Institute of Psychology
Polish Academy of Sciences

Intentionality attribution to social robots

by

Katarzyna Zerka

Supervisors:

Rober Balas, Associate Professor

Grzegorz Pochwatko, PhD

Warsaw 2023

Acknowledgements

I am thankful to my supervisors, for supporting me on this journey. The greatest adventure with excellent company.

Many thanks to Lukas, my mum, brother and friends, for your love and support.

Abstract

Concerning human-technology interactions, much attention is being paid to evaluating how we are going to be perceived by AI. Little attention has been paid to how we react to it and how our biases influence interactions. This is particularly relevant for social robots, as they could be perceived as active social players. Examining intentionality bias in the perception of social robots in this line of research contributes to filling this gap. The pilot study aimed to adapt an intentionality bias paradigm to research human–robot interactions. The goal of the first study was to compare the intentionality bias towards humans and robots with low or high levels of anthropomorphic features. In addition, a time pressure condition was introduced as a between-subjects factor for bias enhancement. The goal of the second study was to evaluate whether priming could alter the perception of robots and influence intentionality bias. Priming in the opposite direction of the anthropomorphic features of the robot was introduced as a between-subjects factor, expanding the precedence from the first study. Considering ambiguous accidental behaviours, intentionality bias is present for robots at the same level or higher than humans, suggesting a mental model of the robot as a deliberate, programmed machine. This mental model will influence the user experience of a new class of social robots. Priming should be investigated further as a means to influence it. This research contributes to the design of human-robot interactions that should allow for the control of perceptual modes triggered in users.

Keywords: intentionality bias, intention attribution, intentional stance, Human-Robot Interaction, social robots

Contents

Introduction

Our life has become so intertwined with technology that it is hard to think of any day where it does not play a vital role. Looking at the list of the most valuable brands (Forbes Magazine, n.d.) on our planet, the first five positions are occupied by technology-focused companies. Geek culture has found its way to pop culture, and engineers have become new role models for success in life. However, it is not only the complex technical landscape that has been changing. Our relationship with technology has been evolving, and the impact on our society has been more profound. Robots are like no other technological creation. They inspire such a contradictory mixture of excitement, admiration, and fear. Whether we favour this development or not, robots are becoming not only utilitarian machines but social actors as well. Robots have started appearing at airports and shopping malls, and they have found their way into our homes. Some of them are specifically designed for the emotional responses of humans, like those to comfort patients or residents of nursing homes (Lee et al., 2019), and the recent COVID-19 pandemic has accelerated the use of robots for contactless services.

The rapid development of robots has spawned its own field, as it brings a new level of complexity and a truly eclectic challenge. This dissertation introduces the concept of a robot and social robots and describes recent applications. The current picture of human-robot interaction (HRI) design practice is presented, its roots in academia and current influences from the industry.

In the field of HRI, much attention is still being paid to evaluating the technological boundaries and expanding the capabilities of our robots. However, we can not only ask the question of what technology has to offer us but, more importantly, it can be asked how we are currently reacting to it and how we may react to it in the future. There has been a pronounced shift from technology-oriented development to the user-centred design of technology-based

products. One of the most fascinating topics in social robotics concerns the possibility that a robot can be perceived as having a mind, and/or being an active social player. There has been a growing interest in this perspective, employing a psychological approach (e.g., Bossi et al., 2020; Marchesi et al., 2021; Pochwatko et al., 2015).

This thesis explores the topic of attributing intentions to robots, its philosophical background, plausibility, and current research approaches. Reviewing the recent program from one of the most important conferences in the field, Human-Robot Interaction, we can see that the topic of intentionality perception has started to be discussed. For example, the recent workshops 'Social Cognition for HRI: Exploring the relationship between mindreading and social attunement in human-robot interaction' (Pérez-Osorio et al., 2020) or 'Human vs Humanoid. A behavioural investigation of the individual tendency to adopt the intentional stance (Marchesi et al., 2021) addressed precisely the topic of how intentionality can be ascribed to machines that interact with humans.

As this thesis focuses on the intentionality perception of technology, the concept of anthropomorphism is employed because it involves the attribution of human-like characteristics, such as emotions and inner mental states (e.g., motivations and intentions), to animals and non-living things (Epley et al., 2008). Anthropomorphism provides a sense of understanding of a non-human agent (Waytz et al., 2010), where the knowledge of human social interactions plays the role of a heuristic. The presented research line explores characteristics of a robot that makes it more likely to be perceived as anthropomorphic, as well as individual differences of people potentially influencing this relation, like empathy, the tendency to anthropomorphise, attitudes towards robots, or beliefs in human nature uniqueness.

Finally, priming as a method of altering the reactions to robots is introduced. There are signals that priming with more human-like robots will affect subsequent interactions with a less human-like robot (Castelli, 2002), which is a technique that may be utilized by human-robot interaction designers to shape the experiences of robot users.

