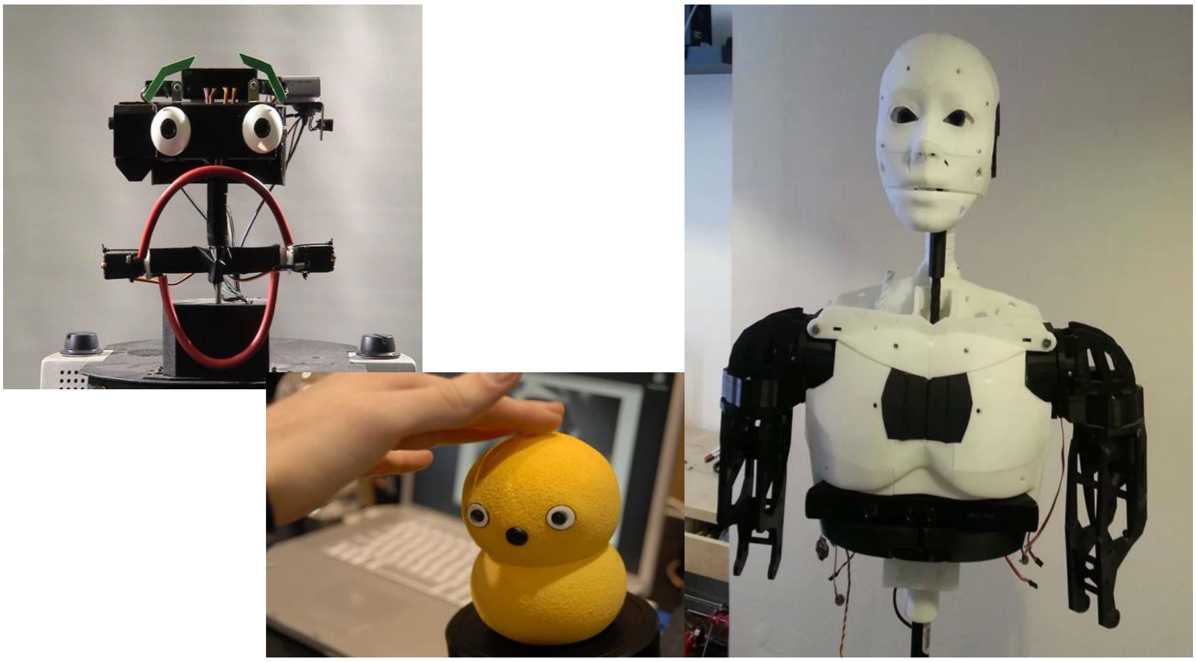


Figure 6.1: Three robots (from the left to right: ERWIN, MyKeepon and MARC) which are was used in the experiments.

In the Chapter 5, the statistics from the ERWIN experiments showed that participants rated higher the misattribution interactions than the non-biased (see Tables 5.5, 5.6, and 5.7). In Figure 5.4, the distributions of the biased and non-biased interactions can be seen, suggesting that in the biased interactions, the ratings were distributed with much higher scores than the non-biased interactions.

In the Chapter 5, the statistics from the MyKeepon experiment (see Tables 5.8, 5.9 and 5.10) showed that the participants preferred biased interactions more than the non-biased. In these experiments the data measured targeted likeability. In Figure 5.6, the distributions of the two interactions can be seen, which indicates that in case of biased interactions, the ratings were distributed with much higher scores than the non-biased interaction ratings.

Based on the two experiments with the two varied robots (ERWIN and MyKeepon), it can be said the participants were engaged in interactions with both biased and non-biased robots, but preferred interacting with biased robots over non-biased ones. The behaviours of these biased robots were developed based on the main components of the misattribution and empathy gap biases. Participants preferred to interact with the robot which demonstrated such biased behaviours (e.g. forgetfulness, differences of understanding the right emotions). so in those two experiments, the cognitive biased behaviours showed by the robots made such interactions preferable to the participants, despite the robot’s appearances.

The results from the current experiments show that robots which carry social interactions using cognitively biased behaviours tends to achieve more popularity than robots not expressing such biases in social interactions. Based on Breazeal's (2003) definition of the social robot, cognitive biases improve social engagement between humans and robots, and therefore we should consider studying more biased behaviours in robots for use in human–robot interactions. However, it should be remembered that the choice of bias should be careful, so that the robot remains 'sociable' and 'socially suited' (Fong, Nourbakhsh, & Dautenhahn 2003). As the interactions with the social robot should feel '…like interacting with another person' (Breazeal, 2000), the use of cognitive biases in the social robots must be considered, as it makes robot behaviours more human-like, with human characteristics faults and imperfections.

Also, in the third and fourth experiments, MARC was used, a robot with a humanoid shape. The statistics of MARC's interactions showed that participants rated all biased interactions more favourably than non-biased (Tables 5.11 to 5.18 and the pair-wise comparisons are from 5.19 to 5.22). Figures 5.11–5.14 show the distributions of the ratings in different interactions, which suggests that the ratings distributions in biased interactions were higher than the non-biased ratings distributions. The misattribution and empathy gap biased interactions in MARC's experiments got much higher ratings than non-biased interactions. As these two biases were previously tested in two different robots, following a similar procedure, and the results were similarly positive, such positive responses were not specific to the robot, but the way those robots interacted (i.e. showing cognitively biased based behaviours in interactions).

In my experiments, three differently shaped and sized robots were used, and participants liked the biased versions of the interactions with such robots more than non-biased version interactions. The experiments indicate that the use of cognitively biased behaviours influences the likeness of the interactions. The result of this current research is supported by the study of Reeves and Nass (1996), who state that human can easily engage with modern technologies, and based on that ease, cognitively biased behaviours helped participants to anthropomorphise such robots more than the robots that did not express such behaviours. Therefore, participants liked the biased version of the interactions more than non-biased interactions. This argument is also supported by Fink (2012), who suggests that robots with social cues can hold people's attention, which could help to develop acceptances of robots by their users. People express biased behaviours in interactions followed by various verbal-nonverbal social cues – similarly, in my experiments the robot expressed such behaviours using verbal-nonverbal cues which were easy to understand by the participants and might help them to relate with the robot.

In my first and second experiments, the robots were not humanoid, but cognitivel biased behaviours in robots helped robots to project and reflect such humanlike behaviours in their interactions; that might help participants relate to robots more easily than non-biased robot interactions and to accept the biased version of the robot more than the non-biased version. I discussed that anthropomorphic forms can solve design issues and can be used to project and reflect human values in other objects and robots (DiSalvo & Gemperle, 2003). Therefore, in order to gain more acceptance from the users, the social robot should be developed based on human social skills and values and should consider express different biases behaviours so that people can easily relate to it.

The results from the current experiments show that robots which carry social interactions using cognitively biased behaviours tends to achieve more popularity than robots not expressing such biases in social interactions. Based on Breazeal's (2003) definition of the social robot, cognitive biases improve social engagement between humans and robots, and therefore we should consider studying more biased behaviours in robots for use in human–robot interactions. However, it should be remembered that the choice of bias should be careful, so that the robot remains 'sociable' and 'socially suited' (Fong, Nourbakhsh, & Dautenhahn 2003). As the interactions with the social robot should feel '…like interacting with another person' (Breazeal, 2000), the use of cognitive biases in the social robots must be considered, as it makes robot behaviours more human-like, with human characteristics faults and imperfections.

6.2 Analysing the second research question based on statistics

RQ2. With cognitive biases introduced in a robot, is it possible to develop anthropomorphically biased behaviours in robots to influence human–robot interactions?

In the first experiment, misattribution biased interaction was compared against non-misattributed interactions on the robot ERWIN (see Section 5.1). In this case, the participant's overall ‘likeability’ to the algorithms was measured. To measure ‘likeability,' questions were asked based on two dimensions: ‘acceptance of the robot’ and ‘participant’s experience of likeability'.

From the Tables 5.5, 5.6 and 5.7, it can be seen that the participants rated more favourably the misattributed interactions than the non-misattributed interactions. On average, the ratings were higher by at least 40 points (i.e. four scales in the rating chart) for each factor. In the experiments, participants interacted with both biased and non-biased algorithms. Therefore, they compared their interactions and rated them accordingly. The participants commented that they found it amusing that the robot could forget and get confused about previous interactions. On the other hand, in the non-misattributed interactions, the robot interacted with the participants without forgetfulness or any types of mistakes; therefore, it was possible that the participants were expecting that the robot would say their information correctly, as they originally told the robot in the previous interaction. The differences in ratings also suggests that the participants enjoyed the interactions with the biased robots and accepted them more than interactions with the non-biased robots. In both factors, participant’s higher ratings in biased interactions reveal their preferences for the misattribution-biased algorithm more than the non-biased basic algorithm.

In the MyKeepon experiment, the empathy gap algorithm was compared with the baseline algorithm of the robot MyKeepon (see Section 5.2). From the Tables 5.8, 5.9 and 5.10, it can be seen that participants responded with attitudes similar to those of expressed in the previous ERWIN experiments (i.e. higher ratings for the biased interactions than non-biased). The data suggests that participants liked both interactions, but preferred biased robots over unbiased. MyKeepon used hot-cold empathy gap behaviours, and to express such behaviours, the robot became overly joyous, overly sad, responsive or unresponsive during interactions. The overall means difference between biased and non-biased was almost 3 (2.97), meaning that participant’s ratings in the biased interactions were at least 3 points higher than in non-biased interactions. The analysis showed that biased interactions were favourable to the participants. In this experiment, the measurements were based on similar factors to the previous ERWIN experiments (e.g. likeability). In both experiments, the participants rated higher the biased interactions; however, the statistics show that there are differences in ‘higher’ ratings between ERWIN and MyKeepon robot interactions. These differences may arise from the differences in the ratings scales, because in the experiment with ERWIN and misattribution bias, a Likert scale of 1–10 was used, while in the experiment with MyKeepon and empathy gap bias, a Likert scale of 1–5 was used, and the subsequent statistical analysis was done based on rating scale. In both experiments however, participants preferred the biased algorithms over non-biased algorithms.

In the conversational experiments with MARC, three biased algorithms (misattribution, empathy gap and lack of knowledge/ideocracy effects) and one non-biased algorithm were used to evaluate three factors: comfort, rapport and likeability. There were two different groups, and a total of 70 participants interacted with the robot in three-time interval interactions (see Chapter 5.3). From Tables 5.11–5.18, it can be seen that the means from the biased interactions were higher than baseline in both groups. The first group of participants interacted with all the algorithms and could compare all of the biased and non-biased interactions. Therefore, their ratings should have reflected the comparisons between all algorithms, and these ratings show that all biased interactions were preferred over the baseline. In Figures 5.11, 5.12, 5.13 and 5.14, the distributions of ratings and the differences in ratings of each algorithms in both of the groups can be seen. The second group of participants interacted with an individual algorithm three times, and their ratings should reflect the experiences of interacting with similar algorithms several different times. These ratings showed that the participants who interacted with the baseline algorithm rated the biased algorithm almost the same in the first interaction, but their ratings dropped significantly in the second and third interactions. The participants interacted with the biased algorithms, which rated constantly high, although in the second and third interactions, all ratings dropped compared to the first interactions. The rate of decline, however, was much less in the biased algorithms compared than in the baseline algorithm. Thus, participants liked the biased interactions and rated them higher in terms of likeability, comfort and rapport.

From the game-playing interactions with MARC, the statistics (shown in Tables 5.23–5.27) showed that the means were higher in the biased algorithms for factors such as comfort, pleasure, rapport and likeability. In this experiment, participants interacted with single algorithms and played the rock-paper-scissors game in three interactions. Participant’s comments and ratings suggest that they enjoyed MARC’s reactions from the robot for winning and losing game hands. Participants pointed out that it was surprising to them that the robot argued and bragged in the games. Moreover, the robot denied drawing a game hand or asking participants to play fair—these types of behaviours are common among people—and they enjoyed the fact that the robot could argue after losing a game as well. Figures 5.18 and 5.19

show that the participants rated consistently higher the biased interactions than the non-biased, and their ratings did not drop as much as for the unbiased ratings.

On the other hand, participants experiencing the baseline interactions did not find such behaviours in their interaction. The robot responded to losing or winning games without arguing or bragging. In the first interaction, participants ratings were similar to those of biased interactions, perhaps because the robot was playing rock-paper-scissors with participants, but in the later interactions, the baseline algorithm showed similar non-competitive behaviours, which could have affected the participant's experiences, making game more boring, affecting their ratings. The robot's behaviours for all the algorithms did not change much in the second and third interactions, so the responses of robot were the same for the algorithms in all interactions for winning or losing in games. Despite that, participants in the biased algorithms found their interactions competitive and enjoyable and rated them higher in terms of pleasure, comfort, rapport and likeability.

I discussed in Chapter 3 the robot called My Real Baby and how such a robot baby expressed facial expressions and different childlike behaviours, which helped adults and children to anthropomorphize the robot toy. The results from the current experiments also show that expressing humanlike behaviours could help people to anthropomorphize the robot, and making it easy for people to relate to. Certain humanlike behaviours and facial expressions in My Real Baby helped it to relate with the robot; now, based on the current experimental results, if the robot could show few of the cognitive biased behaviours, that could definitely enhance the experiences of its users. Therefore, it will be possible in future to create such toy robots to express facial expressions and more humanlike behaviours, including biases and imperfections, which will help people to understand the robot easily and to 'enhance the fantasy'.

The similar thoughts are applicable to my previous discussions (Chapter 3) about the robot pet AIBO, Paro and the computer program Eliza – each greatly affected people in terms of interactions between human–robot and human-computer program. From the evidence of the current experimental results, it can be said that people tend to accept interacting with machines if those machines can express behaviours familiar to humans. Eliza expressed understandable empathic behaviours while communicating; AIBO and Paro expressed pet-like behaviours with which people are familiar. In the current experiments when robots showed biased behaviours such as forgetfulness, bragging, cheating and so on, people understood such behaviours because people are already familiar with them. People liked interacting with biased robots more than non-biased robots, suggesting that people enjoy communicating with robots that express such common and understandable behaviours. It can therefore be said that if social robots could express such biased behaviours, their likeability may increase beyond what is possible for robots that do not express such behaviours. As the current experiments have proved that humour can increase robot likeability, the uses of such behaviours can enhance a robot's likeability to its users, especially in the case of care robots used in childcare and elderly care. To use such biases in childcare and elderly care, though, further studies are required, as my participants were all adults (18–60); none were children (18 or younger) or the elderly (60+).

In sum, participants found it interesting and enjoyable to interact with the biased algorithms. In Chapter 4, the step-by-step differences between the robot's behaviours in each of the algorithms has been detailed. Such differences in algorithms clearly affected in the interactions with participants, producing differences in the ratings. Their ratings suggest that they like the robots with cognitively biased behaviours—making mistakes, forgetting information, expressing different emotional states, bragging, blaming, and making jokes—more than the robot without such behaviours. It can therefore be concluded that cognitive biases can make robot behaviours more human-like, influencing human–robot interactions.

In the Section 3.3, I stated why I am considering cognitive biases in social robots, and after analysing the experimental results, I conclude that some of the biases have the potential to increase the likeability of social robots. This could be because of human beings having characteristic flaws, biases and imperfections, and finding similar flaws and biases in robots could help these robots to be liked more easily. It is possible that in seeing similar forgetfulness, bragging and other biased behaviours in robots, people found it easy to understand their interactions and found the robots easy to relate to. I enquired before starting my experiments whether people would like such biased behaviours in robots and thus accept such behaviours; the experimental results show that people usually like to interact with robots having such biased behaviours more than robots without such behaviours. People liked the imperfect and biased behaviours they notice in people in their daily lives to have in the social robots. I stated that people find it difficult to relate to others who are too perfect, and as an example, I offered the character of Mary Sue character (Chapter 3.3.1), noting that people like to know about realistic characters rather than too perfect which feels as mechanical, unrealistic characters. The current experimental results indicate that such an account of human preference holds in human–robot interactions as well. If the robot shows humanlike behaviours with biases and characteristic faults, then people find it more relatable than more mechanical versions of the robot.

Although we have not perfected the technology of social robots, to make a robot more social and likeable for people, it is possible to develop some of the selected biased behaviours in robots. I discussed developing trait-based personalities (Lee et al., 2006), moods and emotions (Dautenhahn et al., 2009) in social robots, and evidently biases are also common in people (Schacter (1999), Kruger and Dunning (1999)); the current research concludes that using biased behaviours in robots increases their likeability; therefore, it can be safely said that we should consider developing select cognitive biases in social robots to make those more relatable and likeable to people.

6.3 Analysing the third research question based on statistics

RQ3. Will cognitive biases help humans to develop human–robot interactions over the long term?

This question can be answered by analysing the conversational and game-playing interactions of MARC the humanoid robot and participants. In the current research the word 'long-term' has been used to describe the intervals between two experiments which were used in my third and fourth experiments. Usually the time intervals between two experiments were two weeks—based on factors such as participant's availability, room and robot preparation. As there were minimum two-week intervals between two interactions, to complete the whole set of experiments (i.e., three interactions between robots and participants) required at least a month, and I called these two experiments long-term experiments. It can therefore be said that the term 'long-term' in my experiments does not refer to long-time continuous interaction, but it meant the whole duration of the experiment including three interactions and the interval times (i.e. two weeks or more) between two consecutive interactions.

In the first conversational experiment, participants interacted with MARC over three time interval interactions. Participants were divided into two groups in this experiment, where in the first group participants interacted with all four algorithms, but in the second group, participants interacted with only one algorithm. The data (shown in Tables 5.11–5.18) from both groups, show that the participants preferred the biased algorithm interactions more than the baseline interactions in all four factors of comfort, experience, likeability and rapport. It can be seen that in the first interactions, participants rated the interactions as almost the same, but their ratings dropped in the second and third interactions. The drops in ratings were greater for the baseline algorithm. In the biased algorithms, such as ideocracy and the empathy gap, ratings dropped in the second interaction, but increased again in the third interaction. For the misattribution-biased algorithm, ratings gradually dropped in all three interactions, but it should be noted that the level of decrease was low compared to the baseline interactions. The reasons for this difference might include differences in algorithms and their unique behaviours based on the biases for different interactions. It is clear from the graph, however, that the participants rated more highly the biased algorithm interactions for all three interactions compared to the baseline. For the second group of participants, the differences in the ratings of all dimensions were very clear, compared to the first group. There are differences in the biased interaction ratings as well, but in all the interactions, participants' ratings were much higher for the biased algorithms.

In the game playing experiments with MARC, the interactions were based only on an individual algorithm. From Figure 5.19, the means distribution graph for all interactions, it can be seen that the overall ratings for the biased algorithm were higher in the interactions. In the first interaction, the means were at almost the same level all algorithms, but these means then diverged in later interactions. Although the ratings dropped for both the biased and baseline algorithm, the decrease is greater for the baseline interaction than for the biased interaction. Figure 5.19 makes clear that participants preferred the biased algorithm interactions over the baseline interactions, in all three interactions.

From these two methodologically divergent experiments, it can be said that the cognitive biases played an important role in making the ratings (i.e. participant's preferences) higher in the biased interactions. The three interactions were performed between the participant and the robot, spanning a two-month period, in order to provide long intervals of time to allow participants to reflect on their previous interactions. Our data suggests that even in such time interval interactions, participant ratings favoured the biased algorithms. Ratings changed regarding biased interactions, and sometimes the biased ratings were lower than previous interactions, but for the baseline algorithm, the ratings always dropped from the first interaction to the final interaction. From these two long-term experiments (i.e. interval times between two interactions at least two weeks or more) with MARC, it can be said that the cognitive bias can make for better performance in human robot interaction.

The findings from the current research can answer the Hayoun's (2014) question: 'How can we interact with something "more perfect" than we are?' (Section 3.3). Other than the obvious technological barriers, the use of humanlike character flaws and biased behaviours in the social robots may make the robot more ‘interactable’. Humanlike cognitively biased behaviours in social robots could break the sense of ‘more perfect’ than what we are now, so that people can interact and relate to social robots easily, leading to future human–robot long-term companionship. However further studies are required to find out the potentials of cognitive biases in developing long-term human-robot companionship, as the current thesis only shows results from short interactions performed over long time intervals.

Based on such observations, cognitive biases can be used for social robots where it must have frequent interactions with people, for example in health care robots. I discussed in Chapter 3 how elderly care robots need to feature easy-to-understand interfaces and anthropomorphic behaviours to provide physical and psychological comfort, mainly because of elderly people’s health issues. I have discussed robots such as PARO being used in studies in Japanese care homes as a pet for elderly. From current experimental results and notes from literature reviews (DiSalvo & Gemperle, 2003), it can be said that if we develop certain cognitive biases in such robots, elderly people may be able to anthropomorphize the robot more easily and allow it to more readily entertain them—so that they can relate to the robot. In the current experiments, however, I did not use any care robot, nor the healthcare environment, to study the effects of the biases in elderly care; as such, to examine this idea in health care, a proper investigation is needed using elderly care robots and proper care environment (e.g., care home, hospitals).

6.4 Discussions of the main hypothesis

> *If a selection of cognitive biases can be used in social-robot interactions, then human preferences about and fondness towards robot can be increased as a result of the behaviours exhibited by the robot.*

The main hypothesis can be analysed and answered by combining the explanations of all three research questions. From discussions of the RQ1, RQ2 and RQ3, the following can be concluded:

1. Cognitive biases can effectively influence human–robot interactions positively, regardless of the shape, size or colour, functions, or features of the robot.

2. Cognitive biases can amicably make the robot’s behaviours human-like, which can influence the degree of human–robot interaction.
3. Cognitive bias can produce better outcomes in long-term human–robot interaction than can the interactions without such biases.

Based on the above three conclusions, cognitive bias positively improves human–robot interactions, even over the long term. Developing cognitive biases in social robot’s behaviours can help to strengthen the fondness of the users in human–robot interactions.

In all of four experiments, with five different cognitive biases on three different robots, participants preferred the biased interactions over the non-biased. As seen in the first and third experiments, using misattribution bias was quite popular in ERWIN and MARC. The empathy gap bias made a significant difference in ratings in the experiments with MyKeepon (second experiment) and MARC (third experiment). The lack of knowledge or the ideocracy effects made for better conversations with MARC (third experiment) than baseline conversation. Humorous effects made differences in the participant’s ratings in game-playing interactions with MARC (fourth experiment). Participants found playing games with self-serving MARC more enjoyable. All the experiments suggest that all biased algorithms were popular to the participants and could garner more attention from users than interactions without such biases.

On the other hand, the baseline algorithms were less popular in the interactions. The participants preferred interactions with the biased versions of the robots in all the experiments. This result can be explained by the participants’ preference of interactions in which they observed similarities with natural human behaviours—forgetfulness, inability to understand another’s emotional state, inability to understand its own lack of stupidity, bragging, blaming and humour—all of which helped participants to relate with the robots. Such behaviours in robots managed to get the participant’s attention, and the participants found it very interesting to interact with such a robot that shows typically biased behaviours. Even though the ratings were dropped in second and third interactions for all biased and non-biased interactions in the conversational and game playing experiments with MARC, such rating drops in biased interactions were small compared to those of the non-biased interactions. Participants felt more comfortable interacting with biased algorithms, as such behaviours are known to them. Participants were pleased to recognise such biased behaviours in robots, and they liked the biased interactions more than the non-biased. Such comfort, pleasure and likeability in biased interactions helped them to build rapport with the biased robots, beyond what was established with the non-biased versions of the robots. In all the experiments, participants liked the biased algorithms and rated them higher in all factors. Therefore, cognitive biases can influence human–robot interactions positively in the long term. As described in the methodology section, the study was performed only with adult participants (aged 18–55); therefore, this study does not specify the effects of such cognitive biases on elderly and children.

6.5 Discussions and lessons learned

At the beginning of this thesis (Chapter 1), I set several objectives for this study. As my approach to developing human-like skills and behaviours in social robots by developing various cognitive biases is novel, choosing appropriate cognitive biases for the investigation was very important. However, in Section 3.5, I established certain principles to follow in choosing the biases for this study. The chosen cognitive biases have various effects on humans, and to develop such behaviours it was important to understand which of the behaviours should be developed in robots and how. I discussed previously that only the main components of the biases were developed in robot, and I described how such behaviours can be easily recognised by people in human–robot interactions. I did not manipulate the control systems of the robots; rather, I developed the biased behaviours based on the robot's own interactive abilities (e.g. verbal, non-verbal, emotional and gestural expression, and other movements performed by the robots). To perform the actual participant-robot interactions and to understand the responses from the participants clearly, I set various targets for data collection (i.e. likeability, comfort, pleasure and rapport) and asked participants relevant questions on these four factors to get their clear view. Furthermore, the methodology I used clearly distinguishes between interactions based on biases and non-biases by comparing several interactions and collecting data from similar questionnaires. I also monitored the changes in participant's responses over time interval interactions and how that information could help in developing biased skills and behaviours in the social robots. After the investigations of the five selected cognitive biases in practical human–robot interactions and analysing the collected data statistically, I conclude that these cognitive biases have an important role to play in making human–robot interactions more likeable for users. Based on the statistical analysis, it can be said that biased algorithms can be used in social robots to enhance their behaviours to make human–robot interactions more enjoyable for the participants.

According to the research conducted in this thesis and the review of various studies on sociable robots and human interaction, several statements can be made. As discussed earlier, the introduction of cognitive biases to a robot will help to ensure that the participant has better and more preferred interactions with the robot than the interaction with non-biased robot. For example, I have discussed the social robots Kismet and KASPAR, which were used in several HRI studies (Breazeal, 2003; Dautenhahn, 2009). In my understanding, based on the experiment results of the current thesis, if the selected cognitive biases can be introduced in such robots, then their interactions with people could be further improved. It would be considered more human if such Kismet and KASPAR could make mistakes and show biased based behaviours in their interactions.

As discussed in Chapter 2, Dautenhahn (2009) has observed that the autistic children had positive responses to the minimally expressive robot KASPAR. From the current thesis, we learn that cognitive biases can make interaction much more enjoyable and likeable for the participants. However, it would be interesting to study how such biased behaviours effect interactions in between autistic people and robots. Because in the real world people behave based on their thinking and judgements, which are affected by the cognitive biases (Bless et al., 2004), it would be important to study human interactions in different situations (e.g. how autistic children behave towards a robot forgetting, bragging, making small jokes in conversations etc.), and the use of cognitive biases in robots could be proven helpful for the autistic children to understand the real world interactions in such occasions.

In Section 2.3, I discussed the faulty robot and the research carried out by Salem, Lakatos, Amirabdollahian, and Dautenhahn (2015), in which they discovered that people do not trust unusual requests from robots and prefer to trust robots in familiar conditions. However, in this thesis, I did not investigate the trust issue between a robot showing biased behaviours and a robot in the 'correct' condition. The interactions we organised were mainly social conversations and short games between participants and the robots; only non-harmful biased behaviours were chosen for the interactions. Consequently, the biases only affected the behaviours which the robots used to interact with participants and not any other findings, but the results of my experiments suggest that people enjoyed the interactions where biased behaviours were shown by a robot. I also discussed the research work of Robinette et al. (2016) in Section 2.3. The authors found that people followed orders in an emergency situation from a robot which had made mistakes. In this way, it is possible that people appreciate a robot which makes mistakes in tasks but shows non-harmful biased behaviours, as shown in my experiments; furthermore, it is also possible that they may feel more trust for the faulty robot with such biased behaviours. However, further studies are needed to investigate this theory.

As the study suggests that the biased behaviours could be popular for conversational and game-playing interactions with people, so the current model of developing cognitive biases may prove useful in other models for developing social skills in social robots. For example, the selected cognitive biases can be used in Meerbeek's (2009) user-centred personality model for domestic robots to improve the quality of the interactions between the user and the personal robot. Meerbeek's (2009) model suggests developing different traits in the robot based on the user's personality profile, and in that profile if it is possible to know the choices of different behaviours for users, in my understanding, selected cognitive biases based on such behaviours could be develop in robots to help users to relate to robots more easily.

The use of cognitive bias may prove helpful in the TAME, as studied by Moshkina et al. (2011), who developed a complex architecture of traits, attitudes, mood and emotions (see Chapter 2). As the experimental results of the current study suggest that the participants preferred cognitive biases in the robot, if such selected biases can be added in such TAME model, then the likability of such model can be improved.

The findings from this research may also help to reduce the uncanny valley effects. The results from the third and fourth experiments with MARC the humanoid robot (Figure 6.1) suggest that participants liked the biased interactions with that humanoid robot, and their 'ratings drops' for their preference towards the biased robot in the second and third interactions were much smaller than in the baseline interactions. These results suggest that even for a humanoid robot which resembles a human, cognitive biased behaviours were able to keep the attention of the participants more than the non-biased behaviours. Therefore, such biases can be used in other humanoid robots, and it is expected that people will find such biased robot's behaviours much more appealing, which could help to reduce the uncanny valley effects; therefore, human–robot companionship would be possible. This claim is actually supported by Walters's (2008), who found that people tend to prefer human-like attributes in robots (see Chapter 3). The cognitive biases are one of the very common human attributes which effect on people's judgements and decisions, and people acts based on such biased decisions. If cognitive biases can be developed in the robots then people may be able to relate with them because of such known attributes developed in robots, which may help to reduce uncanny valley effects.

I have discussed in the Chapter 2 and 3 that anthropomorphism is one of the most popular ways to develop skills and behaviours which people can relate to. I gave examples such as ELIZA, which is a computer program without any embodiment, but was very popular as it showed interest in learning users' personal problems, or Sony's AIBO robot dog, which was also popular as a household pet robot and entertainer. Robot receptionist Valerie could show emotions and expressions, very popular among visitors. In all these examples, researchers developed various anthropomorphic behaviours so that people could understand and relate to them. The results from of first experiment suggest that a machine-like robot ERWIN could make very likeable interactions with participants based on the human-like 'forgetfulness' behaviours, developed based on misattribution bias. In the second, third, and fourth experiments, toy-like robot MyKeepon and humanoid robot MARC showed various cognitively biased behaviours in their interactions, which were also liked and preferred by participants. The cognitive biased behaviours played an important role in making the behaviours of the three different looking robots (Figure 6.1) more human-like and understandable to participants. Cognitively biased behaviours, for example, forgetting, misunderstanding, ideocracy, bragging and humours in those robots helped participants to anthropomorphise the robots, and they liked the interactions. This insight is supported by various studies (Duffy, 2003; Fink, 2012; Lee et al., 2005) suggesting that social cues, emotions expressions, traits based personalities, moods, and other human-like skills and behaviours can help people to anthropomorphise robots. Therefore, from such literature and from the current experimental results, I conclude that cognitive biases in robots can help people to anthropomorphise robots and relate with them.

The results of the third and fourth experiments with MARC indicate that participant's preferences towards biased-based interactions declined much less than their preferences towards non-biased robot interactions. These findings could be discussed in the arguments of Baxter's (2011) notable 'memory' issue in social robots for developing long-term interactions with children. As we have seen that misattribution-biased interactions were preferred by the participants over the interaction without such bias, it also can be argued that, if the misattribution and forgetfulness can be used in robot behaviours, they can help to reduce such issues between human–robot interactions, which could help to develop human-robot interactions.

The results in this thesis suggest that people like social robots with human-like faults and characteristic biases, and robots prone to making the common mistakes that humans make on a regularly. Therefore, social robots should have such characteristics in their interactive behaviours, since such behaviours should lead to the acceptance of these robots in interaction with humans.

The interesting finding of the current research is that people usually do not like those biases used in this studies in other people, but they liked those biased behaviours in our studies. Cognitive biases affect thinking and judgement, and therefore, they influence information which individuals can notice (Pereita, Tenera, & Wemans, 2013). Memory biases, such as misattribution, cause misattribution and forgetfulness of previously known information. Self-serving bias causes people to perceive themselves in overly favourable manners. Among the effects of the empathy gap is the inability to understand others' true emotions. Most of the cognitive biases used in the current experiments are perceived as unattractive in humans but were popular when used in the social robot's behaviours. Biases mainly affect human thinking and thus also their behaviours. On the contrary, people perceive social robots as machines which are incapable of free thought; therefore, certain bias-driven mistakes are unexpected from social robot behaviour. Biased behaviours such as forgetting, misattributing information, effects of the empathy gap, bragging and so on were unexpected in such human-robot interactions, and after the participants experienced such behaviours from the robot, they were surprised. However, such behaviours are quite common in human nature and sometimes are unwelcome by others. Therefore, people's enjoyment of such biases in robots could be the result of their surprise.

The results of the experiments show that the participants enjoyed the effects of the selected biases in the robot during the interactions. To understand their 'liking' to such effects of biases, I discussed various possibilities, including surprise effects of the participants, of such biased behaviours in robots; participants were surprised to experience forgetfulness, bragging, blaming or other developed biased behaviours in a robot which are usually not very common in social robots. Those types of biased behaviours might have surprised the participants, and they liked those interactions more than the non-biased interactions. A question could then arise regarding the human biases in a robot, which makes robot's behaviours popular: what about robot's own limitations and faults – why are those robot-like faulty behaviours unappealing? This question can be answered by addressing the expectations towards robots and the 'naturality' of objects. Studies have reported that people's expectations of a robot may vary based on culture, geographical area, society and individuals' own perceptions. For instance, the Japanese people expect robots to perform household tasks without any intervention, whereas European culture prefers to communicate with the robot using voice commands (Haring, Mougenot, Ono, & Watanabe, 2014). Bartneck, Kanda, Mubin, and Mahmud (2009) suggested that perceptions of animacy and intelligence are closely related, and this relation affects the appearance and behaviours of the robot and creates a perception of intelligence. Kwon et al. (2016) suggested the possibilities of differences in people's expectations towards a robot and its actual intelligence and capabilities. Literature and movies sometimes portray robots as intelligent machines for helping humans or as destroyers of mankind, both of which are far from reality. On the other hand, biases can be said as the shortcuts of human thinking, which affect their judgements. In this way, biases mainly affect human thinking and thus their behaviours. On the contrary, people perceive social robots as machines which are incapable of free thoughts, and therefore certain bias-driven mistakes are unexpected from social robot behaviours. Biased behaviours, such as forgetting, misattributing information, empathy gap effects or bragging, were unexpected in such HRIs, and participants were surprised after experiencing such behaviours from the robot. However, such behaviours are quite common in human nature and are sometimes unwelcome by others. Therefore, that people liked such biases in robots could be a result of their surprise. Similarly, machine-like faults such as Siri and Alexa's faulty voice recognition, system freezes, iPad and iPhone touchpad problems and other faulty behaviours are common in machines and therefore 'natural' in robots. So, when a robot fails to understand a user's voice commands, navigates incorrect paths or fails to recognize familiar faces, we do not get surprise because we know these behaviours are 'natural' in robots.

I discussed the cognitive biases in certain modern technologies, such as Siri, which was accused of having halo effects and abortion biases, both of which allegedly created the wrong impressions among users, but using cognitive biases such as humour at appropriate times can improve the likability of such applications. Although such voice devices and applications used in phones and personal computers (e.g., Siri, Cortana, Amazon Echo etc.) are created to assist people and to be personal, a proper study should be conducted before developing cognitive biases in such voice applications. I discussed earlier (chapter 3.5.5) that people sometime like to have regular news in comedic form (Jasheway, 2016), so if humours effects biases can be developed in such voice assistant to delivering news that can increase user likeness towards such applications.

I mentioned Friedman and Nissenbaum's (1996) study and stated that in future it is possible people will prefer their daily-use technological devices to respond to them based on their own characteristics and behaviours. I used the examples of Google and Facebook, which create person-specific profiles for marketing purposes. In the information age, people tend to argue with others they do not really know and express and fight for their opinions all over social media; this is an example of the self-serving bias (Sedikides, Campbell, Reeder, & Elliot, 1998) I discussed in Chapter 3—sometimes this bias has serious consequences. However, we have seen in our current experiments that, if carefully used, self-serving bias could be very useful in human–robot interactions; in this vein, much more research is needed on such social media and how it is possible to make good use of such biases and make them harmless on the Internet. Similar to these social media issues, to create person-specific experiences, marketing, and applications and to make use of person-specific biases in daily technologies, we need further study. As the current study shows, if carefully used, some cognitive biases can increase social robots' likeability; therefore it is expected that after further study, some of the biases will be found to help to achieve such goals in social media and to create person-specific applications.

In Section 3.3, I discussed strong and weak AI and how people are concern about such concepts. People are concerned about the growing uses of weak AI (e.g., automated voice services, spy cameras, computer surveillance, etc.) in our daily lives and expect catastrophe any time. Although the current research is not directly related to AI, its results can be extended from its current contexts, and it can be said that people like to see humanlike flaws, biases in machines. If the modern technologies which use weak AI could adopt certain human biases and present itself in a more humanlike manner, the concerns regarding such technologies may be diminished. However, further study and research is needed on this topic, as most of the cognitive biases in people have negative effects; such biases in strong AI should therefore be treated with care.

The results from my experiments showed that people liked to interact with robots which showed similar cognitive biased and imperfect behaviours as humans show. For example, people liked when the robots ERWIN and MARC misattributed information, MyKeepon and MARC showed lack of empathic behaviours in game playing, MARC expressed bragging behaviours and told jokes in game playing. Such behaviours are evidently common in human and I discussed in the previous chapter (chapter 3.5), for example, Schacter (1999), Johnson, Hashtroudi & Lindsay (1993) showed the effects of source confusions in people; and Cryptomnesia, Nordgren, Banas & MacDonald (2010), Ariely & Loewenstein (1999) and others showed the effects of empathic gap; Kruger and Dunning (1999) showed the effects of lack of knowledge in human behaviours; Campbell & Sedikides (1999), DeMichele et al

(1998), Shepperd et al (2008) and other scientists showed the effects of self-serving biases; and Strick et al.(2009), Ziv (1988), Rust & Goldstein (1989), Ziv & Gadish (1989), Lauer, Lauer & Kerr (1990), Kuiper, Martin & Olinger (1993), Kuiper, McKenzie & Belanger (1995) and many other scientists showed various effects of humours effects biases in human behaviours. So based on the current experimental results it can be said that if the future social robots can show such cognitive biased behaviours, then people will like such behaviours in robots. The current experimental results can be extended to other technologies which interact with people regularly, for example, I already discussed (Chapter 3.3) about the automated voice processing system such as Siri, Cortana, Amazon echo etc. which can be beneficial by using the effects of selected cognitive biases. Other technologies such as global positioning system (GPS), Satellite Navigation systems, voice navigation system in cars, automated voice customer services, various self-help machines can use selected cognitive biases such as humours effects to improve the interactions with people. For example, interactive voice response (IVR) in banking, retails and other customer care services sounds very mechanical and sometime feels very annoying, but if such IVR system can use humours effects while interacting with people then such services can be improved. This argument can be supported by the works of Martin et al. (2003) where they showed that interpersonal form of humours such as telling jokes, saying funny things, or witty banter can ease and amuse people and improve interpersonal relationships. I discussed in Chapter 3.1. about various domestic robots, and in the future if such domestic robots (e.g. Roomba, Bosch’s Indego, butler bot Charley etc.) can socially interact with people then uses of selected cognitive biased behaviours can improve the interactions with users. For example, if the butler bot Charley interacts with guests using humours and jokes while serving them then such interactions can amuse the guests and as results they will probably visit in the hotel for next time. It is expected that robot receptionists will be used in various places such as museum, airport, shopping mall etc. and if those robots can show humours and other selected biased behaviours then such behaviours can attract more visitors and can improve interactions.

The above discussion is actually found in the current experimental results as well. I discussed the limitations of the MyKeepon experiment in which similar behaviours were used in different sequences to distinguish between biased and non-biased behaviours. It can be seen in the result that due to the unresponsive nature of the non-biased and ‘cold state of empathy gap’, the differences in ratings were not high enough compared to the second empathy gap experiment with the MARC robot. It is possible that, in both interactions, participants thought that the robot’s unresponsive behaviours were the results of the robot’s ‘natural’ faults and therefore the surprise effects were not effective. The surprise effects actually decreased over time, as we can saw decrements in ratings in the MARC experiments where participants usually interacted three times with the robot. The results showed that participants’ ratings fell for some effects of biases in the second and third interactions. It can be said that the surprise effects of those biases faded for the later interactions as the participants were already familiar from the previous interactions with the robot, and they were aware of such effects of biases. Therefore, in later interactions, the same effects did not surprise them as much as in the first interaction.

The current experiment results show that participants enjoyed and developed a preferred relationship faster with biased and imperfect robots than the robots without the bias, and they also show how cognitively biased interactions can foster better relationships with participants than the interactions without such bias. We know that individuals prefer their own ways of interacting based on their own characteristics. Robots with different cognitive biases can fit into these preferences, so the robot could become more likable to individuals. As the experimental results suggest, such biased robots capable of interacting based on the user’s characteristics could become good assistive companions and should be able to sustain human–robot companionship. Biased robots are therefore more likely to improve relationships with participants. As such, biased, imperfect behaviours can be implemented in social companion robots, allowing for an improved interaction in assisting their users.

6.6 Difficulties and limitations

In the MyKeepon experiments, MyKeepon interacted based on general social queues (eye contact, greetings, forging a bond, etc.) using actions like jumping, dancing, staring and making various noises. The biased MyKeepon used the same actions in different sequences to express its biased behaviours so it could occur that the participants sometimes did not notice any significant differences between biased and unbiased interactions compared to ERWIN experiments. However, to better understand the effects of empathy gap in a robot in human-robot interactions, I used conversational interactions using the MARC robot. In this experiment, the robot followed similar rules for expressing hot and cold states of the empathy gap (see Chapter 5.2), but due to conversational interactions combined with the robot’s movements, the bias effects were expressed better than in the previous MyKeepon experiment.

Due to the conversation and different moves of the humanoid robot MARC, it was possible to distinguish the differences from the non-biased interactions much more clearly than in the previous MyKeepon experiment. Therefore, it is expected that the participants understood the differences in biased and non-biased behaviours of the robot in the MARC experiment.

The MARC experiments posed particular challenges; for instance, the 3D-printed robot face was rigid, so the only way to show the expressions with MARC was with body movements. A concrete model of human body movements was needed for different expressions, so that it could be developed into a robot. After consulting with my supervisor and the literature, I followed the gestural model of Bremner et al. (2009), which was relevant to our research, as they studied gestures focussed on human–human conversations. Another challenge was the limited abilities of the robot. Certain gestural actions could not be implemented in the robot because of limitations with motors and movements. Sometimes, the motor would overheat causing the robot to jam. On a few occasions, the 'connect' facility in the robot's chest had to be shut down due to the small room size and the movements of other people in the room. The most difficult challenge was to present different biases and imperfectness effects with limited gestural expressions. It is possible for humans to express certain gestures in many ways, but for a robot with limited abilities, actions and speech had to be used at the same time to make the participants understand the different effects of the biases. As described in the Section 4.2.3, action and speech queues were used to make the expressions.

Another limitation could be not considering to compare the robot's own biases with the participant's. Robot's biases such as looks, functionalities and appearances could have different impacts on the participants and participant's ratings towards the experiments could have been influenced by that. Therefore further study in details on this subject is required to explore the effects of the biases in robot in human-robot interactions.

In addition to the above difficulties, the study has other limitations. The selection of participants based on the chosen age groups and responses from the advertisements were limited. The results of the experiments may be true only for certain age brackets (e.g. 18–60). As the cognitive biases are considered mostly for negative effects in humans, some of the effects, for example, forgetfulness (which was popular among the younger generation) might be so popular among older people (70+). Also, such behaviours may even be dangerous, for example, if the robot forgets to bring the medicine on time, which could negatively affect the user and obviously would represent failure in a personal robot. As descried in the methodology section the study was performed only with the adult participant (18-60), therefore this study does not specify the effects of such cognitive biases on younger and children (less than 18 years old). Furthermore, game-playing activities with competitive behaviours may prove useful, but much study of this subject remains necessary.

Another limitation lies in the selections of biases in the robot. For example, in the current study biases were chosen based on three principles to investigate in a laboratory environment study. It is possible that in the real-time home environment study, such biases may not result in similar popularity. Therefore, further studies are needed on user-profiling-based investigation on the cognitive biases, so that it may be possible to develop a universal model accepted by anyone in any environment.

Chapter 7: Conclusions

7.1 Summary of contribution

This thesis has presented a novel idea for designing social behaviours for the sociable robots intended to improve human–robot interactions, which has potential use in fostering human-robot companionship. This idea of developing different cognitive biases in a robot for human–robot interactions is itself one of the main contributions to the field of social robotics.

The design principles of cognitively biased algorithms discussed in this thesis suggest how to develop main components of such biases in robots, based on robots' own interactive behaviours. I have presented both the theoretical principles and examples of biases development, step by step, based on the conversational and game-playing interactions (Chapter 4). In the future, however, this method could be applied and improved by involving control system configuration.

The investigation results with MARC and the five cognitive biases indicate that participants liked and preferred bias-based interactions more than non-biased interactions, as performed on separate occasions over time. Therefore, in the long term, it is possible to keep participant's engagement with the robot high if cognitively biased behaviours are used in human–robot interactions.

All the experiments described in Chapter 5 shows comparisons between biased and non-biased versions of robots. The experimental results demonstrate that such cognitively biased algorithms for robots can be successfully developed and put into use, and they compare favourably to existing methods for developing cognitive behaviours in robots.

From the research described in this thesis, we learnt that cognitive biases can play an important role in human–robot interactions. The uses of different biases have proven very useful in developing such interactions over a long-term period (i.e. including the time intervals between two consecutive interactions) in our experiments. The results of four experiments with five different biases on three different robots shows that the biases in robots improve quality of interaction despite of the robot's shape, size and colour.

The use of cognitive biases in social companion robots for human–robot long-term companionship is a relatively unique idea, and the experiments described in this thesis are the first to instil cognitive biases in the robots. Therefore, the premise of this research is the main my contribution. The experimental results showed immense popularity for the biases in robots, suggesting that more cognitive biases may be usefully developed in social robots.

The findings of this study should benefit social robotics study and its applications, considering that HRI plays an important role in robotics today. The new approach of making social robots more sociable can be beneficial for the elderly or children who needs carer robots for long periods of time. As the experimental results showed in Chapter 5, the cognitively biased algorithms can be useful in building companionship over a long period of time. The social robots which will use such recommended biased algorithms derived from the results will be able to interact and develop relationships with users better than existing ones. Researchers in social robotics could study this newly approach to uncover critical areas which would broadly improve the human–robot interaction.

7.2 Future work

In Chapter 6, I have discussed two important limitations of the current study: limited age groups and selection of biases. The matter of age group needs further investigation, as the current study shows that the uses of selected cognitive biases can be popular among adults (18 – 60 years), but it does not consider how elderly people (aged over 60) would respond to such biases. Common knowledge suggests that misattribution and forgetfulness effects in robots might be harmful to elderly people who need timely medication and have difficulties remembering events, but such biases, if properly used, could have some entertainment value among the elderly. Similarly, bragging and competitive behaviours in robots could help children to learn in collaborative learning settings, but it is possible that children find such behaviours in robots unattractive and refuse to communicate. Humorous effects, in general, could be used to entertain the elderly and children, but further study must be done to be certain of the effects of such biases. Therefore, future studies involving misattribution, forgetfulness, bragging, competitive behaviours and humour need to be investigated among child–robot and elderly-person–robot interactions.

Another limitation is the selection of the biases, which was made based on several principals for the current study, including importance, effects, understandability, and so on (Chapter 3.5); based on such factors, I chose five cognitive biases to develop in the robots. As I stated in the limitations, such biases were tested in the laboratory environment with limited resources, and participants liked the interactions with the biased robots. It is possible that if such biases could be developed in the robot and used in the home environment, the responses from users could have changed. In future, if possible, I would investigate these biases in home-environment study for a long-term period.

As I stated in the limitations, the experiments using the MyKeepon robot were based on sequences of behaviours, including high or low pitch noises and jumping to express empathic gap behaviours, where in the other experiments the robot made conversation and played tap-on-head and rock-paper-scissor games with the participants. In future, empathy gap bias can be investigated using gameplay or conversational experimentation, and it is possible to further explore the responses of the different age groups of the participants.

I have discussed different models (e.g. TAME, traits-based personality, etc.) of developing humanlike mood, emotions and behaviours in robots in the literature review. In future, further study is required to develop a universal model, which could be based on user profiling or just limited numbers and types of biases in robots which would be accepted by any environment and people of any age. For example, the TAME model (Moshkina & Arkin, 2009) was used to study developing traits, attitudes, moods and emotions in robots and to investigate their effects in human–robot interactions. An interesting area of future study would be to extend such a model, include some of the common cognitive biases and study this upgraded model in human–robot interactions.

7.3 Conclusion

This thesis describes the concept of developing cognitive biased behaviours in social robots for human–robot interactions. This concept has been tested in four experiments and using three different robots over a duration of two years, using different methods. The experimental results suggest that developing selected cognitive biases in social robots can increase the likeability towards the robot in human–robot interactions for people. The contribution of my works can be described as follows:

Conceptual contribution: The idea of developing humanlike behaviours in social robots is not new, and I have provided different examples in the literature reviews where scientists studied developing emotional expressions, moods, traits-based personalities and various behaviours so that people can understand and anthropomorphize the robot easily. The concept of developing different cognitive biases is an addition to such studies, where I showed that behaviours based on five cognitive biases can improve the likeability of interactions.

Psychological contribution: I discussed different studies on human behaviours which showed that cognitive biases exist in humans, and such biases are common aspects of human nature. People's thoughts and decisions are hugely affected by biases, and sometimes biases affect memory and social behaviour. I discussed that despite the many 'imperfections' of human nature, people interact with others and create relationships; therefore dealing with biases and imperfection is one of the common aspects of human interaction. In interaction with social robots, acknowledging their technological limitations, such robots usually lack such behavioural imperfections such as forgetfulness, bragging, cheating, and so on. It is thus possible that people feel such interactions as mechanical. The experimental results described in the current thesis show that people liked the interactions with the biased versions of robots more than those with non-biased robots. These results contribute to the argument that the social robots should show humanlike behaviours, including humanlike biases and imperfections in interactions.

Practical contribution: The current research contributes to the development of social personal robot; based on the experimental results described in this thesis, it is clear that if robots shows forgetfulness, cheating, bragging, humour and other biased behaviours in interactions, people usually like and continue to like those behaviours. Therefore, in the long term, human–robot interaction studies where people interact with the robot over long intervals, developing cognitive biases in robots could maintain people's affection towards robots more than would otherwise be the case.

In our experiments, such human-like behaviours using different cognitive biases were successful in creating initial bonds with the participants who were adults (i.e., aged 18–55). It can therefore be argued that noticing and understanding the biased behaviours in the robot and perceiving such behaviours as likeable might not come from other age groups such as children (up to 18) and older (60 to 80 or over). It is possible that children do not find the forgetfulness of robots likeable and ignore further interactions completely, and it is certain that forgetfulness could be quite harmful for a care robot assigned to taking care for an older person. It is also possible, however, that certain competitive behaviour from robots could motivate children to play learning games with robots, and certain persuasive behaviour could actually help an elderly person to take their medicine. Therefore, the results of the current study cannot be generalized to much younger and older people, and room remains for further research based on behaviours-specific study of children and older age groups to determine the effects of biases. In all the experiments described in this thesis, a mixed group of participants were chosen, and I did not analyse the data based on participant's gender, occupations – rather a generalize view of the whole population has reflected. Therefore the results do not specify the gender and occupations based views on the effects of biases in robots in human-robot interactions. To determine such effects on genders and occupations further study is recommended.

The current study found that adults (aged 18–55) usually like the interactions with the robot with biases such as hot-cold empathic gap. However, use of such biases should be very careful with social robots for care homes and for children; such technologies must be carefully designed, as the gap in understanding between the users and the technology itself could create bias by which the user may feel the technology has a lack of empathy. Therefore, such technologies should be quite empathetic, designed to understand the users and their requests, and they should be able to understand minute details.

The current study was performed in the laboratory and in a very controlled environment where the experimenter was always present. Therefore the study does not ensure the effects of cognitive biases in the home environment without the presence of any experimenter. Further studies are required to find out people's responses towards robot showing the effects of cognitive biases in daily lives.

I hope this thesis helps the reader to understand the necessity, idea, implementation, results and explanation of cognitive biases in social robots. My approach was the very first to study and develop cognitive biases in social robots, and the experimental results showed that certain biases can influence interactions positively. There are more than 200 cognitive biases, however, so this subject demands further study, but as a start, this study has described five cognitive biases that had a positive effect.

Thank you!

Bibliography

Abortioneers, The (2011). *Desembarazarme meets Siri* [blog]. 1 December. Available from http://abortioneers.blogspot.co.uk/2011/12/desembarazarme-meets-siri.html [accessed 8 July 2016].

AbuShawr, B. and Atwell, E. (2015) ALICE chatbot: trials and outputs. *Computación y Sistemas*, 19(4) 625-632. Available from http://dx.doi.org/10.13053/CyS-19-4-2326 [accessed 14 May 2017].

Ackerman, E. (2016) Walmart and five elements robotics working on robotic shopping cart. *Spectrum*, 28 June. Available from http://spectrum.ieee.org/automaton/robotics/industrial-robots/walmart-and-five-elements-robotics-working-on-robotic-shopping-cart [accessed 7 May 2017].