The main research chapter addresses the question of whether intentionality bias is specific to the perceptions of humans or whether it is present in the perceptions of social robots' behaviours. In addition, it examines whether the level of anthropomorphic features of a robot affects the level of intentionality attribution and whether the tendency to anthropomorphize, empathy, negative attitudes towards robots, and a belief in human nature uniqueness have effects on intentionality perceptions. Moreover, this project examines the possibility of influencing a robot's intentionality perception by employing priming. Results of a pilot study and two experiments are presented and discussed.

The presented work is an attempt to adapt protocols from experimental psychology to HRI research, as a psychological perspective focusing on human tendencies can play a vital role in explaining and influencing our future interactions with social technology.

Human-robot interactions

In the words of social robotics pioneers, 'the design of social robot technologies and methodologies are informed by robotics, artificial intelligence, psychology, neuroscience, human factors, design, anthropology, and more' (Breazeal et al., 2016, p. 1935). Human-robot interaction (HRI) is a relatively new, eclectic, and fast-growing area of research. It is rooted in Human-Computer Interaction and adopting design and evaluation techniques coming from current User Experience (UX) practices (Dautenhahn, 2007). As the presence of social robots in everyday life is growing, there is a growing need for a constant evaluation of our relationships

with technology, and conscious and user-centric design is becoming more pronounced. This thesis aims to contribute to the field of designing deliberate Human-Robot Interaction (HRI). This chapter covers the context and definition of a social robot, the roots of the HRI discipline, and the domains of practice that are closely related to it.

What is a social robot?

The history of robots as we know them today started in 1920 with a Czech writer, Karel Čapek, and his play called *Rossum's Universal Robots.* The word robot comes from Czech, which stands for forced labour. It was invented by Čapek's brother Josef, a painter and a writer himself. In the play, robots were mass-produced workers created from artificially synthesized organic material. The plot gave a start to the dystopian imagination of the future with the first robot uprising (Wilson, 2015).

The first robot was an industrial machine and was installed in a Swedish metalworks plant in 1959. It was a two-tone arm controlled by a program on a magnetic drum, and it operated between a few pre-set angles with precision. By the 1970s, there were 3,000 industrial robots in operation, and 800,000 in 2003. In 2015, more than 1.3 million industrial robots were used in various manufacturing industries, including automotive, electronics, rubber and plastics, cosmetics, pharmaceutical, and food and beverage. Their market value at that time was $9.5 billion (Wilson, 2015). It was not until the 1980s that robots found space outside of factories and laboratories. This is when Honda launched its humanoid robotics program, developing the P3, which could walk and shake your hand. Its successor, Asimo, became known for playing soccer with Barack Obama, the president of the United States at the time. The robotics industry has taken a huge leap in recent years, mainly due to developments in related fields, namely sensors, actuators, and artificial intelligence (AI; Simon, 2018). Sensors have boosted the capability of

robots to comprehend the environment, preventing them from crashing into things and allowing them to act with precision. Actuators are responsible for smoother movement that can withstand a lot of pressure. These components can be thought of as joints for robots that have enabled the dancing robot videos that have become popular in the last few years and have allowed the robot Atlas from Boston Dynamics to do backflips. What is really compelling is the robots becoming smarter, not only because of algorithms and the use cases we give to them but because they can learn on their own, figuring out the rules of their operations. As most people might be exposed to social robots in the coming years, psychological studies looking into this topic are gaining importance.

A robot is ‘an actuated mechanism programmable in two or more axes with a degree of autonomy moving within its environment to perform intended tasks’ (International Organization for Standardization [ISO], 2014). In addition, it has some range of mobility and an interface of some sort to interact with us. Autonomy ISO (2014) defines it as ‘the ability to perform intended tasks based on current state and sensing, without human intervention.’

There are different types of robots and many different frameworks trying to classify them. Onnasch and Roesler (2021) suggest a taxonomy that enables a systematic comparison of different robots and their effectiveness. Summarising, there are three variables, each having a few stages: task specification (Information exchange, Precision, Physical load reduction, Transport, Manipulation, Cognitive stimulation, Emotional stimulation, Physical stimulation), robot morphology (anthropomorphic, zoomorphic, technical) and degree of robot autonomy (information acquisition, information analysis, action selection, and action implementation; for each stage, the level of robot autonomy can vary from low/none to high/complete).

The everyday usage of the word robot may bring confusion about what a robot is and what it

is not. A laptop is not a robot, as it is not ‘programmable in two or more axes’ (International Organization for Standardization [ISO], 2014), so it cannot make changes in the physical environment. It also does not have enough autonomy, it needs a human to trigger every action. Popular chatbots are not robots either. Even though they include Artificial intelligence algorithms, software is not embodied. Therefore, it is not a part of robotics (Owen-Hill, n.d.).