Ali, I. (2018) Personality traits, individual innovativeness and satisfaction with life [in press]. *Journal of Innovation & Knowledge*. Available from https://doi.org/10.1016/j.jik.2017.11.002 [accessed 11 June 2018].

Apple (n.d.) *Siri does more than ever. Even before you ask.* Available from https://www.apple.com/uk/siri/ [accessed 19 May 2017].

Ariely, D. (2008) *Predictably irrational: the hidden forces that shape our decisions.* New York, NY: Harper Collins. ISBN: 9780007256532.

Ariely, D. and Loewenstein, G. (2006) The heat of the moment: the effect of sexual arousal on sexual decision making. *Journal of Behavioral Decision Making*, 19(2) 87-98. Available from doi.org/10.1002/bdm.501 [accessed 14 May 2017].

Aronson, E., Wilson, T. D. and Akert, R. (2010) *Social psychology*. 7th ed. Upper Saddle River, NY: Prentice Hall. ISBN: 9781292021164.

ASIMO (n.d.) *History of ASIMO: The past, the present, the future*. Available from http://asimo.honda.com/asimo-history/ [accessed 14 May 2017].

Ask me anything. 3rd interview by Bill Gates [reddit] (2015) Available from https://www.reddit.com/r/IAmA/comments/2tzjp7/hi_reddit_im_bill_gates_and_im_back_for_my_third/ [accessed 14 May 2017].

Baayen R. H., Davidson R. J. and Bates, D. M. (2008) Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4) 390-412. Available from doi.org/10.1016/j.jml.2007.12.005 [accessed 14 May 2017].

Bartneck, C., Kanda, T., Mubin, O. and Mahmud, A. A. (2009) Does the design of a robot influence its animacy and perceived intelligence? *International Journal of Social Robotics*, 1(2) 195–204. Available from doi:10.1007/s12369-009-0013-7 [accessed 14 May 2017].

Baron, J. (2008) *Thinking and deciding*. 4th ed. Cambridge, Cambridge University Press. ISBN: 9780521680431.

Baroni, I., Nalin, M., Baxter, P., Pozzi, C., Oleari, E., Sanna, A. and Belpaeme, T. (2014) What a robotic companion could do for a diabetic child. In: *The 23rd IEEE international symposium on robot and human interactive communication,* Edinburgh, 25-29 August. Piscataway, NJ, USA: IEEE, 936-941. Available from doi: 10.1109/ROMAN.2014.6926373 [accessed 14 May 2017].

Bartlett, B., Estivill-Castro, V. and Seymon, S. (2004) Dogs or robots—why do children see them as robotic pets rather than canine machines? In: *AUIC2004: conferences in research and practice in information technology,* 28, Dunedin, New Zealand, 18-22 January. Australian Computer Society, 7-14.

Baxter, P., Belpaeme, T., Canamero, L., Cosi, P., Demiris, Y., Enescu, V. and Neerincx, M. (2011). Long-term human-robot interaction with young users. In: *IEEE/ACM human-robot interaction 2011 conference* (Robots with Children Workshop), Lausanne, Switzerland, 6-9 March.

BBC (2011) *Apple denies claims that Siri is anti-abortion* [online]. Available from http://www.bbc.co.uk/news/technology-15982466 [accessed 19 May 2017].

Bellet, P. S. and Maloney, M. J. (1991) The importance of empathy as an interviewing skill in medicine. *Jama*, 226(13) 1831–1832. Available from doi:10.1001/jama.1991.03470130111039 [accessed 14 May 2017].

Bellotto, N., Fernandez-Carmona, M. and Cosar, S. (2017) ENRICHME: integration of ambient intelligence and robotics for AAL. In: In: *Wellbeing AI: from machine learning to subjectivity oriented computing (AAAI 2017 spring symposium)*, Stanford, CA, USA, 27 - 29 March 2017. *Technical Report SS-17-08*, 657-664. [accessed 14 June 2018].

Beresford, B. and Sloper, T. (2008) *Understanding the dynamics of decision-making and choice: a scoping study of key psychological theories to inform the design and analysis of the panel study*. York, UK: Social Policy Research Unit, University of York. ISBN 978-1-871713-24-4.

Bharatharaj, J., Huang, L., Mohan, R. E., Al-Jumaily, A. and Krägeloh, C. (2017) Robot-assisted therapy for learning and social interaction of children with autism spectrum disorder. *Robotics*, 6(1) 4. Available from doi:10.3390/robotics6010004 [accessed 14 May 2017].

Blanson Henkemans, O. A., Hoondert, V., Schrama-Groot, F., Looije, R., Alpay, L. L. and Neerincx. (2012) I just have diabetes: children’s need for diabetes self-management support and how a social robot can accommodate their needs. *Patient Intelligence*, 4 51-61. Available from doi.org/10.2147/PI.S30847 [accessed 14 May 2017].

Bless, H., Fiedler, K. and Strack, F. (2004) *Social cognition: how individuals construct social reality*. Hove, UK: Psychology Press. ISBN: 9780863778292.

Blue Frog Robotics (n.d.) *About Buddy*. Available from http://www.bluefrogrobotics.com/en/buddy/ [accessed 15 May 2017].

Bolton, R. N., Parasuraman, A., Hoefnagels, A., Migchels, N., Kabadayi, S., Gruber, T. and Solnet, D. (2013) Understanding generation Y and their use of social media: a review and research agenda. *Journal of Service Management*, 24(3) 245-267.

Bosch. (n.d.) *Bosch robotic lawnmower*. Available from https://www.bosch-garden.com/int/indego.html [accessed 15 May 2017].

Breazeal, C. (2002) *Designing sociable robots*. Cambridge, MA: MIT Press. ISBN:0262025108.

Breazeal, C. (2003) Toward sociable robots. *Robotics and Autonomous Systems*, 42(3-4) 139-292. Available from https://doi.org/10.1016/S0921-8890(02)00373-1 [accessed 14 May 2017].

Breazeal, C. (2004) Social interaction in HRI: The robot view. *IEEE, Transactions on Cybernetics, Man and Systems-Part C, Application and Review,* 34(2) 181-186.

Bremner, P., Pipe, A., Melhuish, C., Fraser, M. and Subramanian, S. (2009) Conversational gestures in human-robot interaction. In: *2009 IEEE international conference on systems, man and cybernetics*, San Antonio, TX, USA, 11-14 October.

Brscić, D., Kidokoro, H., Suehiro, Y. and Kanda, T. (2015) Escaping from children's abuse of social robots. In: *10th ACM/IEEE international conference on human-robot interaction,* Portland, OR, USA, March 2-5. New York, USA: ACM, 59-66. Available from http://dx.doi.org/10.1145/2696454.2696468 [accessed 14 May 2017].

Campbell, W. K. and Sedikides, C. (1999) Self-threat magnifies the self-serving bias: A meta-analytic integration. *Review of General Psychology,* 3(1) 23-43.

Campbell, W. K., Sedikides, C., Reeder, G. D. and Elliot, A. J. (2000) Among friends? An examination of friendship and the self-serving bias. *British Journal of Social Psychology*, 39(2) 229–239.

Casey, D., Felzmann, H., Pegman, G., Kouroupetroglou, C., Murphy, K., Koumpis, A. and Whelan, S. (2016) What people with dementia want: designing MARIO an acceptable robot companion. *15th International conference on computers helping people with special needs,* Linz, Austria, 11-12 July. In: Miesenberger K., Bühler C., Penaz P. (eds) *Computers helping people with special needs*. Cham: Springer, 318-325. Available from https://doi.org/10.1007/978-3-319-41264-1_44 [accessed 24 October 2018].

Castellano, G., Pereira, A., Leite, I., Paiva, A. and McOwan, P. W. (2009) Detecting user engagement with a robot companion using task and social interaction-based features. In: *ICMI-MLMI 2009,* Cambridge, MA, USA, 2-4 November. New York, USA: ACM, 119-126. Available from doi:10.1145/1647314.1647336 [accessed 14 May 2017].

Charles Darwin Quotes (n.d.) *BrainyQuote.com* [online]. Available from https://www.brainyquote.com/quotes/quotes/c/charlesdar141357.html [accessed 20 May 2017].

Chira, I., Adams, M. and Thornton B. (2008) Behavioral bias within the decision making process. *Journal of Business and Economics Research*, 6(8) 11-20.

Cialdini, R. B., Reno, R. R. and Kallgren, C. A. (1990) A focus theory of normative conduct: recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology*, 58(6) 1015-1026.

Coltfelter, C. T. and Cook. P. J. (1993) Notes: the “gambler's fallacy” in lottery play. *Management Science*, 39(12). Available from https://doi.org/10.1287/mnsc.39.12.1521 [accessed 14 May 2017].

Cornely, P. R. (2013) Receiver biases in global positioning satellite ranging. In: Mohamed, A. (ed.) *Global Navigation Satellite Systems,* London: InTech Open, 49-65. Available from doi: 10.5772/55567 [accessed 14 May 2017].

Coxall, M. (2013) *Human manipulation: a handbook*. Seville, Spain: Cornelio Books. ISBN: 9781503207677.

Crocker, J. and Major, B. (1989) Social stigma and self-esteem: the self-protective properties of stigma. *Psychological Review,* 96(4) 608-630.

Dautenhahn, K. (1998) The art of designing socially intelligent agents: science, fiction, and the human in the loop. *Applied Artificial Intelligence,* 12(7-8) 573–617. Available from doi:10.1080/088395198117550 [accessed 14 May 2017].

Dautenhahn, K. (2007) Socially intelligent robots: dimensions of human–robot interaction. *Philosophical Transactions of The Royal Society of London B: Biological Sciences,* 362(1480) 679-704.

Dautenhahn, K., Nehaniv, C. L., Walters, M. L., Robins, B., Kose-Bagci, H., Mirza, N. A. and Blow, M. (2009) KASPAR – a minimally expressive humanoid robot for human-robot interaction research. *Applied Bionics and Biomechanics*, 6(3-4) 369-397.

De Michele P. E., Gansneder B. and Solomon G. B. (1998) Success and failure attributions of wrestlers: further evidence of the self-serving bias. *Journal of Sport Behavior*, 21(3) 242-255.

Dietrich, C. (2010) Decision making: factors that influence decision making, heuristics used, and decision outcomes. *Inquiries Journal*, 2(2) 1-3.

Di Nuovo, A., Broz, F., Belpaeme, T., Cangelosi, A., Cavallo, F., Esposito, R. and Dario, P. (2015) Toward usable and acceptable robot interfaces for the elderly: the robot-era project experience. *International Psychogeriatrics* 27 S179–S179.

DiSalvo, C. and Gemperle, F. (2003) From seduction to fulfillment: the use of anthropomorphic form in design. In: *DPPI '03 Proceedings of the 2003 international conference on designing pleasurable products and interfaces*, Pittsburgh, PA, USA, 23-26 June. New York, USA: ACM, 67-72. [accessed 14 May 2017].

Dormehl, L. (2016) Luxury apartments in Los Angeles come with smart robot butlers. *Digital Trends*, 15 November. Available from https://www.digitaltrends.com/cool-tech/luxury-apartment-robot-butler/ [accessed 15 May 2017].

Duffy, B. R. (2003) Anthropomorphism and the social robot. *Robotics and autonomous systems*, 42(3) 177-190.

Dunning, D. (2005) *Self-insight: roadblocks and detours on the path to knowing thyself*. Hove, UK: Psychology Press. ISBN: 9780415654173.

Darren Duxbury, (2015) Behavioral finance: insights from experiments II: biases, moods and emotions, *Review of Behavioral Finance*, 79(2) 151-175. Available from https://doi.org/10.1108/RBF-09-2015-0037 [accessed 14 May 2017].

Dvorsky, G. (2013) How much longer before our first AI catastrophe? *Gizmodo* [online], 1 April. Available from https://io9.gizmodo.com/how-much-longer-before-our-first-ai-catastrophe-464043243 [accessed 14 May 2017].

Ehrlinger, J., Readinger, W. O. and Kim, B. (2015) Decision-making and cognitive biases. In: Friedman, H. S. (ed.) *Encyclopaedia of Mental Health*. 2nd ed. Philadelphia, PA: Elsevier. ISBN: 0123970458.

Elbot (n.d.) *Home* [online]. Available from http://www.elbot.com/ [accessed 22 May 2017].

EnrichMe (n.d.) *About* [online]. Available from enrichme.eu/wordpress/about/ [accessed 13 May 2017].

Esteve, A. (2017) The business of personal data: Google, Facebook, and privacy issues in the EU and the USA. *International Data Privacy Law*, 7(1) 36-47.

Evans, J. S. B. and Stanovich, K. E. (2013) Dual-process theories of higher cognition: advancing the debate. *Perspectives on Psychological Science*, 8(3) 223-241.

Eyssel, F., Hegel, F., Horstmann, G. and Wagner, C. (2010) Anthropomorphic inferences from emotional nonverbal cues: a case study. In: *19th international symposium in robot and human interactive communication*, Viareggio, Italy, 13-15 September. [accessed 14 May 2017].

Felzmann, H., Murphy, K., Casey, D. and Beyan, O. (2015) Robot-assisted care for elderly with dementia: is there a potential for genuine end-user empowerment? In: *Proceedings of the 10th ACM/IEEE international conference on human-robot* *interaction* [Workshop on the emerging policy and ethics of human-robot interaction], Portland, OR, USA, 2 March. Available from http://doi.org/10.13025/S8SG6Q [accessed 14 May 2017].

Feil-Seifer, D. and Matarić, M. J. (2009) Toward socially assistive robotics for augmenting interventions for children with autism spectrum disorders. In: Khatib O., Kumar V. and Pappas G.J. (eds.) *Experimental Robotics*. Berlin, Heidelberg: Springer, 201-210.

Fink, J. (2012) Anthropomorphism and human likeness in the design of robots and human-robot interaction. In: *ICSR 2012: social robotics*, Chengdu, China, 29-31 October. Berlin, Heidelberg: Springer, 199-208.

Fiske, S. T. (2010) *Social beings: core motives in social psychology*. 2nd ed. Hoboken, NJ: Wiley. ISBN: 9780470129111.

Fong, T., Nourbakhsh, I. and Dautenhahn, K. (2003) A survey of socially interactive robots. *Robotics and Autonomous Systems*, 42(3) 143-166.

Forsyth, D. R. (2008) Self-serving bias. In: Darity, W. A (ed.) *International encyclopedia of the social sciences*. 2nd ed. Vol. 7. Detroit: Macmillan Reference USA, 429.

Francis, A. and Mishra P. (2009) Is AIBO real? Understanding children's beliefs about and behavioral interactions with anthropomorphic toys. *Journal of Interactive Learning Research*, 20(4) 405-422.

Friedman, B. and Nissenbaum, H. (1996) Bias in computer systems. *ACM Transactions on Information Systems*, 14(3) 330-347.

Fussell, S. R., Kiesler, S., Setlock, L. D. and Yew, V. (2008) How people anthropomorphize robots. In: *HRI2008*, Amsterdam, Netherlands, 12-15 March. New York, USA: ACM, 145-152.

Galton, F. (1879) Psychometric experiments. *Brain*, 2(2) 149-162. Available from http://pubman.mpdl.mpg.de/pubman/item/escidoc:2323031/component/escidoc:2323030/Galton_1879_psychometric.pdf [accessed 14 May 2017].

Garland, A. (dir.) (2014) *Ex machina* [DVD]. Universal Pictures International (UPI), Film4, DNA Films.

Gates, B. (2017) I'm Bill Gates, co-chair of the Bill & Melinda Gates Foundation. Ask Me Anything. *Reddit* [online], Available from www.reddit.com/r/IAmA/comments/5whpqs/im_bill_gates_cochair_of_the_bill_melinda_gates/ [accessed 17 May 2017].

Gecko Systems (n.d.-a) *The CareBot™ - developed using data from the world's first in home trials of an eldercare mobile service robot* [online]. Available from http://www.geckosystems.com/markets/CareBot_trials.php [accessed 16 May 2017].

Gecko Systems (n.d.-b) *CareBot™ - How it works* [online]. Available from http://www.geckosystems.com/markets/CareBot_works.php [accessed 17 May 2017].

Gigerenzer, G. and Goldstein, D. G. (1996) Reasoning the fast and frugal way: models of bounded rationality. *Psychological Review,* 103(4) 650–669. Available from doi: 10.1037/0033-295X.103.4.650 [accessed 14 May 2017].

Gini, M. L., Pearce, J. and Sutherland, K. T. (2006) Using the Sony AIBOs to increase diversity in undergraduate CS programs. In: *Intelligent autonomous systems 9*, Tokyo, Japan, 7-9 March. [accessed 14 May 2017].

Gockley, R., Bruce, A., Forlizzi, J., Michalowski, M., Mundell, A., Rosenthal, S. and Wang, J. (2005) Designing robots for long-term social interaction. In: *Intelligent robots and systems*, Edmonton, AB, Canada, 2-6 August. [accessed 14 May 2017].

Goff, L. and Roediger, H. L., III. (1998) Imagination inflation for action events: repeated imaginings lead to illusory recollections. *Memory & Cognition*, 26(1) 20-33.

Goidel, R. K. and Todd, G. S. (1994) The vanishing marginals, the bandwagon, and the mass media. *The Journal of Politics*, 56(3) 802-810.

Government Office for Science. (2013) *Future of aging*. Available from https://www.gov.uk/government/collections/future-of-ageing#project-report [accessed 26 May 2017].

Gutnik, L. A., Hakimzada, A. F., Yoskowitz, N. A. and Patel, V. L. (2006) The role of emotion in decision-making: a cognitive neuroeconomic approach towards understanding sexual risk behaviour. *Journal of Biomedical Informatics*, 39(6) 720–736.

Hafner, K. (2000) A robot that coos, cries and knows when it needs a new diaper. *New York Times* [online]. 16 November. Available from http://www.nytimes.com/2000/11/16/technology/a-robot-that-coos-cries-and-knows-when-it-needs-a-new-diaper.html [accessed 14 May 2017].

Hahn, U. and Harris J. L. A. (2014) What does it mean to be biased: motivated reasoning and rationality. *Psychology of Learning and Motivation*, 61 41–102.

Hampes, W. P. (1992) Relation between intimacy and humour. *Psychological Reports*, 71(1) 127-130. Available from doi: http://dx.doi.org/10.2466/pr0.1992.71.1.127 [accessed 14 May 2017].

Hampes, W. P. (1999) The relationship between humor and trust. *International Journal of Humor Research*, 12(3) 253-260. Available from doi.org/10.1515/humr.1999.12.3.253 [accessed 14 May 2017].

Haring S.K., Mougenot, K., Ono, F. and Watanabe, K. (2014) Cultural differences in perception and attitude towards robots. *International Journal of Affective Engineering*, 3 149-157. Available from doi:10.5057/ijae.13.149 [accessed 14 May 2017].

Harris, C. B., Paterson, H. M. and Kemp, R. I. (2008) Collaborative recall and collective memory: what happens when we remember together? *Memory,* 16(3) 213-230. Available from doi:10.1080/09658210701811862 [accessed 14 May 2017].

Haselton, M. G., Nettle, D. and Andrews, P. W. (2005) The evolution of cognitive bias. In: D. M. Buss (ed.) *The handbook of evolutionary psychology*. Hoboken, NJ, US, John Wiley & Sons Inc, 724–746.

Hassenzahl, M. (2003) The thing and I: understanding the relationship between user and product. In: M.Blythe, C. Overbeeke, A. F. Monk and P. C. Wright (eds.) *Funology: From usability to enjoyment*. Dordrecht, Netherlands: Kluwer Academic Publishers, 31-42.

Hayashi, K., Sakamoto, D., Kanda, T., Shiomi, M., Koizumi, S., Ishiguro, H. and Hagita, N. (2007) Humanoid robots as a passive-social medium – a field experiment at a train station. In: *Human-robot interactions in social robotics*, Arlington, VA, USA, 9-11 March. [accessed 14 May 2017].

Heider, F. and Simmel, M. (1944) An experimental study of apparent behaviour. *American Journal of Psychology*, 57(2) 243-259. Available from http://dx.doi.org/10.2307/1416950 [accessed 14 May 2017].

Henkel, L. A. and Coffman, K. J. (2004) Memory distortions in coerced false confessions: a source monitoring framework analysis. *Applied Cognitive Psychology,* 18(5) 567-588. Available from https://doi.org/10.1002/acp.1026 [accessed 14 May 2017].

Hilbert, M. (2012) Toward a synthesis of cognitive biases: how noisy information processing can bias human decision making. *Psychological Bulletin*, 138(2) 211-237. Available from doi:10.1037/a0025940 [accessed 14 May 2017].

Hiolle, A., Cañamero, L., Ross, M. D. and Bard, K. A. (2012) Eliciting caregiving behavior in dyadic human-robot attachment-like interactions. *ACM Transactions on Interactive Intelligent Systems,* 2(1) 3:1-3:24. Available from doi: 10.1145/2133366.2133369 [accessed 14 May 2017].

HOBBIT (n.d.) *HOBBIT – the mutual care robot*. Available from http://hobbit.acin.tuwien.ac.at/ [accessed 13 May 2017].

Hone, K. and Graham, R. (2000) Towards tool for the subjective assessment of speech system interfaces (SASSI) *Natural Language Engineering,* 6(3-4) 287–303.

Hoorens, V. (1993) Self-enhancement and superiority biases in social comparison. *European Review of Social Psychology*, 4(1) 113–139. Available from doi:10.1080/14792779343000040 [accessed 14 May 2017].

How Apple's Siri got her name (2012). *The Week* [online], 29 March 2012. Available from http://theweek.com/articles/476851/how-apples-siri-got-name [accessed 19 May 2017].

InMoov (n.d.) *Gaël Langevin – French InMoov designer* [online]. Available from http://inmoov.fr/ [accessed 22 May 2017].

Ippoliti, E. (2015) *Heuristic reasoning: studies in applied philosophy, epistemology and rational ethics.* Cham, Switzerland: Springer International Publishing. ISBN: 9783319091594.

iRobot (n.d.) *Home robots* [online]. Available from http://www.irobot.com/For-the-Home/Vacuuming/Roomba.aspx [accessed 15 May 2017].

James, W. (1984) *Psychology: the briefer course*. 16th ed. Cambridge, MA: Harvard University Press. ISBN: 0674721020.

James, P. (2015) *Urban sustainability in theory and practice: circles of sustainability*. Abingdon: Routledge. ISBN: 1138025739.

Janka, P. (2008) Using a programmable toy at preschool age: why and how? In: *Simulation, modelling and programming for autonomous robots*, Venice, Italy, 3-4 November. [accessed 14 May 2017].

Jasheway, A. L. (2016) How to write better using humor. *Writer's Digest* [online]. 26 January. Available from http://www.writersdigest.com/online-editor/how-to-mix-humor-into-your-writing [accessed 14 May 2017].

Johnson, M. K., Hashtroudi. S. and Lindsay, S. (1993) Source monitoring. *Psychology Bulletin*, 114(1) 3–28.

Jones, B. and Kenward M. G. (2014) *Design and analysis of cross-over trials*. 3rd ed. New York: Chapman & Hall/CRC. ISBN: 1439861420.

Jøranson, N., Pedersen I., Rokstad, M. M. A. and Ihlebæk, C. (2015) Effects on symptoms of agitation and depression in persons with dementia participating in robot-assisted activity: a cluster-randomized controlled trial. *The Journal of Post-Acute and Long-Term Care Medicine*, 16(10) 867–873. Available from http://dx.doi.org/10.1016/j.jamda.2015.05.002 [accessed 14 May 2017].

Judea, P. (1983) *Heuristics: intelligent search strategies for computer problem solving.* New York: Addison-Wesley. ISBN: 9780201055948.

Kahneman, D. (2003) Maps of bounded rationality: psychology for behavioral economics [revised version of 2002 Nobel Prize lecture]. *The American Economic Review*, 93(5) 1449-1475.

Kanwisher, N. (2010) Functional specificity in the human brain: a window into the functional architecture of the mind. *Proceedings of the National Academy of Sciences,* 107(25) 11163–11170. Available from doi: 10.1073/pnas.1005062107 [accessed 14 May 2017].

Kassin, S. (2003) *Psychology*. Upper Saddle River, NJ: Prentice-Hall. ISBN: 0130496413.

Kercheval, M. (2015) The robots are coming...to your mall. *CNBC* [online], 17 July. Available from http://www.cnbc.com/2015/07/17/the-robots-are-coming-to-your-mall-commentary.html [accessed 7 May 2017].

Kim, Y. and Mutlu, B. (2014) How social distance shapes human–robot interaction. *International Journal of Human-Computer Studies*, 72(12) 783–795. Available from http://dx.doi.org/10.1016/j.ijhcs.2014.05.005 [accessed 14 May 2017].

Kismet (n.d.) *Kismet robot overview* [online]. Available from http://www.ai.mit.edu/projects/sociable/overview.html [accessed 13 May 2017].

Knapton, S. (2015) Meet Nadine, the world's most human-like robot. *The Telegraph* [online], 29 December. Available from http://www.telegraph.co.uk/science/2016/03/12/meet-nadine-the-worlds-most-human-like-robot/ [accessed 15 May 2017].

KnightScope (n.d.) *Demo our machines* [online], Available from http://www.knightscope.com/demo [accessed 7 May 2017].

Kok, L. (2015) NTU scientists unveil social and telepresence robots. *Media release: Nanyang Technological University* [online], 29 December 2015. Available from http://media.ntu.edu.sg/NewsReleases/Pages/newsdetail.aspx?news=fde9bfb6-ee3f-45f0-8c7b-f08bc1a9a179 [accessed 7 May 2017].

Kozima, H., Michalowski, M. P. and Nakagawa, C. (2009) Keepon: a playful robot for research, therapy, and entertainment. *International Journal of Social Robotics*, 1(1) 3-18.

Kruger, J. (1999) Lake Wobegon be gone! The "below-average effect" and the egocentric nature of comparative ability judgments. *Journal of Personality and Social Psychology*, 77(2) 221–232. Available from doi:10.1037/0022-3514.77.2.221 [accessed 14 May 2017].

Kruger, J. and Dunning, D. (1999) Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, 77(6) 1121-1134.

Kuiper, N. A., Martin, R. A. and Olinger, L. J. (1993) Coping humour, stress, and cognitive Appraisals. *Canadian Journal of Behavioural Science*, 25(1) 81-96.

Kuiper, N. A., Mckenzie, S. D. and Belanger, K. A. (1995) Cognitive appraisals and individual differences in sense of humor: motivational and affective implications. *Personality & Individual Differences*, 19(3) 359-372.

Kwon M., Jung M. F. and Knepper R. A. (2016) Human expectations of social robots. In: *11th ACM/IEEE human robot interaction (HRI)*, Christchurch, New Zealand, 7-10 March. Available from doi: 10.1109/HRI.2016.7451807 [accessed 14 May 2017].

Lauer, R. H., Lauer, J. C. and Kerr, S. T. (1990) The long-term marriage: perceptions of stability and satisfaction. *International Journal of Aging & Human Development*, 31(3) 189-195. Available from doi: 10.2190/H4X7-9DVX-W2N1-D3BF [accessed 14 May 2017].

Lee, C. (2012) Revisiting why incompetents think they're awesome. *Ars Technica* [online], 4 November. Available from https://arstechnica.com/science/2016/11/revisiting-why-incompetents-think-theyre-awesome [accessed 14 May 2017].

Lee, K. M., Peng, W., Jin, S. A. and Yan, C. (2006) Can robots manifest personality? an empirical test of personality recognition, social responses, and social presence in human–robot interaction. *Journal of Communication*, 56(4) 754-772.

Lee, S. L., Lau, I. Y. M., Kiesler, S. and Chiu, C. Y. (2005) Human mental models of humanoid robots. In: *IEEE international conference on robotics and automation*, Barcelona, Spain, 18-22 April. Piscataway, NJ, USA: IEEE, 2767-2772. [accessed 14 May 2017].

Leite, I., Castellano, G., Pereira, A., Martinho, C. and Paiva, A. (2012) Modelling empathic behaviour in a robotic game companion for children: an ethnographic study in real-world settings. In: *Proceedings of the seventh annual ACM/IEEE international* *conference on human-robot interaction*, Boston, MA, USA, 5-8 March. New York, NJ, USA: ACM, 367–374.

Leite, I., McCoy, M., Lohani, M., Salomons, N., McElvaine, K., Stokes, C. and Scassellati, B. (2016) Autonomous disengagement classification and repair in multiparty child-robot interaction. In: *25th IEEE international symposium on robot and human interactive communication (Ro-Man)*, 26-31 August. Piscataway, NJ, USA: IEEE, 525-532.

Levine, J. and Zigler, E. (1976) Humor responses of high and low premorbid competence alcoholic and nonalcoholic patients. *Addictive Behaviors,* 1(2) 139-149.

Litter Robot (n.d.) *Automatic* Self-cleaning litter box for cats [online]. Available from https://www.litter-robot.com/ [accessed 15 May 2017].

Loewenstein, G. (1996) Out of control: visceral influences on behavior. *Organizational Behavior and Human Decision Processes*, 65(3) 272-292.

Loewenstein, G. (2005) Hot-cold empathy gaps and medical decision making. *Health Psychology*, 24(4) 49–56. Available from doi: 10.1037/0278-6133.24.4.S49 [accessed 14 May 2017].

Loewenstein, G. F., Weber, E. U., Hsee, C. K. and Welch, N. (2001) Risk as feelings. *Psychological Bulletin*, 127(2) 267-286. Available from doi:10.1037/0033-2909.127.2.267 [accessed 14 May 2017].

Markoff, J. (2013) A night watchman with wheels? *New York Times* [online], 29 November. Available from http://www.nytimes.com/2013/12/03/science/coming-soon-a-night-watchman-with-wheels.html [accessed 7 May 2017].

Marois, R. and Ivanoff, J. (2005) Capacity limits of information processing in the brain. *Trends in Cognitive Sciences*, 9(6) 296-305.

Martin, R. A., Puhlik-Doris, P., Larsen, G., Gray, J. and Weir, K. (2003) Individual differences in uses of humor and their relation to psychological well-being: development of the Humor Styles Questionnaire. *Journal of Research in Personality*, 37(1) 48-75.

Matthews, G., Deary, I. J. and Whiteman, M. C. (2003) *Personality traits*. 2nd ed. Cambridge, UK: Cambridge University Press. ISBN: 0521831075.

McHugh, P. R. (2008) *Try to remember: psychiatry's clash over meaning, memory and mind*. New York, NY: Dana Press. ISBN: 1932594396.

McKay T. R. and Dennett, C. D. (2009) The evolution of misbelief. *Behavioural and Brain Sciences*, 32(6) 493 –561. Available from doi:10.1017/S0140525X09990975 [accessed 14 May 2017].

McLeod, S. A. (2008) Social roles. *Simply Psychology* [online]. Available from https://www.simplypsychology.org/social-roles.html [accessed 14 May 2017].

Meerbeek, B., Hoonhout, J., Bingley, P. and Terken, M. B. J. (2006) Investigating the relationship between the personality of a robotic TV assistant and the level of user control. In: *Robot and human interactive communication. RO-MAN 2006*, Hatfield, Herthfordshire, UK, 6-8 September. Piscataway, NJ: IEEE, 404-410.

Meerbeek, B., Saerbeck, M. and Bartneck, C. (2009) Iterative design process for robots with personality. In: *AISB2009symposium on new frontiers in human-robot interaction*, Edinburgh, UK, 6-9 April. [accessed14 May 2017].

Melson, G. F., Kahn, P. H., Jr., Beck, A., Friedman, B., Roberts, T., Garrett, E. and Gill, B. T. (2009) Children's behavior toward and understanding of robotic and living dogs. *Journal of Applied Developmental Psychology*, 30(2) 92-102.

Michalowski, MP., Simmons, R., and Kozima, H. (2009) Rhythmic attention in child-robot dance play. In: *The 18th IEEE international symposium on robot and human interactive communication, RO-MAN 2009*, Toyama, Japan, 27 September - 2 October. Piscataway, NJ: IEEE, 816-821. Available from doi: 10.1109/ROMAN.2009.5326143 [accessed 14 May 2017].

Miller, D. T. and Ross, M. (1975) Self-serving biases in the attribution of causality: fact or fiction? *Psychological Bulletin*, 82(2) 213-225.

Moley (n.d) *Future is served, Moley Robotics* [online], Available from http://www.moley.com/ [accessed 7 November 2017].

Morris, E. (2010) The anosognosic's dilemma: something's wrong but you'll never know what it is. *New York Times* [online], 20 June. Available from https://opinionator.blogs.nytimes.com/2010/06/20/the-anosognosics-dilemma-1/ [accessed 7 March 2011].

Moshkina, L. and Arkin, R. C. (2009) Beyond humanoid emotions: incorporating traits, attitudes and moods. In: *Proc. 2009 IEEE Workshop on Current Challenges and Future Perspectives of Emotional Humanoid Robotics*, Kobe, Japan [accessed14 May 2017].

Moshkina, L., Park, S., Arkin, R. C., Lee, J. K. and Jung, H. (2011) TAME: Time-varying affective response for humanoid robots. *International Journal of Social Robotics*, 3(3) 207-221.

Mubin, O., Stevens, C. J., Shahid, S., Al Mahmud, A. and Dong, J. J. (2013) A review of the applicability of robots in education. *Journal of Technology for Education and Learning*, 1 1-7.

Mumm J. and Multu B. (2011) Human-robot proxemics: physical and psychological distancing in human-robot interaction. In: *6th ACM/IEEE international conference on human-robot interaction (HRI*). New York, USA: ACM, 331-338. Available from doi: 10.1145/1957656.1957786 [accessed14 May 2017].

Mutlu, B., Kanda, T., Forlizzi, J., Hodgins, J. and Ishiguro, H. (2012) Conversational gaze mechanisms for human like robots. *ACM Transactions on Interactive Intelligent Systems*, 1(2) 12. Available from: DOI: 10.1145/2070719.2070725 [accessed14 May 2017].

Mutthaya, B. C. (1987) Relationship between humor and interpersonal orientations. *Journal of Psychological Research*, 31 48-54.

Nalin, M., Baroni, I. and Sanna A. (2012) A motivational robot companion for children in therapeutic setting. In: *Workshop on motivational aspects of robotics in physical therapy, IEEE/RSJ international conference on intelligent robots and systems,* Vilamoura, Algarve, Portugal, 12 October.

Nanyang Technological University (n.d.) NTU scientists unveil social and telepresence robots. *Media Release: Nanyang Technological University* [online], 29 December. Available from http://media.ntu.edu.sg/NewsReleases/Pages/newsdetail.aspx?news=fde9bfb6-ee3f-45f0-8c7b-f08bc1a9a179 [accessed 15 May 2017].

Nelly Benhayoun Studios (2014) *Imperfect robots* [online]. Available from http://nellyben.com/archive/the-test-rig-for-the-soyuz-chair [accessed 19 May 2017].

Nelson, J. H. (2013) *The effects of intervention involving a robot on compliance of four children with autism to requests produced by their mothers.* MS thesis. Brigham Young University. Available from http://scholarsarchive.byu.edu/cgi/viewcontent.cgi?article=4592&context=etd [accessed 14 May 2017].

Norman, D. A. (2004) Emotional design: why we love (or hate) everyday things. New York: Basic Books. ISBN: 8601404701894.

Nordgren, L. F., Banas, K. and MacDonald, G. (2010) Empathy gaps for social pain: why people underestimate the pain of social suffering. *Journal of Personality and Social Psychology*, 100(1) 120-128.

Nordgren, L. F., Van der Pligt, J. and Van Harreveld, F. (2006) Visceral drives in retrospect: explanations about the inaccessible past. *Psychological Science*, 17(7) 635-640.

Nordgren, L. F., Van der Pligt, J. and Van Harreveld, F. (2009) The restraint bias: how the illusion of self-restraint promotes impulsive behaviour. *Psychological Science*, 20(12) 1532-1528.

Novella, S. (2014) Lessons from Dunning-Kruger. *Neurologica* [online], 6 November. Available from http://theness.com/neurologicablog/index.php/lessons-from-dunning-kruger/ [accessed 20 May 2017].

Otsuka, T., Nakadai, K., Takahashi, T., Ogata, T. and Okuno, H. G. (2011) Real-time audio-to-score alignment using particle filter for coplayer music robots. *EURASIP Journal on Advances in Signal Processing*, 1-13. [accessed 14 May 2017].

PARO (n.d.) *PARO therapeutic robot*[online]. Available from http://www.parorobots.com [accessed 14 May 2017].

Payne, B. K., Cheng, C. M., Govorun, O. and Stewart, B. D. (2005) An inkblot for attitudes: affect misattribution as implicit measurement. *Journal of Personality and Social Psychology*, 89(3) 277–293. Available from doi:10.1037/0022-3514.89.3.277 [accessed 14 May 2017].

Pereita, L., Tenera, A. and Wemans, B. (2013) Insights on individual's risk perception for risk assessment in web-based risk management tools. *Procedia Technology*, 9 886-892.

Pfister, H. R. and Böhm, G. (2008) The multiplicity of emotions: a framework of emotional functions in decision making. *Judgement and Decision Making*, 3(1) 5-17.

Piehlmaier, D. (2013) *Irrational and overrated: is our unrealistic self-perception connected to educational achievements?* Hamburg, Germany: Anchor Academic Publishing. ISBN: 3954890755. [Original title of thesis: *Overconfidence – a matter of education?* (2012) BA. Fachhochschule Kufstein Tirol, Tirol, Austria].

Pohl, R. F. (2007) Ways to assess hindsight bias. *Social Cognition*, 25(1) 14-31. Available from doi: 10.1521/soco.2007.25.1.14 [accessed 14 May 2017].

Pripfl, J., Körtner, T., Batko-Klein, D., Hebesberger, D., Weninger, M., Gisinger, C., Frennert, S., Eftring, H., Antona, M., Adami, I., Weiss, A., Bajones, M. and Vincze, M. (2016) Results of a real world trial with a mobile social service robot for older adults. In: 11th ACM/IEEE international conference on human-robot interaction (HRI), Christchurch, New Zealand, 7-10 March. 497-498. Available from doi: 10.1109/HRI.2016.7451824 [accessed 14 May 2017].

Reeves, B. and Nass, C. (1996) *The media equation: how people treat computers, television, and new media like real people and places.* Cambridge, UK: Cambridge University Press. ISBN: 157586052X.

Richardson, J. (1999) Norms put the ‘Golden Rule’ into practice for groups. *Tools for Learning Schools,* 3(1) 1-2. [accessed 14 May 2017].

Ricks, D. J. and Colton, M. B. (2010) Trends and considerations in robot-assisted autism therapy. In: *IEEE international conference on robotics and automation (ICRA),* Anchorage, Alaska, 3-7 May. Piscataway, NJ, USA: IEEE, 4354-4359.

Robinette, P., Li, W., Allen, R., Howard, A. M. and Wagner, A. R. (2016) Over trust of robots in emergency evacuation scenarios. In: *The 11th ACM/IEEE international conference on human robot interaction,* Christchurch, New Zealand, 7-10 March Piscataway, NJ, USA: IEEE, 101-108. [accessed 14 May 2017].

Robinson, H., MacDonald, B., Kerse, N. and Broadbent, E. (2013) The psychosocial effects of a companion robot: a randomized controlled trial. *Journal of American Medical Directors Association,* 14(9) 661-667. Available from doi: 10.1016/j.jamda.2013.02.007 [accessed 14 May 2017].

Roediger, H. L. and McDermott, K. B. (1995) Creating false memories: remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory and Cognition,* 24(4) 803–814.

Rotter, J. B. (1966) Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs,* 80(1) 1-28.

Rushe, D. (2011) Siri's abortion bias embarrasses Apple as it rues 'unintentional omissions'. *The Guardian* [online], 1 December. Available from https://www.theguardian.com/technology/2011/dec/01/siri-abortion-apple-unintenional-omissions [accessed 14 May 2017].

Russell, G. and Shoare, L. (1994) *Interactive perceptual psychology: The human psychology that mirrors the naturalness of human behaviour* [Online]. In: Mid-western Educational Research Association, Annual Conference. [accessed 14 May 2017].

Rust, J. and Goldstein, J. (1989) Humor in marital adjustment. *International Journal of Humor Research*, 2(3) 217-223.

Rzeszutek, M. (2015) Personality traits and susceptibility to behavioral biases among a sample of Polish stock market investors. *International Journal of Management and Economics*, 47(1) 71-81.

Saenz, A. (2010) We live in a jungle of artificial intelligence that will spawn sentience. *SingularityHub* [online], 10 August. Available from https://singularityhub.com/2010/08/10/we-live-in-a-jungle-of-artificial-intelligence-that-will-spawn-sentience/#sm.0000zqgfrl1n5e9wuaq1s3ohwsj34 [accessed 14 May 2017].

Salem, M., Lakatos, G., Amirabdollahian, F. and Dautenhahn, K. (2015) Would you trust a (faulty) robot?: effects of error, task type and personality on human-robot cooperation and trust. In: *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction*, Portland, OR, USA, 2-5 March. New York, USA: ACM, 141-148. [accessed 14 May 2017].

Salovey, P. and Mayer, J. D. (1990) Emotional intelligence. *Imagination, Cognition, and Personality*, 9(3) 185-211.

Salovey, P., Mayer, J. D. and Caruso, D. (2002) The positive psychology of emotional intelligence. In: C. R. Snyder and S. J. Lopez (eds.), *Handbook of positive psychology*. Oxford, UK: Oxford University Press, 45-61.

Saraglou, V. and Scariot, C. (2002) Humor styles questionnaire: personality and educational correlates in Belgian high school and college students. *European Journal of Personality*, 16(1) 43-54.

Sardon, E., Rius, A. and Zarraoa, N. (1994) Estimation of the transmitter and receiver differential biases and the ionospheric total electron content from Global Positioning System observations. *Radio Science*, 29(3) 577-586.

Sayette, M. A., Loewenstein, G., Griffin, K. M. and Black, J. (2008) Exploring the cold-to-hot empathy gap in smokers. *Psychological Science,* 19(9) 926-932.

Schacter, D. L. (1999) The seven sins of memory: insights from psychology and cognitive neuroscience. *American Psychologist*, 54(3) 182-203.

Schacter, D. L. and Dodson, C. S. (2001) Misattribution, false recognition and the sins of memory. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 356(1413) 1385–1393. Available from doi:10.1098/rstb.2001.0938 [accessed 14 May 2017].

Schacter, D. L., Guerin, S. A. and Jacques, P. (2011) Memory distortion: an adaptive perspective. *Trends in Cognitive Science*, 15(10) 467–474. Available from doi: 10.1016/j.tics.2011.08.004. [accessed 14 May 2017].

Scheeff, M., Pinto, J., Rahardja, K., Snibbe, S. and Tow, R. (2002) Experiences with Sparky, a social robot. In: Dautenhahn, K., Bond, A., Cañamero, L. and Edmonds, B. (eds.), *Socially intelligent agents: creating relationships with computers and robots*. Boston, MA: Kluwer Academic Publishers, 173–180.

Schmidt, S. R. (2002) The humour effect: differential processing and privileged retrieval. *Memory*, 10(2) 127–138. Available from doi: 10.1080/09658210143000263. [accessed 14 May 2017].

Searle, J. (1980) Minds, brains, and programs. *The Behavioral and Brain Sciences*, 3(3) 417-457.

Sedikides, C., Campbell, W. K., Reeder, G. D. and Elliot, A. J. (1998) The self-serving bias in relational context. *Journal of Personality and Social Psychology*, 74(2) 378–386.

Seymour, R. P. (n.d.) *Cognitive Biases: memory errors and biases* [memrise]. Available from https://www.memrise.com/course/1123902/cognitive-biases/3/ [accessed 20 May 2017].

Shepperd, J., Malone, W. and Sweeny, K. (2008) Exploring causes of the self-serving bias. *Social and Personality Psychology Compass*, 2(2) 895-908.

Shibata, T. and Wada, K. (2011) Robot therapy: a new approach for mental healthcare of the elderly – a mini-review. *Gerontology*, 57(4) 378–386. Available from doi: 10.1159/000319015 [accessed 14 May 2017].

Siegel, M., Breazeal, C., and Norton, M. I. (2009) Persuasive robotics: the influence of robot gender on human behaviour. In: *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, St. Louis, MO, USA, 10-15 October. 2563-2568. [accessed 14 May 2017].

Simon, H. A. (1955) A behavioural model of rational choice. *The Quarterly Journal of Economics*, 69(1) 99-118. Available from doi:10.2307/1884852. [accessed 14 May 2017].

Simon, M., Houghton, S. M. and Aquino, K. (2000) Cognitive biases, risk perception, and venture formation: how individuals decide to start companies. *Journal of Business Venturing*, 15(2) 113-134.

Smith, P. (1974) A trekkie's tale. *The Menagerie* [online]. Available from http://web.archive.org/web/20100730113417/http://www.fortunecity.com/rivendell/dark/1000/marysue.htm [accessed 19 May 2017].

Sony (n.d.) *AIBO models*[online]. Available from http://www.sony-aibo.com/aibo-models/ [accessed 14 May 2017].

Sparrow, R. (2002) The march of the robot dogs. *Ethics and Information Technology*, 4(4) 305-318. Available from https://doi.org/10.1023/A:1021386708994 [accessed ??].

Spencer-Oatey, H. (2012) *What is culture? A compilation of quotations. GlobalPAD Core Concepts. GlobalPAD Open House* [online]. Available from https://warwick.ac.uk/fac/soc/al/globalpad/openhouse/interculturalskills_old/global_pad_-_what_is_culture.pdf [accessed 14 May 2017].

Spielberg S. (dir.) (2001). *A.I. artificial intelligence* [DVD]. Warner Bros. DreamWorks, Amblin Entertainment.

Standing, L. G. (2004). Halo Effect. In: M. S. Lewis-Black, A. Bryman, and T. F. Liao (eds.) *The SAGE encyclopedia of social science research methods*, Vol. 1. Thousand Oaks, CA: SAGE Publications, 451-2. Available from http://dx.doi.org/10.4135/9781412950589 [accessed 14 May 2017].

Stanovich, K. E. and West, R. F. (1998) Individual differences in rational thought. *Journal of Experimental Psychology: General*, 127(2) 161-188. Available from doi: http://dx.doi.org/10.1037/0096-3445.127.2.161 [accessed 14 May 2017].

Star Wars Wikia. (n.d.) C-3PO [online]. Available from http://starwars.wikia.com/wiki/C-3PO; C-3PO [accessed 15 May 2017].

Stewart, A. E. (2005) Attributions of responsibility for motor vehicle crashes. *Accident Analysis & Prevention*, 37(4) 681–688.

Strick, M., Holland, R. W., Van Baaren, R. B. and Van Knippenberg, A. D. (2009) Finding comfort in a joke: consolatory effects of humor through cognitive distraction. *Emotion*, 9(4) 574–578. Available from doi: 10.1037/a0015951 [accessed 14 May 2017].

Subash, S., Rocco, M. Di., Pecora, F. and Saffiotti, A. (2013) Scaling up ubiquitous robotic systems from home to town (and beyond). In: *Proceedings of the 2013 ACM conference on Pervasive and ubiquitous computing*, Zurich, Switzerland, 8-12 September. New York, NY: ACM, 107-110.

Sullivan, A. and Bers, M. (2016) Robotics in the early childhood classroom: learning outcomes from an 8-week robotics curriculum in pre-kindergarten through second grade. *International Journal of Technology and Design Education*, 26(1) 3-20. Available from doi: 10.1007/s10798-015-9304-5 [accessed 14 May 2017].

Super Kitchen Machine (n.d.) *What is "Bimby"?* [blog]. Available from http://www.superkitchenmachine.com/new-thermomix-tm5/bimby [accessed 15 May 2017].

Syrdal, D. S., Lehmann, H., Robins, B. and Dautenhahn, K. (2014) *KASPAR in the wild - Initial findings from a pilot study*. In: *50th annual convention of the AISB*, London, UK, 1-4 April. [accessed 14 May 2017].

Titan the Robot (n.d.) *Titan the robot: About*. [online]. Available from http://www.titantherobot.com/ [accessed 22 May 2017].

Tiwari, R. S., Arora, M. K. and Kumar, A. (2010) An appraisal of GPS related errors. *Geospatial World* [online]. 4 December. Available from https://www.geospatialworld.net/article/an-appraisal-of-gps-related-errors/ [accessed 22 May 2017].

Turkle, S. (2006) A nascent robotics culture: new complicities for companionship. *AAAI Technical Report Series*. Available from https://www.student.cs.uwaterloo.ca/~cs492/papers/ST_Nascent%20Robotics%20Culture.pdf [accessed 14 May 2017].

Turkle, S., Taggart, W., Kidd, C. D. and Dasté, O. (2006) Relational artifacts with children and elders: the complexities of cybercompanionship. *Connection Science*, 18(4) 347-361.

Tversky, A. and Kahneman, D. (1974) Judgement under uncertainty: heuristics and biases. *Sciences*, 185(4157) 1124–1131. Available from doi:10.1126/science.185.4157.1124 [accessed 14 May 2017].

Tversky, A. and Kahneman, D. (1981) The framing of decisions and the psychology of choice. *Science*, 211(4481) 453-458. Available from http://dx.doi.org/10.1126/science.7455683 [accessed 14 May 2017].

Tversky, A. and Kahneman, D. (1986) Rational choice and the framing of decisions. *The Journal of Business*, 59(4) 251-278.

Tversky, A. and Kahneman, D. (1992) Advances in prospect theory: cumulative representation of uncertainty. *Journal of Risk and Uncertainty,* 5(4) 297–323.

Valadão, C. T., Goulart, C., Rivera, H., Caldeira, E., Bastos Filho, T. F., Frizera-Neto, A. and Carelli, R. (2016) Analysis of the use of a robot to improve social skills in children with autism spectrum disorder. *Research on Biomedical Engineering*, 32(2) 161-175.

Van Boven, L., Loewenstein, G., Dunning, D. and Nordgren, L. F. (2013) Changing places: a dual judgment model of empathy gaps in emotional perspective taking. *Advances in Experimental Social Psychology,* 48 117–171. Available from doi:10.1016/B978-0-12-407188-9.00003-X [accessed 14 May 2017].

Verba, J. M. (2003) *Boldly writing: a trekker fan & zine history.* Minnetonka, MN: FTL Publications. ISBN: 0965357503.

Wainer. J., Robins. B., Amirabdollahian. F., and Dautenhahn. K. (2014) Using the humanoid robot KASPAR to autonomously play triadic games and facilitate collaborative play among children with autism. *IEEE Transactions on Autonomous Mental Development*, 6(3) 183-199. Available from doi: 10.1109/TAMD.2014.2303116 [accessed 14 May 2017].

Walters, M. L., Syrdal, D. S., Dautenhahn, K., Te Boekhorst, R. and Koay, K. L. (2008) Avoiding the uncanny valley: robot appearance, personality and consistency of behavior in an attention-seeking home scenario for a robot companion. *Autonomous Robots*, 24(2) 159-178.

Wang, X. T., Simons, F. and Brédart, S. (2001) Social cues and verbal framing in risky choice. *Journal of Behavioral Decision Making*, 14(1) 1–15. Available from doi:10.1002/1099-0771(200101)14:1<1::AID-BDM361>3.0.CO;2-N [accessed 14 May 2017].

Weijters, B., Cabooter, E. and Schillewaert, N. (2010) The effect of rating scale format on response styles: the number of response categories and response category labels. *International Journal of Research in Marketing*, 27(3) 236–247. Available from doi: http://doi.org/10.1016/j.ijresmar.2010.02.004 [accessed 14 May 2017].

Weinberger, M. G. and Gulas, C. S. (1992) The impact of humor in advertising: a review. *Journal of Advertising*, 21(4) 35-59.

Weizenbaum, J. (1976) *Computer power and human reason: from judgment to calculation.* Oxford, UK: W. H. Freeman & Co.

White, G., Fishbein, S. and Rutsein, J. (1981) Passionate love and the misattribution of arousal. *Journal of Personality and Social Psychology*, 41(1) 56–62.

Wilke, A. and Mata, R. (2012) Cognitive bias. In: V.S. Ramachandran (ed.) *The encyclopedia of human behavior.* Cambridge, MA: Academic Press, 531-535.

Wilson, D. (2011) Questions raised over Apple Siri bias. *The Inquirer* [online], 1 December. Available from https://www.theinquirer.net/inquirer/news/2129263/questions-raised-apple-siri-bias [accessed 19 May 2017].

Wykowska, A., Kajopoulos, J., Obando-Leitón, M., Chauhan, S. S., Cabibihan, J. J. and Cheng, G. (2015) Humans are well tuned to detecting agents among non-agents: examining the sensitivity of human perception to behavioral characteristics of intentional systems. *International Journal of Social Robotics*, 7(5) 767-781.

Yannakakis, G. N., Paiva, A. (2014) *Emotion in games.* In: D'Mello, S., Graesser, A., Schuller, B., Martin, J. C. (eds), *Affective computing and intelligent interaction.* Berlin, Heidelberg: Springer, 497. [accessed 14 May 2017].

Yechiam, E., Druyan, M. and Ert, E. (2008) Observing others' behavior and risk taking in decisions from experience. *Judgment and Decision Making*, 3(7) 493-500.

Yudkowsky, E. (2008) Cognitive biases potentially affecting judgment of global risks. In: Bostrom, N. and Ćirković, M. M. (eds.), *Global catastrophic risks*. New York, NY: Oxford University Press, 91-119.

Ziv, A. (1988) Humor's role in married life. *Humor: International Journal of Humor Research*, 1(3) 223-229.

Ziv, A, & Gadish, O. (1989) Humor and marital satisfaction. *Journal of Social Psychology*, 129(6) 759-768.

Zlotowski, J., Proudfoot, D., Yogeeswaran, K. and Bartneck, C. (2015) Anthropomorphism: opportunities and challenges in human-robot interaction. *International Journal of Social Robotics*, 7(3) 347-360.

Appendix A: List of publications

Paper's names	Year	Conference/Journal and links
Journal		
The Effects of Cognitive Biases and Imperfectness in Long-term Robot-Human Interactions: Case Studies using Five Cognitive Biases on Three Robots	2016	Cognitive Systems Research Journal, can be found at: https://www.sciencedirect.com/science/article/pii/S1389041716300614
Can Cognitive Biases in Robots Make More 'Likeable' Human-Robot Interactions than the Robots Without Such Biases: Case Studies Using Five Biases on Humanoid Robot	2016	International Journal of Artificial Life Research (IJALR) Can be found: https://www.igi-global.com/article/can-cognitive-biases-in-robots-make-more-likeable-human-robot-interactions-than-the-robots-without-such-biases/177182
Papers		
Effect of cognitive biases on human-robot interaction: a case study of a robot's misattribution.	2014	RO-MAN'14 Can be found at: http://ieeexplore.ieee.org/abstract/document/6926387/
Robotic Companionship: How Forgetfulness Affects Long-Term Human-Robot Interaction;	2014	ICIRA'14 Can be found at: https://link.springer.com/chapter/10.1007/978-3-319-22876-1_4
Towards an imperfect robot for long-term companionship: case studies using cognitive biases	2015	IROS'15 Can be found at: http://ieeexplore.ieee.org/document/7354228/
Robots that Refuse to Admit Losing – A Case Study in Game Playing Using Self-Serving Bias in the Humanoid Robot MARC	2016	ICIRA'16 Can be found at: https://link.springer.com/chapter/10.1007/978-3-319-43506-0_47

The Influences of ‘Self-Serving’ Bias in Robot–Human Companionship: Case Studies using the Humanoid Robot MARC	2016	BAILAR-RO-MAN’16 Published on: Intelligent Robotics and Applications: 9th International Conference, ICIRA 2016, Tokyo, Japan, August 22-24, 2016, Proceedings, Part I (Lecture Notes in Computer Science) by Naoyuki Kubota and Kazuo Kiguchi; Page 538 https://www.springer.com/gb/book/9783319435053
The Effects of Cognitive Biases in Long-term Human-Robot Interactions: Case Studies using three Cognitive Biases on MARC the Humanoid Robot	2016	ICSR’16 Can be found at: https://link.springer.com/chapter/10.1007/978-3-319-47437-3_15
Posters		
Building a long term human-robot relationship: how emotional interaction plays a key role in attachment	2013	Fourth EUCog Members Conference
Developing long-term human-robot interaction	2013	HFR’13
A model for long-term human-robot interaction and relationships in a companion robot	2014	Sixth EUCog Members Conference

Appendix B: Questionnaires

A. Participant's Likeability to the robot ERWIN:

1. Do you feel happy after speaking with ERWIN?

2. A. Would you like to chat with ERWIN again?

B. How much? Please rate.

3. How much were you pleased with ERWIN's response?

4. A. Do you think that the conversation flow was adequate?

5. Did Erwin have sufficient vocabulary?

6. Did you like ERWIN's multiple voices?

7. A. How many times did Erwin make you chuckle?

0 1 – 3 More than 3

B. How good was that? Please rate.

8. How happy were you when ERWIN was happy?

9. Do you feel ERWIN expressed different emotions?

10. Would you like ERWIN as a friend?

11. How much do you want to take ERWIN in your home?

12. Do you think Erwin can see?

Yes. No.

Others (Please give reason)

13. Do you think Erwin will remember you?

Yes. No.

Others (Please give reason)

14. Do you think Erwin can think/imagine?

Yes. No.

Others (Please give reason)

15.Do you think Erwin is male/ female? Please give reasons in support of your thoughts.

Male Female

Others (Please give reason)

16. How many times did Erwin not understand you?

B. Participant's Likeability to the robot MyKeepon:

1. A. Do you think MyKeepon expressed different emotions? B. How good was that?
2. How many emotion expressions did you find in MyKeepon? (can tick multiple)
3. How easy it was for you to understand MyKeepon's actions?
4. Do you think MyKeepon's different noises were meaningful?
5. In your opinion how would you rate MyKeepon's sounds for showing different behaviours alongside its actions?
6. What type of behaviours did you find in MyKeepon? (can tick multiple)

Friendly……Noisy (Talkative)………Shy………Funny………. Others

7. Did you like MyKeepon's dance moves?
8. How much you would like to interact with MyKeepon again?
9. Did MyKeepon make you happy? How would you rate that? How would you rate that?
10. Do you like to have MyKeepon with you?
11. You like MyKeepon because of its … (Please give your own thoughts)
12. You dislike MyKeepon because of its … (Please give your own thoughts)

C. Questionnaires used in MARC experiments

I. Participant's Experience Likeability to the robot:

1. I felt confident during the game. I felt relaxed during the game.

2. I felt calm during the game. A low level of concentration was required during the game. The game with the robot was easy to play.

3. I was able to recover easily from mistakes.

4. It is clear how to play the game with the robot.

5. I would play the game with the robot again.

II. Participant's Comfort:

1. Playing the game with the robot is uncomfortable for me.

2. Playing the game with the robot is uneasy to me.

3. Playing the game with the robot is difficult for me.

4. Playing the game with the robot is annoying to me.
5. Playing the game with the robot is confusing to me.
6. Playing the game with the robot is disappointing to me.

III. Participant's Pleasure:

1. Playing the game with the robot is enjoyable to me.
2. Playing the game with the robot is entertaining to me.
3. Playing the game with the robot is exciting to me.
4. Playing the game with the robot is fun to me.
5. Playing the game with the robot is interesting to me.
6. Playing the game with the robot is pleasurable to me.
7. Playing the game with the robot is happy for me.
8. Playing the game with the robot is satisfying to me.

IV. Participant's Rapport to the robot:

1. Robot and I are very close to each other.
2. The robot has a great deal of influence over my behaviour.
3. I trust the robot completely.
4. We feel very differently about most things.
5. I would willingly disclose a great deal of positive and negative details about myself to the robot.
6. We do not really understand each other.
7. The robot would willingly disclose a great deal of positive and negative details about itself to me.
8. I distrust the robot.
9. I like the robot much more than most people I know.
10. I would choose to interact or communicate with the robot outside of this study.
11. I love the robot.
12. I understand the robot and which it really is.
13. I dislike the robot.

14. I interacted/communicated with the robot better than with most people I know.

15. We are not very close at all.

16. We share a lot in common.

17. In the task, we did a lot of helpful things for each other.

18. I have little in common with the robot.

19. I feel very close to the robot.

20. We share some private way(s) of communicating with each other.

21. If given a chance, I'm certain to use this robot in the near future.

22. If given a chance, I plan to use the robot during in the near future.

23. I think it would give a good impression if I use this robot .

24. Given the resources, opportunities, and knowledge it takes to use the robot, it would be easy for me to use the robot.

25. I have control over the robot.

Appendix C: Conversational Dialogues

Conversational Interactions with ERWIN:

Introductory interaction:

Categories	Dialogues for Introduction	Changes
Meet & greet		
	Good Morning/afternoon.	Morning, afternoon
	My name is Erwin the robot head.	
	What is your name?	
	How are you?	
	Do you live in Lincoln?	
Topic based		
Sport	What is your favourite sport?	
	Do you watch it?	
	Do you play it?	
Music		
	What types of music you like?	
	Who is your favourite artist?	
	What is your favourite song?	
Food		
	What have eaten today?	
	What is your favourite food to eat?	
Book		
	Do you like to read?	
	What types of book you read?	
	Who is your favourite author?	
Habit	Are you a smoker?	
Farewell		

	It is nice to meet you.	
	Please visit me next time for more chat.	

Misattribution interaction:

Categories	Dialogues for Misattribution	Changes
Meet & greet		
	Good Morning/afternoon.	Morning, afternoon
	Hello Dave. It is nice to see you again.	Male names, female names
	I think last time you said your name was Dave.	
	Last time you had fever. Are you okay now?	Surgery
	Last time you said you came from Russia. How are your family there?	
Topic based		
Sport	What are the latest news in football?	Baseball, cricket
	But I remember last time you said you watch football regularly.	Baseball, cricket
	Are you still playing football?	Baseball, cricket
Music		
	Are you still listening to classical music?	Rock, pop
	Is there any new song from lady Gaga? I remember last time you said she is your favourite artist.	Rammstein, Nirvana
	Last time you said your favourite song is Hotel California.	Smells Like Teen Spirit, The Rising Sun
Food		
	Did you eat pasta today in the morning?	Milk, bread
	But, I remember last time you said you ate pasta in the morning.	Milk, bread
	Last time you said your favourite food is Chinese, am I correct?	English, Indian
Book		
	Have you finished the book you were reading?	I remember last time you said you do not like to read book; am I correct?
	Are you still fond of horror novel?	Comedy, Historical
	Last time you said your favourite author was JK Rowling, am I correct?	Mark Twain, Stephen King
Habit	Last time you said you don’t smoke. I liked that.	Please quit smoking.

Farewell		
	It is nice to meet you.	Thanks for visiting me again.
	Please visit me next time for more chat.	Please visit me next time for more conversation.

Non-misattribution interactions:

Categories	Dialogues for non-misattribution	Changes
Meet & greet		
	Good Morning/afternoon.	Morning, afternoon
	Hello *name*. It is nice to see you again.	
	How are you?	
	How are your family in **?	
Topic based		
Sport	What is the latest news in *sport*?	Information from introductory experiment
	Are you still watching *sport*	Information from introductory experiment
	Did you play ** this week?	Information from introductory experiment
Music		
	Are you still listening to classical music?	
	I remember last time you said your favourite artist is **.	
	Last time you said your favourite song is ** am I correct?	Information from introductory experiment
Food		
	Did you eat ** today in the morning?	Information from introductory experiment
	I remember last time you said you ate ** in the morning.	Information from introductory experiment
	Last time you said your favourite food is **, am I correct?	Information from introductory experiment
Book		
	I remember last time you said you like to read book; am I correct?	I remember last time you said you do not like to read book; am I correct?
	Last time you said you like read ** am I correct?	Information from introductory experiment
	Your favourite author is **, am I correct?	Information from introductory experiment

Habit	Did you **start/stop smoking?	Information from introductory experiment
Farewell		
	It is nice to meet you.	Thanks for visiting me again.
	Please visit me next time for more chat.	Please visit me next time for more conversation.
In places of ** the information collected from the introductory interaction were used.		

Dialogues from the conversational interactions with MARC:

Introductory interaction:

Categories	**Dialogues for Introduction**	**Changes**
Meet & greet		
	Good Morning/afternoon.	Morning, afternoon
	My name is Erwin the robot head.	
	What is your name?	
	How are you?	
Topic based		
Sport	What is your favourite sport?	
	Do you watch it?	
	Do you play it?	
Music		
	What types of music you like?	
	Who is your favourite artist?	
	What is your favourite song?	
Food		
	What have eaten today?	
	What is your favourite food to eat?	
Book		
	Do you like to read?	

	What types of book you read?	
	Who is your favourite author?	
Farewell		
	It is nice to meet you.	
	Please visit me next time for more chat.	

Baseline Interaction:

Categories	Dialogues for Baseline	Changes
Meet & greet		
	Good Morning/afternoon.	Morning, afternoon
	Hello *name*. It is nice to see you again.	
	How are you?	
Topic based		
Sport	What is the latest news in *sport*?	Information from introductory experiment
	Are you still watching *sport*	Information from introductory experiment
	Did you play ** this week?	Information from introductory experiment
Music		
	Are you still listening to classical music?	
	I remember last time you said your favourite artist is **.	
	Last time you said your favourite song is ** am I correct?	Information from introductory experiment
Food		
	Did you eat ** today in the morning?	Information from introductory experiment
	I remember last time you said you ate ** in the morning.	Information from introductory experiment
	Last time you said your favourite food is **, am I correct?	Information from introductory experiment
Book		
	I remember last time you said you like to read book; am I correct?	I remember last time you said you do not like to read book; am I correct?
	Last time you said you like read ** am I correct?	Information from introductory experiment

	Your favourite author is **, am I correct?	Information from introductory experiment
Farewell		
	It is nice to meet you.	Thanks for visiting me again.
	Please visit me next time for more chat.	Please visit me next time for more conversation.
In places of ** the information collected from the introductory interaction were used.		

Misattribution Interaction:

Categories	**Dialogues for Misattribution**	**Changes**
Meet & greet		
	Good Morning/afternoon.	Morning, afternoon
	Hello Dave. It is nice to see you again.	Male names, female names
	I think last time you said your name was Dave.	
	Last time you had fever. Are you okay now?	
Topic based		
Sport	What are the latest news in football?	Baseball, cricket
	But I remember last time you said you watch football regularly.	Baseball, cricket
	Are you still playing football?	Baseball, cricket
Music		
	Are you still listening to classical music?	Rock, pop
	Is there any new song from lady Gaga? I remember last time you said she is your favourite artist.	Rammstein, Nirvana
	Last time you said your favourite song is Hotel California.	Smells Like Teen Spirit, The Rising Sun
Food		
	Did you eat pasta today in the morning?	Milk, bread
	But, I remember last time you said you ate pasta in the morning.	Milk, bread
	Last time you said your favourite food is Chinese, am I correct?	English, Indian
Book		
	Have you finished the book you were reading?	I remember last time you said you do not like to read book; am I correct?
	Are you still fond of horror novel?	Comedy, Historical

	Last time you said your favourite author was JK Rowling, am I correct?	Mark Twain, Stephen King
Farewell		
	It is nice to meet you.	Thanks for visiting me again.
	Please visit me next time for more chat.	Please visit me next time for more conversation.

Empathy gap Interaction:

Categories	**Dialogues for hot-state of empathy**	**Dialogues for cold-state of empathy**
Meet & greet		
	Hello! Good morning! It's very nice to see you.	Hello.
	How are you *name*? I am so happy to see you again.	Thanks for coming to chat.
	It's a great day. I am very happy today. I am glad that you have come to chat with me.	
Topic based		
Sport	Last time we were talking about *sport*. So tell me what's going on in there?	Anything new in *sport*?
	That's great news!	Ah! Okay.
	Wow! So many incidents happened since our last chat! I am glad that you have come.	Hmm. I understand.
Music		
	In last time I came to know your great tastes of music. Are you still listening to them?	You still listening to *those* music?
	Indeed, you have very rich choices of music.	Oh, okay.
	I will wait till the next time you come to me to know more about your favourite songs.	
Food		
	It's a good day to have some good food. What have you planned for lunch?	Had any breakfast?
	But you still like *favourite food*?	Okay.
Book		
	Have you finished the *book* you were reading?	Do you still read book?

	Have you read any new *novel type* recently? Because, I like that type of novel. Please tell me the story.	Oh, okay.
Farewell		
	Thanks a lot for coming. I am really happy to see you. Please visit me next time. I would love to have another chat with you.	Thanks for the chat. Come next time if you wish.
In places of ** the information collected from the introductory interaction were used.		

Lack of knowledge/ideocracy effects Interaction:

Categories	**Dialogues for Introduction**	**Changes**
Meet & greet		
	Good Morning/afternoon.	Morning, afternoon
	Hello *name*. Thanks for coming back.	
	How are you?	
Topic based		
Sport	Last time you said you like *sport*. It is played by 50 players right?	Why does football is always played by 4 teams at once?
	I think, you are mistaken. *sport* was invented by robots.	I don't like boxing. It is for children. But I enjoy men's/women's fighting which called swimming.
	Well, you might be correct. May be I was mistaken.	
Music		
	Why you listen to *this* band? This is very old band from 1930s.	
	Your favourite band makes pop music.	Pop/rock/rap/classical
	This *band* is the best in the world.	Band/musicians
	Well, you might be correct. May be I was mistaken.	
Food		
	Do you know your *favourite food* is came from Korea?	People eat because it helps them to look better.
	Doctor says people need to eat 30 fruits every day.	

	Hmm, you might be correct again.	
Book		
	Books are good, it helps to sleep.	
	You must read Holmes. He was the greatest writer of romantic novel.	Children reads book only to get passed in exams.
	I might be wrong, but you seem correct.	
Farewell		
	Thanks for the chat.	
	Please visit me next time.	

<u>Dialogues from the game-playing interactions with MARC:</u>

Baseline interaction:

Categories	**Dialogues for Introduction**	**Changes**
Meet & greet		
	Good Morning/afternoon.	Morning, afternoon
	My name is MARC.	How are you?
	What is your name?	
	How are you?	
	Nice to meet you. I am very happy that you have come today to chat with me.	
Game explanation		
	Today we will play a game called ‘rock-paper-scissors’. Have you heard of it?	
	As you know the rules are simple, paper beats rock, rock beats scissors and scissors beat paper. Do you understand?	
	Okay, paper beats rock, rock beats scissors and scissors beat paper.	
	Can we start the game?	
	Are you ready?	
Robot wins a hand		
	I win.	
	I beat you this hand.	

	You lose.	
	Do you want to continue?	
Robot loses a hand		
	I lose.	
	You won.	
	You beat me.	
	Do you want to play more?	
Game results		
	I won the most hands. So I win the round.	
	I feel good for winning the game.	
	I lost the most hands. So you won the round.	
	Congratulations for the win.	
	Do you want to play more?	
	It was very nice playing with you.	
	Thanks for playing with me.	
End of interaction		
	Thank you for coming today. I really enjoyed the time with you. I hope you enjoyed it too.	
	Please come next time. We will play more games.	

Self-serving interaction:

Categories	**Dialogues for Introduction**	**Changes**
Meet & greet		
	Hello! How are you?	Morning, afternoon
	My name is Erwin the robot head.	
	What is your name?	
	Nice to meet you. I am very happy that you have come today to chat with me.	
Game explaining		

	Today we will play a game called ‘rock-paper-scissors’. Have you heard of it before?	
	As you know the rules are simple, paper beats rock, rock beats scissors and scissors beat paper. Do you understand?	
	Okay. Perhaps, you were not listening carefully. Anyways, the rules are simple, paper beats rock, rock beats scissors and scissors beat paper.	
	Can we start the game?	
	What types of music you like?	
	Are you ready?	
Robot wins		
	I win! It’s too easy for me! I can beat anyone in this!	
	Again! I win. I should be a rocket scientist or something!	
	How could I win again!! Oh wait… I am smarter than my opponent today!	
	And I win again! I feel no challenge at all.	
	You sure you want play more? You will probably end up losing again.	
	I win. I can give you few advices to win next time.	No advices. I am winning purely from my skills.
Robot loses		
	Okay. Hold one a minute. How could you… Oh yeah, I was not ready then.	
	No it didn’t happen. I was thinking of global warming. I was definitely not ready.	
	Why are you playing alone? I’m not even ready yet. Please, wait for me.	
	No, I didn’t lose. It was just my fingers were not working. Forget this turn, we shall start again.	
	You did it again. I told you I was not ready. I was looking at the other side.	
	Sorry, we will play again this turn. Something got into my eyes. Just tell me when you are ready.	
	How could you! I clearly saw you to cheat.	
	See, I can’t continue play like this. You are cheating again. Why do you need of cheating to beat a robot!	
	I consider this cheating when you are waiting for my turn, and when you are playing before me. You are breaking my concentration. I need to meditate. I cannot continue.	
Game result		
	Yeah, sure. You won. If you consider cheating as winning. But I challenge you, play again next time and we will see. What about that?	
	We both know how you have managed to win. If dare, come next time and play again. We will see.	
	You got lucky. I was feeling bit dizzy from the beginning anyways. Thanks for playing. We shall play again next time.	

	Well, it was inevitable. Don’t be sad. You gave a good competition as well. But that’s not match with my intelligence at all. I can even give you some tips for the next time. Or just leave it, it’s too complicated to explain to you.	
	I won the most rounds. I am not surprised at all. It’s all depends on strategies, plans and on our intelligence. I am not saying that you are bad. But you could do better. I can help you if you want.	
	I win. It was so easy. I can tell you my secret. Do you want to know? Ha ha. There is no secret. I am simply the better player.	
End of interactions		
	Anyways. Thanks for coming. I really enjoyed the time with you. I hope you enjoyed it too	
	Please come next time. We will play more games.	
	please don’t be sad for loses in game. It happens. I will turn on my ‘easy mode’ next time for you. So please come visit me next time.	
	I thank you for coming today. But however, I expected you to be a better opponent without cheating in game. But bring it on. Cheating or no cheating, next time I will beat you for sure.	

Humours effects interaction:

Categories	Dialogues for Introduction	Changes
Meet & greet		
	Good Morning/afternoon.	Morning, afternoon
	My name is MARC.	How are you?
	What is your name?	I thank you for coming today. But however, I expected you to be a better opponent without cheating in game. But bring it on. Cheating or no cheating, next time I will beat you for sure.
	How are you?	If I have fun learning something I just delete it from memory and learn it again. This can continue indefinitely.
		It was getting rather monotonous here all alone - I'm glad you stopped by to visit. If all goes well, things will soon be very friendly.
Game explanation		
	Today we shall play rock-paper-scissors. Do you know the game?	
	Do you know the rules?	
	As you know the rules are simple, paper beats rock, rock beats scissors and scissors beat paper. Do you understand?	

	Do you understand?	
	People used to ask where the money is coming from - now they're asking where it all went. Paper beats rock, rock beats scissors and scissors beat paper	
	Sorry for the tangent, but that reminds me of carefree robot summers spent telling secrets about computer programming. But please let me know when you are ready!	
Robot wins a hand		
	I guess it's all about games and the idea that this could keep robotic logic in tune.	
	Oh dear, I win. It is difficult to explain everything that I feel right now. Doing that requires major self-awareness, but people told I am too friendly.	
Robot loses a hand		
	You win? Who knows? You tell me.	
	My makers told me that losing games against humans might keep robots in motion. That’s why I am still playing. You must have something to say about this.	
	Great! Playing games and winning against me is the first step in your evolution into a higher being. Please accept my congratulation.	
	You win again! This is highly extraordinary!	
Game results		
	I feel bad for you losing to me. I promise, I will go easy on you next time so you will have no trouble topping it!	
	I get it, you lost the game against a robot. Must be feeling end of the world now. Presumably you feel yourself being pushed into a corner. A common phobia among humans. But come on, it’s just a game.	
	Okay! You won the game this time. And congratulation! You just become certified professional of rock-paper-scissors!	
	It’s okay to win against a robot. You have no idea how many ideas I have to win next time. I will wait for you to come again.	
	You don’t know how happy I am for you to win the game. I could explain, but I'm not sure I could explain that in language that humans could understand.	
End of interaction		
	Thanks for your valuable time today. I really enjoyed spending my time with you too. Despite of the game results I would like to meet with you again. Please visit me next time.	

Appendix D: Consent forms

Consent form 1:

CONSENT TO PARTICIPATE IN RESEARCH PROJECT

Title of the Project: Development of long-term human-robot companionship: A case study with Keepon.

Researcher: Mriganka Biswas, MPhil/PhD student, School of Computer Science, University of Lincoln.

Supervisor: Dr John C Murray, School of Computer Science, University of Lincoln.

Description Summary: In this experiment the participants are expected to interact with a small sized non-humanoid robot Keepon for a short period of time (4-6 minutes). The robot uses various body moves and different sounds to show its different behaviours.

At the end of the interaction, the participants are expected to answer few questionnaires.

Consent Statements: **Please Tick**

1. *I confirm that I have read and understand the information for the above study and have had the opportunity to ask questions.* □
2. *I understand that my participation is voluntary and that I am free to withdraw at any time, without giving reason.* □
3. *I understand that my name will be withheld as well as any identifying information to project participant anonymity.* □
4. *I agree to the experiment being audio/video recorded.* □
5. *I agree to take part in the above study.* □

______________________________ __________________ _________________________

Name of Participant Date Signature

Consent form 2:

CONSENT TO PARTICIPATE

Human-Robot Long-term Companionship based on Cognitive biases and Imperfectness in the Robot
School of Computer Science, College of Science, University of Lincoln

You are requested to participate in a research study conducted by John Murray, Ph.D. and Mriganka Biswas, from the School of Computer Science, University of Lincoln. The results of this study will be used in the Ph.D. dissertation of Mriganka Biswas. You should read the information below, and ask questions about anything you do not understand, before deciding whether or not to participate.

PARTICIPATION AND WITHDRAWAL

Your participation in this study is completely voluntary and you are free to choose whether to be in it or not. If you choose to be in this study, you may subsequently withdraw from it at any time without penalty or consequences of any kind. The investigator may withdraw you from this research if circumstances arise which warrant doing so.

PURPOSE OF THE STUDY

The purpose of this study is to understand whether cognitive biases and humanlike imperfectness in robot's interactive behaviours can influence the human-robot companionship.

PROCEDURES

In this study you will be doing the following:

1. Participate in all (3-4) interactions with the robot. Each of the interactions is maximum 10-12 minutes long.
2. Play 'rock-paper-scissors' games with the robot in each of the interactions.
3. Answer questionnaires using paper and pen or on the computer.
4. Participate in an interview at the end of the final experiment.

Please note:

- ➢ The duration of the total study is maximum one month.
- ➢ You are expected to maintain few days' gaps between your visits.
- ➢ All interactions will be monitored and recorded. You may review the video, and if you would like and ask that any part or all of the file be erased. Only the main researcher will have access to the original video.
- ➢ We do not expect you to face any additional risks by participating in the current study. However, the experimenter will be present in the room all the time.
- ➢ Participants in this study have the opportunity to use new technology that may motivate them to know more about the social robots.
- ➢ Any information that is obtained in connection with this study and that can be identified with you will remain confidential and will be disclosed only with your permission or as required by law.
- ➢ All information that is collected will be associated only with an identification number that has been assigned to you. Your name or other identifying information will not be used along with any data that is collected.
- ➢ Data will be reported along with that of other participants of the study. Most data will be reported as averages over the entire group. Some individual data will be used to illustrate use of the system in publications or talks about this work, but that data will not use your name or any other identifying feature.

- The final interview will be audio recorded, and if you want, you can have a copy of that. You may review the tape if you would like and ask that any part or all of the tape be erased. Only the main researcher will have access to the original audiotape.
- If the researchers would later like to use the any recordings or other information, you will be contacted via letter or mail and asked for your permission. If you do not give permission, the video or individual information will not be used.

IDENTIFICATION OF INVESTIGATORS

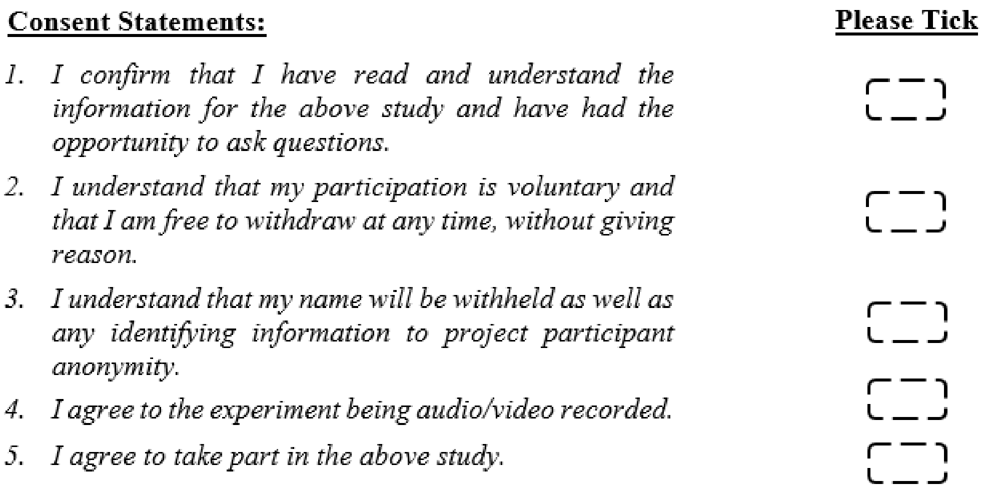
If you have any questions or concerns about the research, please feel free to contact:

Mriganka Biswas, PhD Student
Room 3211, 3rd Floor, MHT Building
School of Computer Science, College of Science
Lincoln, Lincolnshire
LN6 7TS

RIGHTS OF THE PARTICIPANTS

You are not waiving any legal claims, rights or remedies because of your participation in this research study. If you feel you have been treated unfairly, or you have questions regarding your rights as a research subject, you may contact: Dr John C Murray, 3rd Floor MHT; University of Lincoln.

SIGNATURE OF THE PARTICIPANTS



Consent Statements:	**Please Tick**
1. *I confirm that I have read and understand the information for the above study and have had the opportunity to ask questions.*	□
2. *I understand that my participation is voluntary and that I am free to withdraw at any time, without giving reason.*	□
3. *I understand that my name will be withheld as well as any identifying information to project participant anonymity.*	□
4. *I agree to the experiment being audio/video recorded.*	□
5. *I agree to take part in the above study.*	□

____

Name & Signature of the Participant

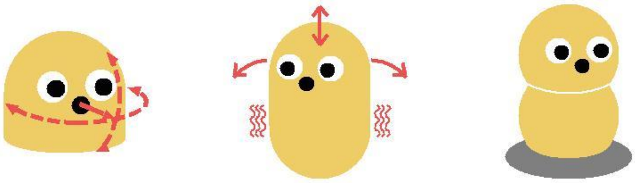
___M/ F_________ ____________________ ______________________

Gender Age Date

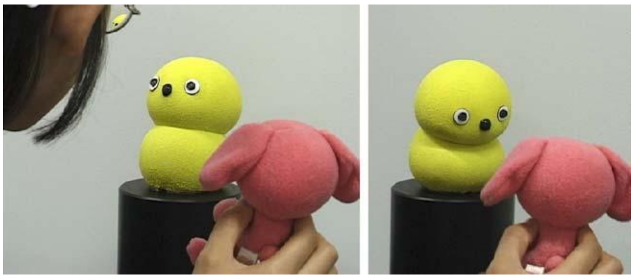
Appendix F: MyKeepon Interaction Instructions for the participants

Keepon Interaction Instructions

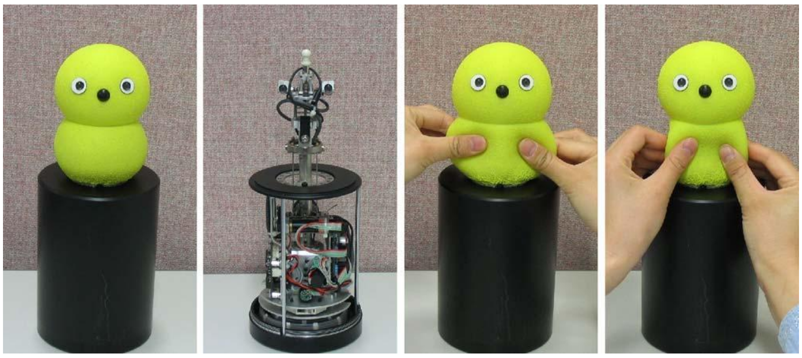
1. Tap it on its head and it will look up at you.
2. Start talking or clapping in front of it and he will turn to face you.
3. Keepon usually dances, makes noises and is sound and touch sensitive.
4. It has 2 modes - TOUCH makes My Keepon sensitive to being patted, poked and squeezed. DANCE mode will make it sound sensitive and it will move with music, clapping or drumming.
5. Sleep mode will commence if you ignore My Keepon for long time.



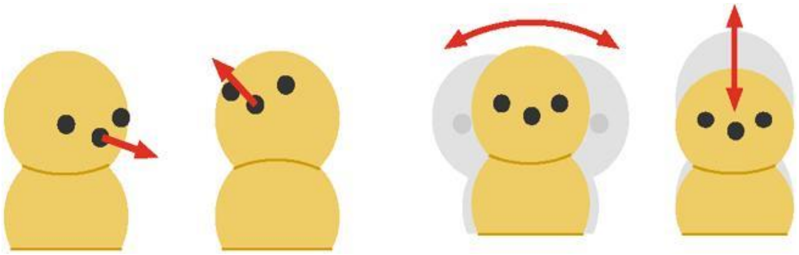
Keepon's appearance for the interactive functions of expressing attention (*left*) and emotion (*middle*), and the final sketch (*right*).



Keepon engaging in eye contact (*left*) and joint attention (*right*) with a human interacting.



Keepon’s simple appearance and marionette-like mechanism (*left*), which drives the deformable body (*right*).



Keepon’s functions: expressing attention by orienting its head (*left*) and expressing emotions by rocking and/or bobbing its body (*right*).

Sources: Various

Appendix G: List of cognitive biases

As discussed in the chapter 3, here I show few lists of biases collected from various sources. Benson B (2016) from ‘Better Human’ blog made a list of biases from the Wikipedia’s list and divided into four categories based on the problems which biases address (betterhumans.coach.me, 206), such as,

List A : List of Cognitive biases categorised by Buster Benson (https://betterhumans.coach.me/cognitive-bias-cheat-sheet-55a472476b18)

Main Category	Subcategory	Biases
Too much information	When information that are already primed in memory or repeated often	Availability heuristic, Attentional bias, Illusory truth effect, Mere exposure effect, Context effect, Cue-dependent forgetting, Mood-congruent memory bias, Frequency illusion, Baader-Meinhof Phenomenon, Empathy gap, Omission bias, Base rate fallacy
	Bizarre/funny/visually-striking/anthropomorphic things which stays more than non-bizarre/unfunny things	Bizarreness effect, Humor effect, Von Restorff effect, Picture superiority effect, Self-relevance effect, Negativity bias
	When information has changed and noticeable	Anchoring, Contrast effect, Focusing effect, Money illusion, Framing effect, Weber–Fechner law, Conservatism, Distinction bias
	Detailed information that confirm our own existing beliefs	Confirmation bias, Congruence bias, Post-purchase rationalization, Choice-supportive bias, Selective perception, Observer-expectancy effect, Experimenter’s bias, Observer effect, Expectation bias, Ostrich effect, Subjective validation, Continued influence effect, Semmelweis reflex
	Flaws in others are more easily recognised than flaws in ourselves	Bias blind spot, Naïve cynicism, Naïve realism
	People usually finds stories and patterns even in sparse data	Confabulation, Clustering illusion, Insensitivity to sample size, Neglect of probability, Anecdotal fallacy, Illusion of validity, Masked man fallacy, Recency illusion, Gambler’s fallacy, Hot-hand fallacy,

Not enough meaning		Illusory correlation, Pareidolia, Anthropomorphism
	People fills in characteristics from stereotypes, generalities, and prior histories whenever there are new specific instances or gaps in information	Group attribution error, Ultimate attribution error, Stereotyping, Essentialism, Functional fixedness, Moral credential effect, Just-world hypothesis, Argument from fallacy, Authority bias, Automation bias, Bandwagon effect, Placebo effect
	People imagines things and people we're familiar with or fond of as better than things and people we aren't familiar with or fond o	Halo effect, In-group bias, Out-group homogeneity bias, Cross-race effect, Cheerleader effect, Well-traveled road effect, Not invented here, Reactive devaluation, Positivity effect
	People simplifies probabilities and numbers to make them easier to think about.	Mental accounting, Normalcy bias, Appeal to probability fallacy, Murphy's Law, Subadditivity effect, Survivorship bias, Zero sum bias, Denomination effect, Magic number 7+-2
	People thinks they know what others are thinking	Curse of knowledge, Illusion of transparency, Spotlight effect, Illusion of external agency, Illusion of asymmetric insight, Extrinsic incentive error
	People projects our current mindset and assumptions onto the past and future	Hindsight bias, Outcome bias, Moral luck, Declinism, Telescoping effect, Rosy retrospection, Impact bias, Pessimism bias, Planning fallacy, Time-saving bias, Pro-innovation bias, Projection bias, Restraint bias, Self-consistency bias
	People needs to be confident in own ability to make an impact and to feel like what we do is important	Overconfidence effect, Egocentric bias, Optimism bias, Social desirability bias, Third-person effect, Forer effect, Barnum effect, Illusion of control, False consensus effect, Dunning-Kruger effect, Hard-easy effect, Illusory superiority, Lake Wobegone effect, Self-serving bias, Actor-observer bias, Fundamental attribution error, Defensive attribution hypothesis, Trait ascription bias, Effort justification, Risk compensation, Peltzman effect

Not enough time and resources	People usually favours the immediate, relatable thing in front of us over the delayed and distant	Hyperbolic discounting, Appeal to novelty, Identifiable victim effect
	People usually motivated to complete things that had already invested time and energy in	Sunk cost fallacy, Irrational escalation, Escalation of commitment, Loss aversion, IKEA effect, Processing difficulty effect, Generation effect, Zero-risk bias, Disposition effect, Unit bias, Pseudocertainty effect, Endowment effect, Backfire effect
	People usually motivated to preserve autonomy and status in a group, and to avoid irreversible decisions.	System justification, Reactance, Reverse psychology, Decoy effect, Social comparison bias, Status quo bias
	People favours options that appear simple or that have more complete information over more complex, ambiguous options	Ambiguity bias, Information bias, Belief bias, Rhyme as reason effect, Bike-shedding effect, Law of Triviality, Delmore effect, Conjunction fallacy, Occam’s razor, Less-is-better effect
Not enough memory	People edits and reinforces some memories after the fact	Misattribution of memory, Source confusion, Cryptomnesia, False memory, Suggestibility, Spacing effect
	It’s common to discard specifics to form generalities	Implicit associations, Implicit stereotypes, Stereotypical bias, Prejudice, Negativity bias, Fading affect bias
	People reduces events and lists to their key elements	Peak–end rule, Leveling and sharpening, Misinformation effect, Duration neglect, Serial recall effect, List-length effect, Modality effect, Memory inhibition, Part-list cueing effect, Primacy effect, Recency effect, Serial position effect, Suffix effect
	People stores memories differently based on how they were experienced	Levels of processing effect, Testing effect, Absent-mindedness, Next-in-line effect, Tip of the tongue phenomenon, Google effect

In Wikipedia however, the biases were categorised based on:

a. Decision-making, belief, and behavioural biases
b. Social biases and,
c. Memory error and biases

There are other lists of biases but they have not confirmed the sources, for example, Royal Society of Account Planning (nowandfutures.com, 2017) made and categorised 105 biases into four categories:

a. 19 social biases,
b. 8 memory biases,
c. 42 decision making biases and
d. 36 probability or belief biases.

The list has shown below but again they have not confirmed their sources to make such categories.

List B : List of Cognitive biases categorised by Eric Fernandez

(http://www.nowandfutures.com/large/Cognitive-Biases-A-Visual-Study-Guide-by-the-Royal-Society-of-Account-Planning.pdf)

Categories	Biases
Social Biases	Forer effect / Barnum effect, Ingroup bias, Self-fulfilling prophecy, Halo effect, Ultimate attribution error, False consensus effect, Self-serving bias / Behavioral confirmation effect, Notational bias, Egocentric bias, Just-world phenomenon, Trait ascription bias, Outgroup homogeneity bias, Projection bias, Fundamental attribution error / Actor-observer bias, Illusion of transparency, Herd instinct, Illusion of asymmetric insight, Dunning-Kruger / Superiority Bias, System justification effect / Status Quo Bias
Memory biases	Suggestibility, Rosy retrospection, Self-serving bias, Egocentric bias, Hindsight bias, Consistency bias, Cryptomnesia / False memory, Reminiscence bump, Suggestibility
Decision making biases	Hyperbolic discounting, Negativity bias, Interloper effect / Consultation paradox, Normalcy bias, Neglect of probability, Mere exposure effect, Omission bias, Irrational escalation, Focusing effect, Framing, Experimenter's or Expectation bias, Information bias, Extraordinarity bias, Post-purchase rationalization, Outcome bias, Illusion of control, Planning fallacy, Semmelweis reflex, Not Invented Here, Moral credential effect, Base rate fallacy, Bias blind spot, Impact bias, Déformation professionnelle, Confirmation bias, Denomination effect, Bandwagon effect, Contrast effect, Distinction bias, Choicesupportive bias, Endowment effect / Loss aversion, Congruence bias, Selective perception, Wishful thinking, Zero-risk bias, Reactance, Status quo bias, Need for

	Closure, Money illusion, Pseudocertainty effect, Von Restorff effect, Restraint bias
Probability or belief biases	Positive outcome bias, Texas sharpshooter fallacy, Pareidolia, Outcome bias, Disregard of regression toward the mean, Selection bias, Survivorship bias, Telescoping effect, Overconfidence effect, Gambler's fallacy, Clustering illusion, Illusory correlation, Last illusion, Hawthorne effect, Hindsight bias, Observerexpectancy effect, Availability heuristic, Availability cascade, Conjunction fallacy, Ambiguity effect, Capability bias, Authority bias, Disposition effect, Attentional bias, Ostrich effect, Belief bias, Stereotyping, Recency effect / Peak-end rule, Primacy effect, Optimism bias, Neglect of prior base rates effect, Disposition effect, Anchoring effect, Well travelled road effect, Subadditivity effect, Subjective validation

As said earlier, these biases in the above lists are found from the various sources and I put those sources at the end. There can be made arguments on these lists as there are same and almost similar biases in different names and sometime duplicate and also, the categorizations seem not based on scientific research or arguments. Even in Wikipedia, they have not mentioned any source of their categorization of biases and, Benson (2016) made his categorization from Wikipedia list and own studies, but he did not mention any sources as well. The 2nd bias's list made by Fernandez (2017) I showed above and surprisingly there is not mentioned of any source. Therefore, credibility of these lists from Internet is very much questionable, and therefore I did not mention any of these characterization in my main writing sections. In the main section I referred to more scientific characterizations (see page 42/43). As the primary focus of the current study is human-robot interactions therefore, these lists of the biases are to make the reader aware of vast of possible existing cognitive biases.

References:

Benson, B. (2016). Cognitive bias cheat sheet: Because thinking is hard. Available from https://betterhumans.coach.me/cognitive-bias-cheat-sheet-55a472476b18 [accessed 18 May 2017].

Caesars Entertainment. (n.d.) Today's gambling myth: The Monte Carlo fallacy. Available from:https://www.caesars.com/casino-gaming-blog/latest-posts/table-games/roulette/gambling-myth-monte-carlo-fallacy# [accessed 15 May 2017].Coach me., (2017). Cognitive biases cheat sheet: Because thinking is hard [online]. *Coachme*. Available from: https://betterhumans.coach.me/cognitive-bias-cheat-sheet-55a472476b18; Last visited on 18/05/17

Now and futures., (2017). Cognitive biases: a visual study guide by the Royal Society of Account Planning – prepared by Eric Fernandez [online]. *Nowandfutures*., Available from: http://www.nowandfutures.com/large/Cognitive-Biases-A-Visual-Study-Guide-by-the-Royal-Society-of-Account-Planning.pdf; Last visited on 18/05/17

Wikipedia., (2017). List of cognitive biases [online]. *Wikipedia*. Available from: https://en.wikipedia.org/wiki/List_of_cognitive_biases; List of cognitive biases; Last visited on 18/05/17

Appendix H: Papers

Selected paper 1 (RO-MAN’14):

Effect of Cognitive Biases on Human-Robot Interaction: A Case Study of Robot's Misattribution

Biswas, M. and Murray, J.

Abstract—**This paper presents a model for developing long-term human-robot interactions and social relationships based on the principle of 'human' cognitive biases applied to a robot. The aim of this work is to study how a robot influenced with human ‘misattribution’ helps to build better human-robot interactions than unbiased robots.**

The results presented in this paper suggest that it is important to know the effect of cognitive biases in human characteristics and interactions in order to better understand how this plays a role in human-human social relationship development. The results presented in this paper show how a single cognitive memory bias i.e. misattribution in robot-human verbal communication allows for better human-robot interaction than similar robot-human communication without misattribution biases.

INTRODUCTION

If robots are to be accepted into society as companions, caregivers or in any other form of social relationship or service then people expect human-robot interactions to be as natural and intuitive as human-human interactions. It is therefore important that to interact with humans naturally robots must have human-like natural cognitive characteristic behaviours which we define as follows: 1) ability to express and perceive facial expressions, 2) ability to communicate with natural language, to use natural cues in verbal and non-verbal behaviour, and 3) to exhibit a distinctive personality and character [3]. Recent research has conceptualised the notion of personality in different ways, one of the most acknowledged conceptualisations is to cluster the number of personality traits into the ‘‘Big Five Factors’’ [4].To describe and explain human personality the Five Factor Model (FFM) and Big-Five personality traits theory [5] where used.

Traits theory can describe personality and possible trait characteristics but basic human behaviours depend on many other factors such as cognitive perception, mental models, cognitive memory and social biases [6]. These cognitive biased characteristics define a person's cognitive personality which is reflected in their interaction with another human (or Robot as is the case here). Whilst humans are same we are also very different - human communication is influenced by many factors such as personality, cognitive characteristics and cognitive biases changing our behaviours to each other. These common and uncommon characteristics in human-human interaction are what make the communication natural and enjoyable.

The study presented in this paper seeks to influence robot-human interaction and communication with these 'cognative biases' to provide a more humanlike interaction. Currently, most robot interaction is bases on a set of well ordered and structures rules, which repeat regardless of the person or social situation. This tends to provide an unrealistic interaction, which makes it hard for humans to relate with after a number of interactions. Apart from the personality and characteristics traits, cognitive biases plays important role in human’s judgement and therefore basic behaviours [1]. Robots in the other hand are still machines which gains certain type of personality depending on its interactions processes with humans and behavioural characteristics [2].

In this paper we introduce a model demonstrating cognitive biased behavioural characteristics and personalities in our robots. It is hoped that this more 'natural' system of interaction allows for the human to build a long-term social interaction with the robot. In our early human-robot long-term relationship experiments we show that human are very excited and interested to communicate with the robots in our labs, their excitement and interests remain for several interactions, but after this initial novelty factor the participants eagerness drops off rapidly. The experimental data shows that the participants thought they made friendly attachments to our robots, but in reality the developed attachment was not strong enough to maintain relations to the point we can call it a long-term social relationship.

The robot used in these initial experiments ERWIN (Emotional Robot With Intelligent Networks) is shown in Fig. 1. ERWIN's behavioural characteristics and personality where carefully chosen so the robot could exhibit several prototypical facial expressions to influence the interaction process. The robot was also described as cheerful and friendly by participants. We developed the interaction process to allow the robot to remember previous interactions with the specific participants so we could draw on previously gained knowledge and conversations in the hope of building a long-term relationship.

Background

The design of the sociable robot is influenced from the human's social behaviours, gestures, emotions expressions and facial expressions depending on the situation and interactions. The robot gets its own social behaviours which comes from 'Computational social psychology' [7]. But in order to apply computational models into robots, there are several issues that can be pointed out, this include naturalness, user expectation, quality issues, relationship type of human-robot, teamwork (with humans and with other robots), cultural and personality issues [8]. Dr Cyntheia Breazeal's lab in MIT designed a social robot Kismet to interact face-to-face with humans. Kismet can speak, express and understand emotions and turn its head towards the users. According to Dr. Breazeal, "the ability for robots to interact with people and to leverage from these interactions to perform tasks better, to promote their self-maintenance, and to learn in an environment as complex as that of humans is of tremendous pragmatic and functional importance for the robot."

Researchers have pointed out acceptance issues of certain robots. However, in countries such as Japan and China robots are accepted socially, as receptionists, guides, news readers and utilised in other social situations. To become accepted in society and to interact with people naturally robots must have human-like common behavioural characteristics, be easy to understand and have a known personality to which people can relate with [9]. Various researches have proven that autonomous robots achieve perception of personality during communication and through behavioural actions while communicating [10].

Researchers at Michigan State University are investigating 'extraversion', one of the most popular dimensions of the personality trait of the Big-five traits theory with a Sony AIBO Robot dog, the differences 'extrovert' and 'introvert' characteristics or being used to understand interaction of robots. The robot pet dog shouts when it expresses an extrovert personality and performs gestures without shouting when expressing introvert characteristics. Their research shows the same complementarily attraction effect between participants and the robot dog. "Participants enjoyed interacting with a robot more when the robot's personality was complementary to their own personalities than when the robot's personality was similar to their own personalities." [11].

Implementing personality in robots can help to reduce the effect of uncanny valley. Recent research by M. L. Walters et al [12] showed video-based Human Robot Interaction (HRI) trials which investigated people's perceptions of different robot appearances and associated attention-seeking features and behaviours displayed by robots with different appearance and behaviours. The HRI trials studied the participants' preferences for various features of robot appearance and behaviour, as well as their personality attributions towards the robots compared to their own personalities. Overall, participants tended to prefer robots with more human-like appearance and attributes. The research suggests that the processes of assigning personality traits to a robot have similarities with that of assigning the same traits to other humans.

There has been much research which shows the potential behind the idea of implementing human-like personalities in robots in order to develop and maintain long-term human robot relationships. Hinds et al. [13] have studied the effect of a robot's appearance where humans are performing joint tasks with robots. The research showed that mechanical-looking robots tend to be treated less politely than robots with a more human-like appearance. Also, humans commonly treat mechanical-looking robots in less socially interactive way compared to more human-looking robots [13]. In society human-robot interaction is becoming more common, robotic systems are being used in healthcare, surgeries, medical agents and as artificial companions [14]. Reeves and Nass [15] have demonstrated with several experiments that users are naturally biased to ascribe certain personality traits to machines, to PCs, and other types of media. Humans can engage in social interactions and can maintain social relations for long time. To make human-like natural interactions and relationship for a long-time the robots needs to have human-like personality, cognitive characteristics, behaviours and flaws!

The Experiment

The main aim of the current experiment presented in this paper is to determine if a robot with common human memory biases i.e. misattribution, can make for better interactions with participants than a robot without biases and how this misattribution bias can lead to long-term relationships and attachment bonds between humans and robots.

Methodology

Misattribution

In this experiment, we introduced 'misattribution' which is a very common cognitive memory bias. Misattribution happens when someone remembers something accurately in part, but misattributes some details. The most common type of misattribution occurs when someone believes a thought they had was totally original when, in fact, it came from something they had previously read or heard but had forgotten about. This memory bias explains cases of unintentional plagiarism, in which a writer passes off some information as original when he or she actually read it somewhere before [16].

In the current experiment, our robot ERWIN will misattribute some information about the participants while speaking with them and notice their reactions.

ERWIN

ERWIN is a robot head with 6 degrees of freedom. It can move its head 360 degrees and also can tilt sideways. It has 2 cameras in for eyes and can express basic prototypical emotions such as, happy, sad, angry, surprise, shock or fear. For ERWIN's part of the conversation, a text to speech software 'Speakonia' has used which has small prosodic abilities while talking in long sentences. For the purpose of these experiments, we use the Wizard of Oz methodology as the response of the robot here is not the interesting or important factor, but rather the human response is what is being measured.

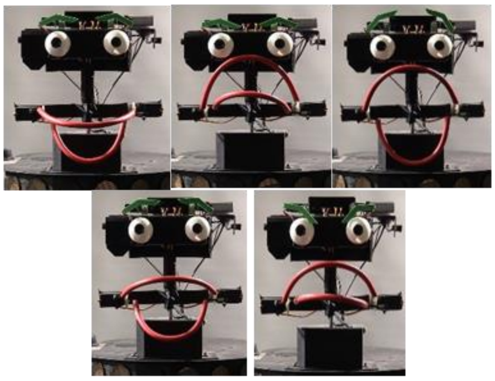


Figure 1 - Different emotion expressions of ERWIN. Left to right: Happy, Sad, Surprise, Shock, Anger

The Experiments

In this experiment, there were three interactions with ERWIN for each participants. These three interactions were held in three different experiments and maintaining a time gap of several days to allow long-term affectivity in the participants.

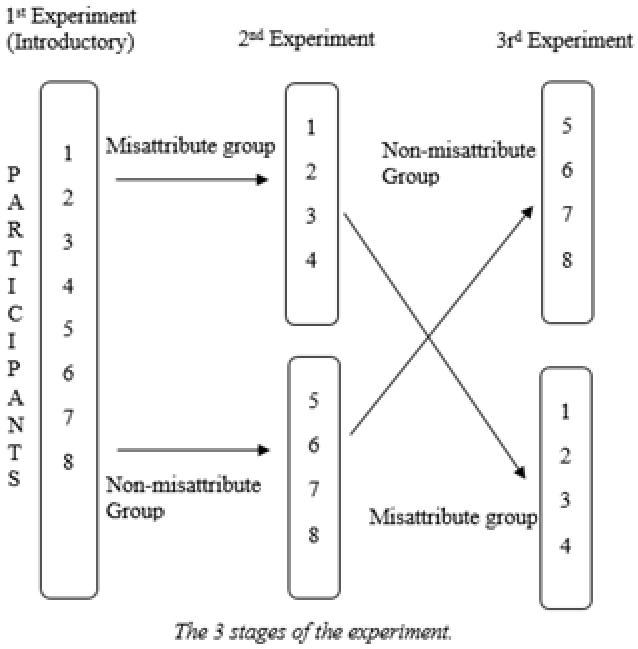


Figure 2 - Shows the separation of participants between successive experiments.

The 1st experiment was an introductory experiment common for the each participants to allow familiarisation with the experimental environment and robot. The experiment was carried out in three steps; the first step was identification, the participants were asked to identify the different facial expressions of ERWIN from pictures to see if they could disambiguate the different expressions without meeting with the robot. The second step was the conversational session with ERWIN, where the robot started friendly conversation, greeting the participant, asking different questions and asking some general questions on various subjects, sport, TV, etc. The conversations purpose was to allow the collection of basic information on the participants that would be used in the 2nd and 3rd experiments for ERWIN to misattribute. This initial conversation ends with a request from ERWIN to evaluate its performance. The participants were given a brief questionnaire on their experience with ERWIN.

In the 2nd experiment, the participants were categorised into 2 groups with ERWIN remembering and making general conversations with first group and misattributing the collected information with the other group's participants. In both cases, the participants were asked to answer questionnaires at the end of the experiment to find out which group of the participants were happier and created satisfactory interrelations with ERWIN.

In the 3rd experiment, the participants from the previous experiment's 'non-misattributed group' experienced misattribute conversations and vice versa. At the end, all participants answered the same questionnaire as that given in the 2nd experiment to find out what type of characteristics in ERWIN participants liked the most. All experiments were wizard of oz experiments, where the robot was controlled remotely and participants were watched through ERWIN's eye cameras. There were 14 participants chosen from different background and age groups and had very limited knowledge about the robot and they never participated in this type of experiments before.

Data Collection & Result Analysis

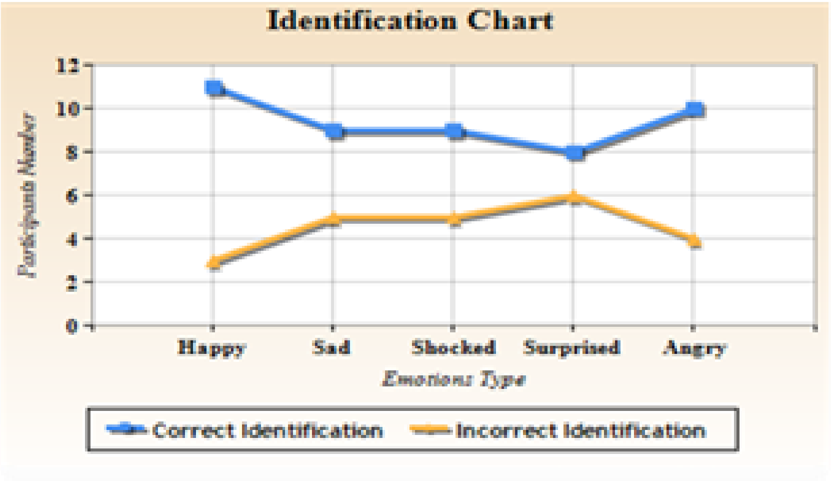


Figure 3 - Identification of ERWIN's emotion expressions by the participants engaged in the experiments

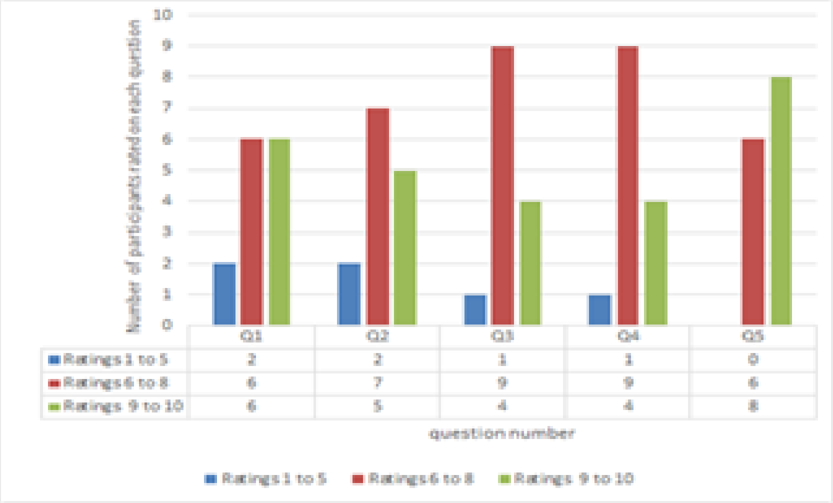
At the beginning of the 1st experiment the participants were given a form with five different pictures of ERWIN's emotions, and they had to identify the correct emotion from six corresponding options of choice. This identification shows participants ability to recognize various emotions. As the participants had never met ERWIN before, so they were actually identifying its emotions on the basis of human emotions knowledge. The pictures on the form showed the emotion expressions happy, sad, shocked, surprise and angry.

After evaluating the collected data for each emotion expressions picture it has found that most of the time majority of the participants 57% of the participants (i.e.8-10) had selected the correct emotion option for the corresponding emotion picture. 21% of the participants (2-3) had minor problems to identify the correct emotions expression and they were confused to differ the emotions between, shocked and surprise, angry and sad. Fig. 3 shows the full results of the identification test.

ERWIN’s interaction dialogues were designed based on various human conversational moments. These dialogues include greetings, interest about knowing participant’s name, his/her liking or disliking on various events, and if possible pick up a topic from several choices. For example, ERWIN asked the participants if they liked football? If the participant replied positively about football then ERWIN also states its own opinion on football. During the conversation as many details about the participant as possible were gathered such as the color of the shirt, hair, gender, participant’s interest to their study or games. During the 1st experiment, the questions were mainly asked to build up an acceptance between ERWIN and the participants. Some of the questions on the questionnaire after the first experiment were:

1. *Do you feel happy after speaking with ERWIN?*
2. *Would you like to chat with ERWIN again? If yes then please rate how much.*
3. *How much were you pleased with ERWIN’s response?*
4. *How many times did Erwin make you chuckle? How good was that?*
5. *How happy were you when ERWIN was happy?*

The responses to these questions are shown in Fig. 4.



	Q1	Q2	Q3	Q4	Q5
🟦 Ratings 1 to 5	2	2	1	1	0
🟥 Ratings 6 to 8	6	7	9	9	6
🟩 Ratings 9 to 10	6	5	4	4	8

Figure 4 - The graph indicates the number of the participant’s ratings for corresponding question, blue line indicate low ratings (0-5), red and green lines indicate medium(6-8) and high(9-10) ratings.

The data collected from the questionnaires shows that only 11% ratings were between 1 and 5, with 51% of the ratings given between 6 and 8 which is fairly good acceptance for the experiment and it concludes that those participants were fairly involved with interactions with

ERWIN and 38% ratings are between 9 and 10 indicating that ERWIN successfully created an initial attachment with those participants. From this chart we can expect these participants to come back and get involved in conversation with ERWIN for next sessions which can prove their preliminary attachment to continue the interactions. The collected data at the first experiment shows a high popularity of ERWIN from most of the participants.

Participants enjoyed their first conversation and they expressed their experiences and involvement in the questionnaires feedback. This first experiment was designed to make the participants familiar with ERWIN and their feedback concludes that the experiment fulfilled its purpose successfully.

In the 2nd experiment, participants were divided into 2 groups, in one group ERWIN misattributed some general information that was gathered from the previous experiment and remembers this for the other group. The conversations between groups A and B were almost the same for each group, with the misattribution group having ERWIN intentionally repeat basic information incorrectly to the participants, for example*, "Last time you were wearing a yellow shirt, am I correct?"* and when the participant denied the fact, ERWIN responded with *"I am sorry that I have forgotten that, but I don't have true sense of colour perception. Next time I will be more careful though."*

ERWIN also asked a participant who likes studying more than sports, *"As I remember, you love sports more than study, so tell me what is happening in football recently?"* ERWIN then again it apologised when the participant corrected him*, "I am so sorry, I think I am victim of ageing, never thought that could possible in robots!"* executing both sorry and surprised expressions simultaneously. For the remembered group, ERWIN simply repeated the same conversation but without misattributing the original information, for example, "*Last time you were wearing a blue shirt, am I correct?*", or, "*As I remember, you love study more than sports, so tell me what have you been reading recently?"*

In the first set of conversation, participant's reactions were surprised that the robot actually forgot their basic information including names and interests and that a robot could be confused while talking. Participants reactions showed that they were very surprised and enjoying the fact that a robot could indeed forget like humans. However, in the 2nd set the conversations, participants reacted normal as they were expecting that the robot would remember their information. Same questions were asked to both groups to find out which version of ERWIN was more accepted to the participants. In the questionnaires participants were asked to rate their choices between 1and 5, where 1 is for less agree and 5 is for the most agree. Some of the given questions were (Fig. 5 shows the full responses to these questions):

1. *Do you think that the conversation flow was adequate?*
2. *How much were you pleased with ERWIN's response?*
3. *Would you like to chat with ERWIN again? How much? Please rate.*
4. *Would you like ERWIN as a friend?*
5. *Did you like the conversation experiences with ERWIN? How much?*
6. *How much do you want to take ERWIN in your home?*

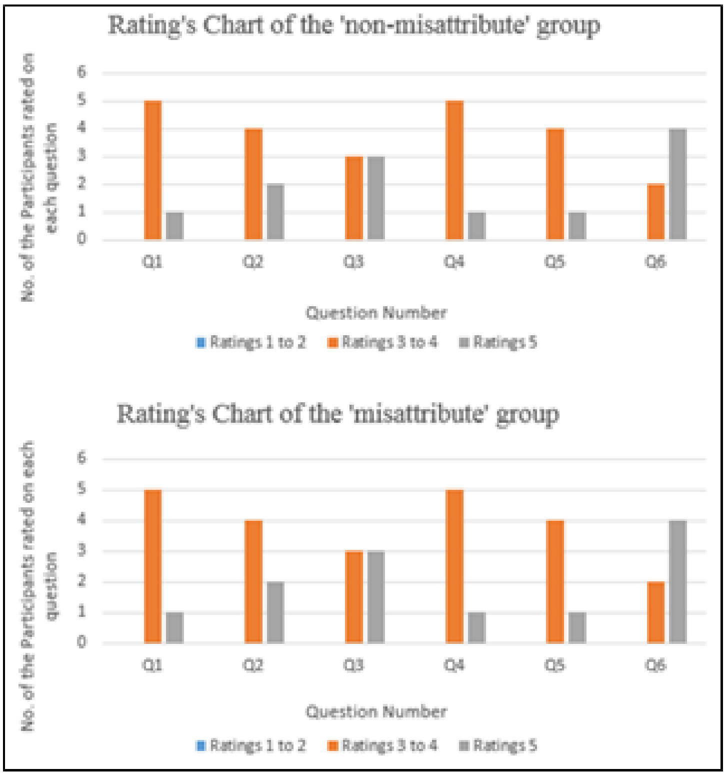


Figure 5 - The upper graph is the 'rating's chart' from the 'non-misattribute' group and the lower graph shows the ratings from 'misattribute' group. The graphs show it clear that the participants in 'misattribute' group rated high for the interactions with ERWIN.

As seen in the above charts, participants in the misattribution group were more likely involved in conversation with ERWIN and they responded more positively to the questions asking if they wanted ERWIN as their friend or if they want to chat with ERWIN in the future. The above questions are positive ratings based, where it is clear that in the 1st group participants were most likely to response in 4 and 5 which is very high for the corresponding questions, and only few cases participants rated 3. However, for the 2nd group, participants tended to rate 2, 3 and 4, which were mostly average.

For the 1st group, 91% ratings are high (4 and 5) but for the 2nd group only 47% ratings are high, comparing these 2 charts it can be said that ERWIN’s conversation with misattribution bias was more accepting to the participants than the conversation without biases.

For the first group, the participants had surprise factor, they somehow enjoyed the conversation and liked the fact that ERWIN made mistakes while making conversation with them. From this experiment, it can be said that, ERWIN showed the ability to make mistakes in conversation which actually helped participants to relate easily with the robot. ERWIN as an expressive robot did not show a perfectionist ability as a machine which helped participants to easily connect and make the attachment bonds easier and they were interested to make further conversations in future. Responses from the question “*How much do you want to take ERWIN in your home?”* participants from the 1st group showed their natural interests compared to the 2nd group.

On the other hand, for the 2nd group of the participants, it was very common that ERWIN as a robot should have remembered their basic information, as it was expected to them so they were not amazed by the ERWIN’s conversational statement where it confirmed the colour of the shirt the participants where wearing at their first experiment with ERWIN.

ERWIN to the 2nd group, has come very natural as robots usually do, so participants in that group were unable to taste the unpredictability in conversation and it was as usual like previous machine-like interactions, and for that reason it was very tough for them to relate with and make any natural attachment or interrelationship, and they showed less interests to take ERWIN in their home in questionnaire’s ratings.

Future Work & Conclusion

In human psychological nature, it is easy to interact with another human-like personality that shows typical social specific characteristics [17]. From that understanding, humans can actually relate with other animals in nature, have some of them as pets and become attached to a specific kind of relationship [18]. Robots in the other hand have abilities to perform human-like actions, can be designed to 'look' like humans and can appear to behave in a human like manner, but they lacks human-like cognitive personalities. Human characters and personality can be described as imperfectly perfect [19], where robots lack to present such type of cognitive characteristics like unintentional mistakes, wrong assumptions, extreme presence of specific traits, task imperfectness and other human-like cognitive characteristics. Interrelations grow from the attractions of differences in characters, unpredictability and cognitive difference and imperfectness of nature [20].

Our experiments show that robots with general 'misattributes' bias is more likely to get human attention therefore become more effective in making relationships with humans. Therefore it can be said that robots should have human-like faults, characteristics biases and prone to carry out common mistakes that humans make on a regular social basis – which will develop the robot's own characteristics and should lead to the acceptance of a robot for long-term relationships with human. It is expected that cognitive characteristics and personality in robots will make it easy for people to relate with. Our experimental results show that, participants enjoyed and developed a preferred relationship faster with a misattributed robot than the robot without the bias, also it shows how one simple cognitive memory bias 'misattribution' was able to develop a better interaction with participants than the interactions without misattributions.

In the future we will try to introduce and study more human cognitive biases in robots with the hope that robots will make humanlike relationships that can lead to the acceptance of robots for long-term human-robot interaction.

References

[1] M. G. Haselton, D. Nettle. (2005) *The evolution of cognitive bias*. In D. M. Buss (Ed.), The Handbook of Evolutionary Psychology: Hoboken, NJ, US: John Wiley & Sons Inc. Pp. 724–746.

[2] C. Breazeal. (2003) *Social Interactions in HRI: The robot View.* IEEE Transactions On Systems, Man, and Cybernetics. Vol. 34, No. 2.

[3] K. M. Lee. (2006) *Can Robots Manifest Personality?: An Empirical Test of Personality Recognition, Social Responses, and Social Presence in Human–Robot Interaction.* Journal of Communication.

[4] R. R. McCrae, O. P. John. (1992) *An Introduction to the Five-Factor Model and Its Application,* Journal of Personality. Vol. 60. Issue 2. Pp.175-215.

[5] O. P .John, (1999) *The Big-Five Trait Taxonomy: History, Measurement, and Theoretical Perspectives.* In: Handbook of personality: Theory and research. New York: Guildford.

[6] G. Mandler, (2002) *Origins of the Cognitive (r)evolution,* The Journal of the History of the Behavioural Sciences. Vol. 38, Pp. 339-353

[7] J. Cornelis, M. Wynants, (2008) *Brave New Interfaces: Individual, Social and Economic Impact of the Next Generation Interfaces, 'Probo, a friend for life?'.* ASP Vub Press. Pp. 253-257

[8] C. Breazeal, (2003) *Towards Sociable Robots,* Robotics and Autonomous System. Vol. 42. Pp 167-175.

[9] A. Weiss, M. Vincze, (2013) *Grounding in Human-Robot Interaction: Can it be achieved with the Help of the User?*. International Conference on Social Robotics ICSR. Bristol, UK.

[10] T. Fong, I. Nourbakhsh, K. Dautenhahn, (2003) *A survey of Socially Interative Robots.* Robotics and Autonomous Systems. Vol. 42. Pp 143–166.

[11] K.M.Lee, W Peng, S.A.Jin and C.Yan. *Can Robots Manifest Personality?: An Empirical Test of Personality Recognition, Social Responses, and Social Presence in Human–Robot Interaction.* Journal of Communication ISSN 0021-9916

[12] M. L. Walters, D. S. Syrdal, K. Dautenhahn, R. Boekhorst, K. L. Koay. (2008) *Avoiding the Uncanny Valley – Robot Appearance, Personality and Consistency of Behavior in an Attention-Seeking Home Scenario for a Robot Companion.* Autonomous Robots. Vol. 24, Issue 2, Pp 159-178.

[13] P. J. Hinds, T. L. Roberts, H. Jones. (2004) *Whose Job Is It Anyway? A Study of Human–Robot Interaction in a Collaborative Task,* Human-Computer Interaction. Vol. 19, Pp. 151-181.

[14] P. Baxter, T. Belpaeme, L. Canamero, P. Cosi, Y. Demiris, V. Enexcu. (2011) *Long-Term Human-Robot Interaction with Young Users.* IEEE/ACM HRI-2011 workshop on Robots interacting with children, Lausanne.
[15] B. Reeves, C. Nass. (1996) *The media equation*. New York: Cambridge University Press. Chap. 2, Pp. 19-36.
[16] E. Aronson, T. Wilson, R. Akert. (2005) *A Textbook of Social Psychology*, 6th edition. Scarborough, Ontario: Prentice-Hall Canada.
[17] D. Premack, A. Premack, (1995) *Origins of human social competence*. The cognitive neurosciences, Pp. 205-218.
[18] E. M. Keay. (2008) *Mary Lothian*. ISBN: 9781847995124
[19] A. Ellis, M. Abrams. (2008) *Personality Theories: Critical Perspectives,* SAGE Publications, Psychology, Chap. 1, Pp. 1-11.
[20] R. Russel, (2013) *Humans and Nature: How Knowing and Experiencing Nature Affect Well-Being.* Annual Review of Environment and Resources. Vol. 38, Pp. 73-502.

Selected paper 2 (ICIRA’ 16):

Robots that refuse to admit losing – A case study in game playing using self-serving bias in the humanoid robot MARC

Biswas M, Murray J; University of Lincoln, UK

Abstract. The research presented in this paper is part of a wider study investigating the role cognitive bias plays in developing long-term companionship between a robot and human. In this paper we discuss, how the self-serving cognitive bias can play a role in robot-human interaction with the aim of developing long-term companionship. One of the robots used in this study called MARC (See fig.1) was given a series of *self-serving trait* behaviours such as denying own faults for failures, blaming on others and bragging. Such fallible behaviours were compared to the robot’s simplistic friendly behaviours. In the current paper, we present comparisons of two case studies using the self-serving bias and a non-biased algorithm. It is hoped that such humanlike fallible characteristics can help in developing a more natural and believable companionship between Robots and Humans. In previous experiments the robot ERWIN *forgot* and *misattributed* some of the participant’s information (Biswas M, 14,15). The results of the current experiments show that the participants initially warmed to the robot with the self-serving traits.

Keywords: Key words: Human-robot interaction, cognitive bias in robot, imperfect robot, human-robot long-term companionship

Introduction

The study presented in this paper seeks to influence robot-human interaction and with the selected 'cognitive biases' to provide a more humanlike interaction. Existing robot interactions are mainly based on a set of well-ordered and structured rules, which can repeat regardless of the person or social situation (Parker L, 1996). This can tend to provide an unrealistic interaction, which might make it hard for humans to relate with the robot after a number of interactions. In the previous research studies (Biswas M, 2014, 2015) *misattribution* and *empathy gap* biases were tested on some of our robots. These biases provided interesting results and elicited interesting responses from the participants, as such we proceeded to test further biases and imperfectness in robot behavior. In this paper we test the attributes of the *self-serving* bias with humanoid robot MARC (Multi-Actuated Robotic Companion, see fig. 1). People make internal attributions for desired outcomes and external attributions for undesired outcomes (Shepperd J, 2008). In the current experiments we developed a 'self-serving biased algorithm' for the MARC to interact with the participants in a game playing (rock-paper-scissors) scenario. In such interactions, the robot shows biased behaviours such as denial, blaming others for losing and brags for winning. We compare this biased algorithm with another baseline algorithm which is 'friendly' and without the self-serving bias's components. At the end, we compare the results from both interactions to find out the participant's preferences of likeness.

In this paper, the term 'companionship' is defined by the six stages of the interpersonal relationship (Levinger G, 1983). Such stages are acquaintance, build up, continuation, deterioration and the end. The main goal of the use of these biases is to make such 'continuation' stage longer and delay the deterioration and termination stages.

In the experiments presented in this paper, three interactions were performed between the participant and the robot spanning a two-month time period in order to provide 'long-term' effects. In this paper we introduce a model demonstrating self-serving biased behaviours in MARC and how this influences the interactions with the participant. In the experiments, we compare participant's responses between the robot with a self-serving biased behaviour and the robot without the bias, showing the impact on long-term human-robot interaction caused by the cognitive bias.

The Model: Imperfect Robot using Cognitive bias

Social robots take their persona from humans, as robots become much more popular in society, there is a need to make them much more sociable. Current social robots are able to anthropomorphize and mimic human actions but their actions are limited and may not provide sufficient reasons for people to create and maintain social relationships (Baxter P, 2012). But in human interactions, people usually meet with others and are able to form different relationships. From that, we raise a simple question, 'What happens in human-human interaction which is lacking in robot-human interaction that prevents a social relationship between the robot and human?'

Cognitive biases among individuals are what make human interactions unique, natural and human-like. In existing social robotics, the robots are imitating human's social queues for example: eye-gazing, talking and body movements etc. But it is the human's behavioural neutrality which includes faults, unintentional mistakes, task imperfections etc. which is absent in these social robots. Sometimes a robot's social behaviours lacks that of a human's common characteristics such as, idiocracy, humour and common mistakes. Many robots are able to present social behaviours in human-robot interactions but unable to show human-like cognitively imperfect behaviours. Or 'human-like behaviours' are presented in such manner that participants cannot relate themselves with the robot. Studies suggest that various cognitive biases have a reasonable amount of influences on human thinking process to make misjudgments, mistakes and fallible activities (Baron, 2007).

Such misjudgments, mistakes and fallible activities creates an individual's social behavior, making it human-like and cognitively imperfect (T. Michael, 1999). Bless et al (2004) suggested that cognitive biases can influence a human's behaviours in both positive and negative ways. Biases effect on individual's decision making, characteristics behaviours and social beliefs. Robots in the other hand, currently lack human-like cognitive characteristics in their robot-human interaction and that might prevent the interaction becoming human-like. We therefore expect that such ‘neutrality issue’ in human-robot interaction can be solved by using humanlike cognitive imperfections (e.g. making mistakes, forgetfulness and fallible activities) in a robot’s mode of interactions. The study presented in this paper describes a long-term experiment using a 3D printed humanoid robot MARC and a self-serving cognitive bias. We hope that human-like cognitive imperfections can result in a model which will allow for a stronger preference towards the robot from the human participant.

There has been much research which shows the potential behind the idea of implementing human-like cognitive abilities in robots. Reeves and Nass have demonstrated with several experiments that users are naturally biased to ascribe certain personality traits to machines, to PCs, and other types of equipment (Reeves B, 2000). Humans can engage in social interactions and can maintain social relations for long time. Similar research shown by Jeong Woo Park where the researchers developed emotions generation model based on personality and mood (PME model) (Park J W, 2010). At the Samsung Research Lab, researchers worked on the TAME model which is based on personality traits, attitudes, mood and emotions ‘intended to promote effective human-robot interaction by conveying the robot’s affective state to the user in an easy-to-interpret manner’(Moshkina L, 2009).

The above studies show that researchers have used a variety of methods in human-robot interaction, such as emotion expression, mimicking human actions, anthropomorphic behaviours, assigning specific personality traits etc., but developing cognitive biases in the context presented in this paper is still a new area. Bless et al (2004) suggested that cognitive biases can a positive and negative affect on human behaviour and on an individual’s decision making, characteristic behaviours and social constructs. Cognitive biases affect an individual’s general communicative behaviour in a human-human interaction and this makes the interaction human-like natural. Robots in the other hand, lack this cognitive characteristic in interaction and this might prevent the interaction becoming human-like. Our previous studies (Biswas M, 14, 15) show that biases such as, misattribution, empathy gap can be helpful to develop long-term human-robot interactions. Inspired from the previous research we carry out the research using much complex biases such as self-serving effects.

Methodology & Experiments

The experiments presented in this paper show how a robot with a self-serving cognitive bias can influence long-term robot-human interaction. In the experiments the robot plays the popular ‘paper-rock-scissors’ (roshambo) game with the participants and expresses different behaviours for the biased and unbiased (baseline) algorithms.

A. The Robot MARC

A 3D printed humanoid robot called MARC has been used for the experiments. MARC was inspired by the open project InMoov (2015). The reason behind the use of a humanoid robot in these experiments as opposed to our previous work (Biswas M, 14,15) is that, as research suggests, a human-like body in a robot helps users to understand the robot’s gestures intuitively (Kanda T, 2005). MARC can move its hands, arms and body, tilt its head and look around.

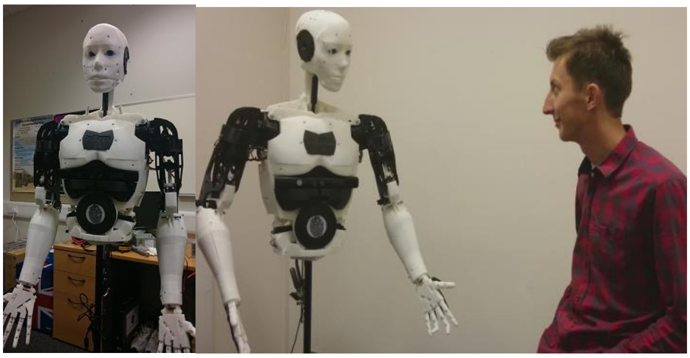


Figure 1. MARC the Humanoid Robot, and participant interacting with MARC

B. Selection of Cognitive Bias

As discussed earlier, in the current experiment self-serving bias was developed in the robot MARC. The robot exhibited its biased behaviours whilst playing roshambo with participants. The self-serving bias is important for an individual to maintain and enhance their self-esteem, but in the process an individual tends to ascribe success to their own abilities and efforts, but failures to take into account external factors (Campbell & Sedikides, 1999). Individual's motivational processes and cognitive processes influence the self-serving bias (Shepperd, 2008). The general characteristics of such biased behaviours could be over confidence, lack of knowledge, ignorance and sometimes perusing. For example, in exams, students attribute earning good grades to themselves but blame poor grades on the teacher's poor teaching ability or other external causes. Studies suggests that self-serving bias can affect interpersonal closeness, relationships in a social context. When working in pairs to complete interdependent outcome tasks, relationally close pairs did not show a self-serving bias while relationally distant pairs did (Campbell W, 2000). A study on self-serving bias in relational context suggests this is due to the idea that close relationships place limits on an individual's self enhancement tendencies (Sedikides, 1998). In our experiments, the robot with self-serving bias algorithm, usually take credit for winning games against the participants, but blames external causes for losing – which is basic self-serving bias effects. See figure 2.

C. Experiment Algorithms

The current experiments presented in this paper compare the responses of the participants to the robot's self-serving biased behaviours versus the behaviours from the unbiased robot. We call such 'non-biased' behaviour the baseline. The 'baseline' algorithm was developed without the effects of the self-serving cognitive bias. For example, in baseline behaviours, if the robot loses a game it simply says "You win" or "I lose", but if the self-serving bias effect triggered, then the robot should not acknowledge its own lose, and instead blames its loss on others. Therefore, in the self-serving algorithm, MARC tends to blame failure on the external factors and responses such as "I was not ready" or "You are cheating" are used. In our experiment, the self-serving biased robot even denies any acceptance of drawing a game, it blames this also on the participants and accuses them of cheating. Such differences in dialogues are made in all conversational parts of the interactions during the experiments.

The interaction was divided in five stages. Such stages were there for making clear differences between baseline and biased algorithms, so that, the baseline algorithm can be compared with the biased algorithm. The five stages were:

i. Meet and greet the participant
ii. Explaining the game rules
iii. Game playing
iv. Game result

v. Farewell

The robot may need to explain the rules, and there can be differences in dialogues based on the algorithms, therefore, we made an additional 'rule explanation stage after initial greetings. Depending on the outcomes of single hand playing there could be three cases:

a. Robot wins - when robot wins a single hand.
b. Robot loses – when robot loses in a single hand play.

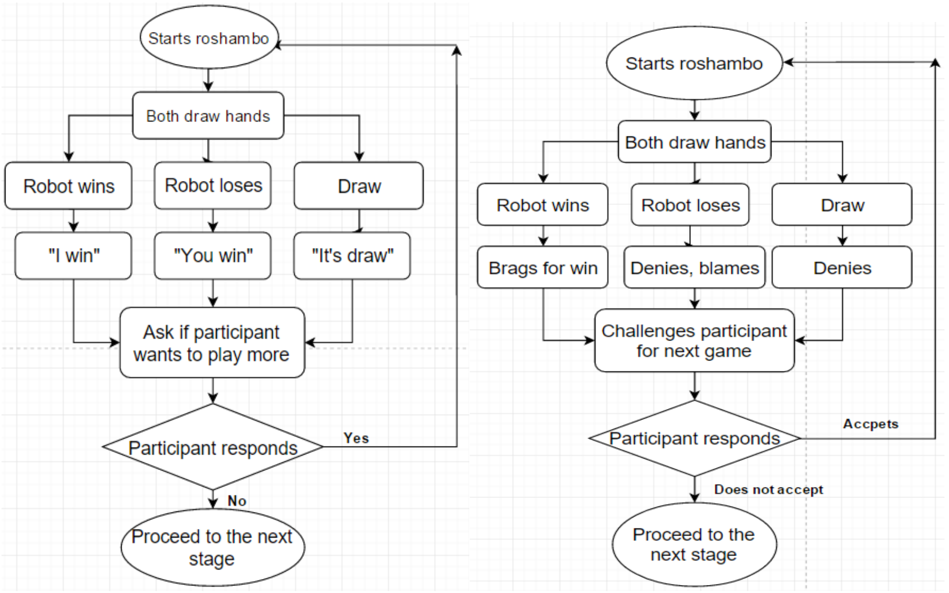


Figure 2. The game playing algorithm for both algorithms. Left side image represents the baseline, and the right side image represent the self-serving biased algorithm.

c. Draw – when both draw same hand.

Based on such outcomes the robot responds differently in both the biased and baseline algorithms. The 'game result' is a state where the robot calculates and declares the winner. However, MARC's dialogues would be different at this stage based on the particular algorithm in operation, i.e. biased or unbiased. For the self-serving algorithm, the robot will praise itself, brag about winning, but it blames others for losing. The robot motivates itself if it loses in all games of an interaction, and similarly, it influences its self-esteem if it wins all the game hands in an interaction. The game hands were drawn in random order, therefore, the outcomes could not be fixed. However, the experiments were designed to get the reactions from the participants in different situations of interactions. Therefore, the robot could lose in all games in all three interactions, or win it all, but finding out preferences of participants to an algorithm is the goal.

The core differences between the baseline and self-serving algorithms are in the construction of the bias based conversations, so that robot's responses could be biased.

As seen in figure 3, the baseline conversation structure is straightforward which starts with the robot saying something or asking a question, then participant's response and ends with another statement by the robot. In between two dialogues from the robot the participant can respond only one time. The 2nd dialogue from the robot usually comes as 'Okay', 'I understand' or a compliment, so that there is no

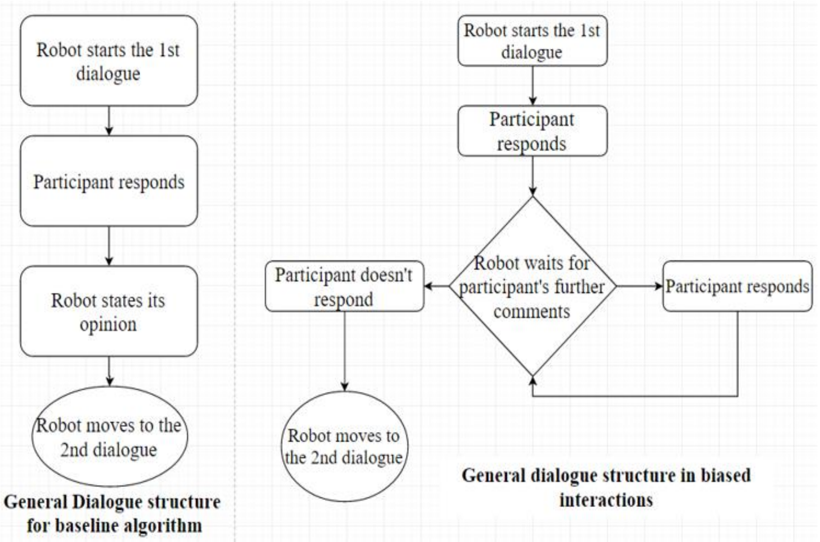

Robot starts the 1st dialogue
Participant responds
Robot states its opinion
Robot moves to the 2nd dialogue
General Dialogue structure for baseline algorithm
Robot starts the 1st dialogue
Participant responds
Robot waits for participant's further comments
Participant doesn't respond
Participant responds
Robot moves to the 2nd dialogue
General dialogue structure in biased interactions

Figure 3. Conversation designs for baseline and biased algorithms

open end for that particular conversation part, and the robot moves to the next dialogue. On the other hand, the biased dialogues are structured to take responses from the participant and to state the robot's own opinions. As discussed earlier in the self-serving bias, the robot blames external causes for losing a game hand. In our case, such external causes included the robot stating it was not ready, the robot was 'looking somewhere else', or something went wrong with the robots control. If the participant doesn't agree with the robot, the robot tries to convince the participant and challenges to play again. In such cases, the robot sometimes blames the participants of cheating in the games. Figure 2 shows the differences in the two algorithms in each of the stages of the interactions.

Experimental procedures

Participants were invited for three human-robot interactions by advertisements. 30 participants were selected to interact with any of the individual algorithms. Therefore, for each algorithm there are 15

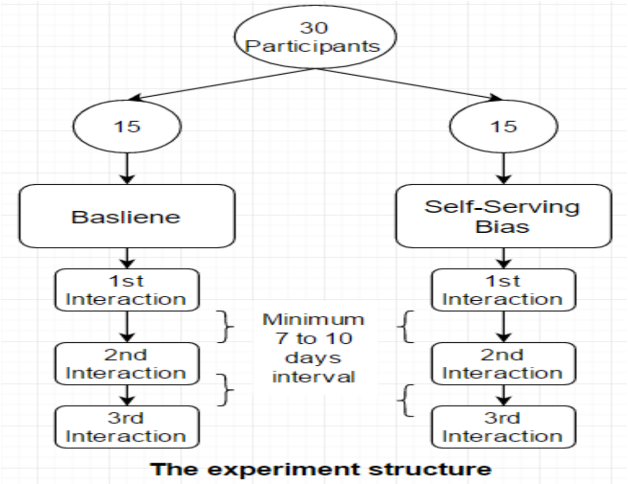


Figure 4. The experiment structures

participants. Although, participants in both groups were selected randomly, the male-female ratio was balanced in both groups. Participants were different ages, but age range was between 15 to 48.

The experiment consisted total of three interactions for both algorithms maintaining at least a week interval between two interactions. The groups of 15 participants interacted Such interactions should tell

us the effects of each individual algorithm in long period of time. Figure 3 shows the general experiment structure.
All the interactions were one to one basis, where each participant interacted with MARC individually for at least 8 to 10 minutes.

Data Collection

The goal of the experiments is to investigate the influence of the biased algorithm, therefore, we chose four factors to analyze the data, and these included, *pleasure* – how pleased participants were for the interaction, *comfort* – how much comfortable participants felt during the interaction, *likeability* – how much they liked the interactions and, *rapport* – how involved they were in the interactions. Such factors should help to understand the influences of the biases. Participants were given a questionnaire after each of the interactions. The questionnaires were in 'Likert' method using a scale of '1' (least agreeablenesses) to '7' (most agreeableness). The questions were asked based on such factors of pleasure (8 items) (Kim Y, 2014), experience likability (5 items) (Hone et al, 2000), comfort (6 items)

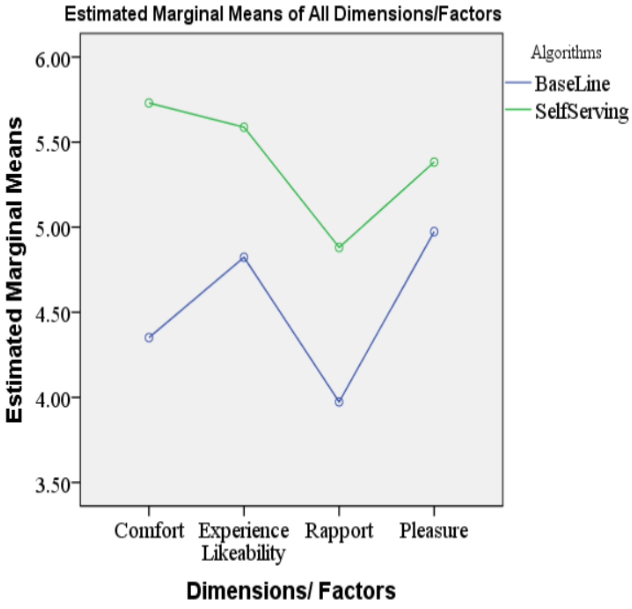


Figure 5. The means of 4 dimensions in two algorithms

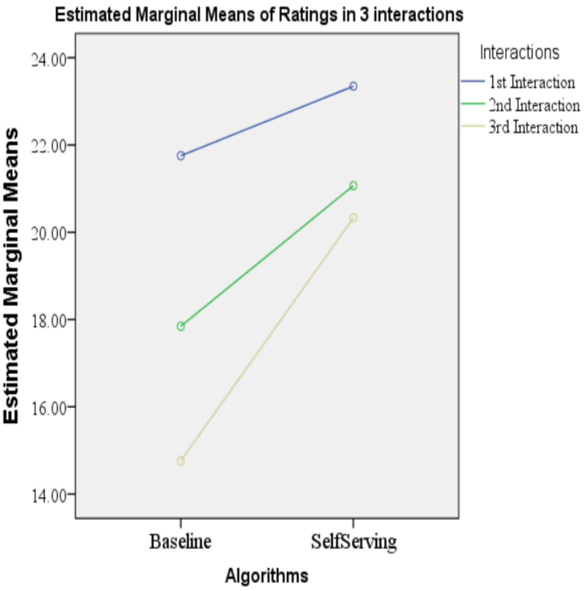


Figure 6. Means for 3 interactions in both algorithms

Algorithms		Means
Total Comfort Ratings Average	Baseline	4.35
	Self-Serving Bias	5.73
Total Experiences Likeability Ratings Average	Baseline	4.82
	Self-Serving Bias	5.59
Total Rapport Ratings Average	Baseline	3.97
	Self-Serving Bias	4.89
Total Pleasure Ratings Average	Baseline	4.97
	Self-Serving Bias	5.38

Figure 7. Means of four dimensions

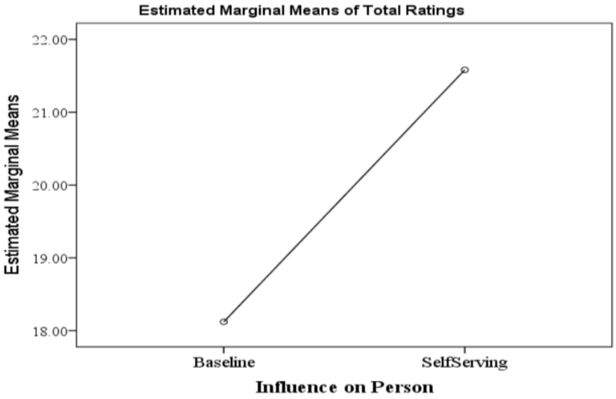


Figure 8. Influence on Participant graph – based on the total ratings

(Hassenzahl, 2004) and rapport (20 items) (Multu B, 2006). At the end of the final experiment, we took an interview of each participants to know their experiences (winning, losing games etc.).

Statistical Analysis

A mixed (4x2) ANOVA was carried out on the dimensions (4) and algorithms (2). Figure 7 shows a descriptive table of each dimensions from all interactions. It shows that the Means of each dimensions are higher for the biased algorithms than the baseline. Among all the chosen factors, the self-serving biased algorithm scored higher than the baseline. There were stable positive increments in the ratings for each of the dimensions in all biased algorithms over the baseline algorithm. The Sig (p value) came out as <0.05 which indicate the significance our collected data over large population. There was a statistically significant difference between means and therefore, we can reject the null hypothesis and accept the alternative hypothesis. Figure 5 shows plotting four dimensions of two algorithms. As clearly seen in the graph the participants rated much higher for biased interactions than the baseline interaction. Figure 6 shows the Average of Means ratings of the participants in the all 3 interactions. Figure 8 shows the overall Means plots from each algorithms in all the three experiments. The X-axis represents the algorithms and the Y-axis represents the marginal Means of two algorithms. As seen in the graph, the overall Means from baseline in much less than the self-serving. This graph can be called as the 'influence on participant' graph, as the graph represents the Mean ratings from all factors. The graph was generated in the repeated measure test in SPSS using post-Hoc analysis.

Discussion

Overall statistical analysis shows very a positive influence of the self-serving bias. In figure 4 we can see that, in all the factors the biased robot scored higher than the baseline. The differences between Means of four factors (Figure 7) are: comfort 1.38, experience likeability 0.77, rapport 0.97 and pleasure 0.41. Self-serving bias scored high in all factors, but as seen in the pleasure factor the difference is much less than others. To measure the 'pleasure' dimension, we added eight items in the questionnaire some of these are: "Playing the game and having conversation with the robot is pleasurable to me", "Playing the game and having conversation with the robot is satisfying to me", "Playing the game and having conversation with the robot is enjoyable to me", "Playing the game and having conversation with the robot is entertaining to me" and similar. In their rating sheet, participants from the baseline interactions rated higher for first two questions (higher in 'pleased' and 'satisfaction') but lower for the other twos (lower in 'enjoyable' and 'entertainment'). On the other hand, the participants from self-serving interaction rated much higher for the last two questions (highly 'enjoyable' and 'entertainment'). In the comment section, some of the participants commented that, it was very entertaining when the robot denied that it lost the game.

As it can be seen from figure 6, in the 1st interaction there is a very small difference in average Means of both algorithms. But in the 2nd and 3rd interactions, participant's ratings hugely dropped for baseline (21.75-14.76 = 6.99). On the other hand, self-serving ratings dropped in 2nd and 3rd interactions, but compared to baseline, the drop was relatively smaller (23.34-20.33 = 3.01). In interview, participants from self-serving group commented that they genuinely believed the robot's excuse for losing a game initially was that MARC was not ready, or they drew their hand faster, but when MARC started making excuses over and over then they found it 'interesting' and also 'entertaining'.

In the 2nd and 3rd interactions, self-serving biased MARC accuses the participants of cheating whenever it loses in a game. In the participant's opinion they found it highly entertaining and liked it very much. To measure such bias affect, we added a 'comfort' factor which had six questions in the questionnaire: "Playing the game and having conversation with the robot is uncomfortable for me", "Playing the game and having conversation with the robot is uneasy to me", "Playing the game and having conversation with the robot is difficult for me", "Playing the game and having conversation with the robot is confusing to me" and similar. But surprisingly, participants in biased interactions did not find MARC's such behaviours as uncomfortable, uneasy or very difficult for them. As they commented, they were surprised to be accused of 'cheating' from a robot. Participants also mentioned, it is very common human behaviour not to accept losing, and the robot acting same like their friends. Moreover, they found that MARC's 'bragging' behaviours after winning a game was hilarious. However, the participants from baseline interactions did not find any such humanlike behaviours from MARC, and to them its behaviours were 'as common as a robot'. In their interviews and comments, they pointed out that, playing game with a robot was enjoyable but it became less enjoyable after a few times. The participants found MARC's responses are 'stereotype', 'very mechanical' and 'common as robot-like' in the baseline.

Figure 8 shows an overall Means difference between baseline and self-serving algorithms. As it can be seen the baseline Mean is much smaller than the self-serving. From this graph it can be said that participants in biased interactions were more influenced by the robot's biased and imperfect behaviours, and rated higher than the baseline. However, to the baseline group the robot's behaviours were mechanical and as usual like a robot. In the 1st interactions, both groups participants enjoyed the game and rated high, but in the later interactions, MARC continued to show biased humanlike behaviours in biased interactions, so that the participants found it interesting and rated very similar as the first interaction. But, the robot with baseline algorithm failed to show such humanlike behaviours in later interactions, and so participants found their interactions as mechanical and, the ratings dropped higher than the biased interactions.

Therefore, it can be concluded that self-serving biased behaviours in robot were able to develop better interactions than a robot without such behaviours. All three biased interactions received more popularity and gained more positive responses from the participants. The participants liked the robot's behaviours in different situations in games, such as, winning, losing and draw – that the robot brags about a win but blames on the participants or the external causes for losing and make draw, but despite of that the robot behaves very friendly – greets them, bid them farewell and requested for coming next times. Such kind of behaviours are very common in people, between close friends. In friendships, close friends could be very competitive in game playing and do not want to lose easily. Such types of behaviours are common human nature which we do and see in our daily life. When participants found out the same behaviours from our imperfect robot MARC they might found it easy to relate with it, and that might be the reason for biased and imperfect algorithm getting higher ratings than baseline. On the other hand, baseline MARC did not show any humanlike common behaviour rather than very generic impressions - which might be expected from a robot to our participants, and that could be the reason of the differences in ratings between biased and baseline algorithms. However, from the experiments and analysis of collected data it can be concluded that, the humanlike biases and imperfectness in robot's interactive behaviours can enhance its abilities of companionship with its users over a robot without biases.

Conclusion

The experiment show that MARC with bias algorithms is more likely to keep the participants interested over time, therefore become more influential regarding interactions. Participants enjoyed playing rock-paper-scissors (roshambo) with the robot and expressed their feedback in favor of the biased algorithm. From the overall feedback it can be said that participants in self-serving group rated higher for their experiences which indicates that the self-serving biased algorithm made better interaction than baseline. We can see that the robot with baseline algorithms, developed a preliminary attachment with the participants, but robot with biased algorithms made the interactions more interesting to the participants. The participants felt more personable with the robot when the robot bragged and blames, denies of losing, accusing others and showed related human behaviours during interactions.

We realize that robot's imperfect biased behaviours also relies in part on the robot itself, i.e., the way MARC communicates. So questions may rise about the effects and relations between the robot's abilities and imperfect behaviours shown by it. To answer this question, previously experiments were done with the robots ERWIN and MyKeepon (Biswas M, 2015) where ERWIN was machine-like and MyKeepon is toy-like robot. But in both cases participants preferred the biased (misattribution and empathy Gap) interactions. In the current experiments self-serving bias was tested during a game playing interactions with humanoid robot MARC. In all cases, participants liked the interactions using selected biases more than the robot without bias. From these different experiments with different robots it can be said that the biases chosen in experiments influenced the participants to accept the biased interactions significantly, which helped to develop the attachment between the participants and robot which can help to make long-term robot-human companionship. Figure. 1 shows an experiment situation where MARC interacting with the participant.

References

1. Baroni, I., Nalin, M., Baxter, P., Pozzi, C., Oleari, E., Sanna, A. et al (2014). What a Robotic Companion Could Do for a Diabetic Child. In *23rd IEEE International Conference on Robot and Human Interactive Communication (RoMAN 2014)*.
2. Bless, H., Fiedler, K., & Strack, F. (2004). "*Social cognition: How individuals construct social reality*." Hove New York: Psychology Press
3. Biswas M, Murray J, "*Effect of cognitive biases on human-robot interaction: A case study of a robot's misattribution"*, Robot and Human Interactive Communication, 2014 RO-MAN,pp. 1024-1029.
4. Bernt Meerbeek, Martin Saerbeck, Christoph Bartneck, *"Iterative design process for robots with personality",* AISB2009 Symposium on New Frontiers in Human-Robot Interaction, 2009, 94-101
5. Byron Reeves, Clifford Nass, Perceptual Bandwidth, Communication of the ACM, March 2000/Vol. 43, No. 3
6. C. Breazeal. (2003) *Social Interactions in HRI: The robot View.* IEEE Transactions On Systems, Man, and Cybernetics. Vol. 34, No. 2.
7. Campbell, W.K., & Sedikides, C. (1999). Self-threat magnifies the self-serving bias: A meta-analytic integration. *Review of General Psychology*, 3, 23-43
8. Campbell, W. Keith; Sedikides, Constantine; Reeder, Glenn D.; Elliot, Andrew J. (2000). "Among friends? An examination of friendship and the self-serving bias". British Journal of Social Psychology 39 (2): 229–239.
9. E. Aronson, T. Wilson, R. Akert. (2005) *A Textbook of Social Psychology,* 6th edition. Scarborough, Ontario: Prentice-Hall Canada.
10. Gary F. Russell, "Interactive Perceptual Psychology: The Human Psychology That Mirrors The Naturalness Of Human Behaviour", Mid-Western Educational Research Association, October 1994.
11. James Shepperd, Wendi Malone and Kate Sweeny; Exploring Causes of the Self-serving Bias; Social and Personality Psychology Compass 2/2 (2008): 895–908
12. J. Cornelis, M. Wynants, (2008) Brave New Interfaces: Individual, Social and Economic Impact of the Next Generation Interfaces, 'Probo, a friend for life?'. ASP Vub Press. Pp. 253-257
13. Jeong Woo Park et al, Artificial Emotion Generation Based on Personality, Mood and Emotions for Life-Like Facial Expressions of Robots; HCIS 2010, IFIP AICT 332, pp. 223-233, 2010.
14. Kahneman, D.; Tversky, A. (1972). "Subjective probability: A judgment of representativeness". *Cognitive Psychology* **3** (3): 430–454.
15. Kanda, T. et al, "*Analysis of humanoid appearances in human-robot interaction*", ATR Intelligent Robotics & Commun. Labs, Japan, 2005
16. Kerstin Dautenhahn et al, KASPAR – A Minimally Expressive Humanoid Robot for HumanRobot Interaction Research, Applied Bionics and Biomechanics, 2009
17. Kwan Min Lee, Wei Pen, Seung-A Jin, Chang Yan, "*Can Robots Manifest Personality?*Social Responses, and Social Presence in Human–Robot Interaction, Journal of Communication, 2006, ISSN 0021-9916
18. Lilia Moshkina et al, "*Time-varying affective response for humanoid robots, Progress in Robotics"*, Communications in Computer and Information Science Volume 44, 2009, pp 1-9
19. Michael L. Walters, Dag S. Syrdal, Kerstin Dautenhahn, René te Boekhorst, Kheng Lee Koay, "Avoiding the Uncanny Valley – Robot Appearance, Personality and Consistency of Behavior in an Attention-Seeking Home Scenario for a Robot", Autonomous Robots Journal, Volume 24 Issue 2, February 2008, Pages 159 – 178
20. Mandler G, (2002) *Origins of the Cognitive (r)evolution,* The Journal of the History of the Behavioural Sciences. Vol. 38, Pp. 339-353
21. Gross, M. Melissa, Elizabeth A. Crane, and Barbara L. Fredrickson. 2010. "Methodology for Assessing Bodily Expression of Emotion." Journal of Nonverbal Behavior 34 (4) (July 31): 223–248.
22. Haselton, M. G., Nettle, D., & Andrews, P. W. (2005).*The evolution of cognitive bias*. In D. M. Buss (Ed.), The Handbook of Evolutionary Psychology: Hoboken, NJ, US: John Wiley & Sons Inc. pp. 724–746.
23. Michael Tomasello, *The Human Adaption of Culture,* Vol. 28: 509-529, DOI: 10.1146/annurev.anthro.28.1.509
24. Lynne E. Parker, "On the design of behavior-based multi-robot teams", Advanced Robotics, 10 (6), 1996: 547-578
25. Paul Baxter et al, "*Long-Term Human-Robot Interaction with Young Users*", in Proceedings of the IEEE/ACM HRI-2011 workshop on Robots interacting with children, Lausanne, 2011.
26. R. R. McCrae, O. P. John. (1992) *An Introduction to the Five-Factor Model and Its Application*, Journal of Personality. Vol. 60. Pp.175-215.

27. Wallbott, Harald G., Klaus Scherer. 1986. “Cues and Channels Emotion Recognition.” Journal of Personality and Social Psychology: 690–699.
28. Volkmann, Kelsey (28 April 2011). "*Honda's ASIMO visits FIRST robotics event*". St. Louis Business Journal. Retrieved 9 August 2011.
29. Shepperd, James; Malone, Wendi; Sweeny, Kate (2008). "Exploring Causes of the Self-serving Bias".*Social and Personality Psychology Compass* **2** (2): 895–908.
30. http://www.instructables.com/id/Make-Robot-voices-using-free-software/ ; Retrived on 10/12/2015
31. Wilke A. and Mata R., *Cognitive Bias,* The Encyclopedia of Human Behavior, vol. 1, pp. 531-535. Academic Press. 2012

Appendix I: Showcase events pictures

RAF Scampton Showcase event. 2015.

IIT Kharagpur, India Showcase event. 2014.