For less complex taxonomy, we can distinguish industrial, service, and personal robots. The International Federation of Robotics defines an industrial robot following the ISO definition as ‘an automatically controlled, reprogrammable, multipurpose manipulator programmable in three or more axes, which may be either fixed in place or mobile for use in industrial automation applications.’ The same norm specifies a service robot as one that ‘performs useful tasks for humans or equipment excluding industrial automation applications’. A personal service robot is a machine that is mostly used by a layperson, excluding medical use cases (International Organization for Standardization [ISO], 2014). Social robots, which are the topic of study in this dissertation, represent the latter category.

As outlined by Darling (2012), social robots are defined as autonomous, embodied beings that are able to interact and communicate with people. This ability to interact with humans is the main feature of these robots, they should be able to express emotions, carry on a simple conversation, comprehend the basic mental model of the interaction partner, and learn with experience. Mainly, they are being designed as humanoids or similar to pets. Robots can also be considered social if they interact with other robots. Indeed, in-group interactions between robots have been modelled after insect societies, which are anonymous, homogeneous groups where an individual does not matter. This type of social structure has proven to be useful in robotics as it allows for complex operations to be performed with relatively simple individuals, like playing

soccer or moving objects (Fong et al., 2003). This thesis focuses on robots designed to interact individually with people and therefore have to fit human society. Fong et al. (2003) called the 'individual-social' robots the socially interactive ones to highlight the focus on their place in our society. According to the authors, these types of robots exhibit human-social characteristics, where they 'express and/or perceive emotions, communicate with high-level dialogue, learn/recognize models of other agents, establish/maintain social relationships, use natural cues (gaze, gestures, etc.), exhibit distinctive personality and character, may learn/develop social competencies.' The authors also refer to different levels of being social, from socially situated robots that can recognize and react to social surroundings, through socially embedded robots that can partially comprehend humans, to socially intelligent machines building deep models of human cognition and social competencies.

Robots capable of complex social interactions carry enormous potential in many fields of application. In 2016, it was estimated that 30 million personal assistive robots would be sold between 2015 and 2018, including around 1.5 million robots that could be categorized as social (KPMG, 2016, p. 5). Social robots are mainly being purchased to help at home; for example, Jibo, which 'helps adults to manage their life, helps seniors to live with greater independence and is a playmate for children' (KPMG, 2016, p. 26). Due to the rising popularity of social robots, it is crucial to constantly evaluate the dynamics of their relationships with people (Ogunyale et al., 2018). Such an evaluation requires an understanding of how individuals conceptualize these robots and their social behaviour, especially if people interpret their behaviour with intention, desire, and opinion attribution (Malle & Hodges, 2006). Research on the human perceptions of robots can uncover mental models and cognitive processes triggered during the interactions (De Graaf & Malle, 2018). This knowledge can then be successfully applied in UX, social

psychology, cognitive science, ethics, and law. The importance of researching this topic grows with the scale of social robot applications.

Robots were envisaged to assist humans in tasks that are hazardous, repetitive, or prone to errors (Takayama et al., 2008). Giving up ‘dirty, dull, and dangerous’ working conditions promote healthy outcomes for people and environmentally friendly processes by increasing production effectiveness. But robots have also found their spots in entertainment, healthcare, and education, where they interact with humans in a social context; thus, how we perceive them becomes crucial.

For the benefit of our health and well-being, robots work in the production of medicines and medical devices, disinfect hospitals, fetch and carry linens and medication in hospitals, and help elderly people live independently for longer. Robots have also proven tremendously helpful for people with physical disabilities, increasing their mobility and helping with everyday life (International Federation of Robotics, 2021). A good example of this is Aibo, the robotic pet dog. Its primary use is to be a companion and to maintain a lifelike appearance (Fujita, 2001). Social robots are used in various therapeutic interventions, for example, robots have been used to help kids with autism spectrum disorder by enabling the practice of joint attention, emotional understanding, or turn-taking (Cabibihan et al., 2013; Dautenhahn, 2007). Some robots have been designed to improve patients' mood, which has many health benefits. An example of this type of robot is Paro, the huggable robot seal. Paro is an alternative to animal-assisted therapy used in elderly care facilities to reduce loneliness (Birks et al., 2016).

As for the robots that are designed to teach through social interactions, they are as effective at improving cognitive and affective outcomes as a human tutor (Belpaeme et al., 2018), which shows robots’ advantages over virtual learning technologies. E.g., when compared with

instructions from virtual characters, videos of robots, or audio-only lessons, instructions from physical robots have produced more rapid learning in cognitive puzzles (Leyzberg et al., 2012). Robots can deal with learning material that requires a physical presence, like teaching handwriting; they can be also more engaging and enjoyable in cooperative tasks than a virtual agent (Belpaeme et al., 2018).

Today, social robots are easily found outside of the clinical and teaching contexts, including in the workplace, serving at conferences, and working at supermarkets or airports (Wiese et al., 2017). Entertainment is also a very interesting use case, as it allows for developing new products that can learn and develop interactions without putting too much risk in mission-critical situations. All of these situations require at least a moderate level of social engagement, where intentionality perception seems to be important.

The COVID-19 outbreak could have been an accelerator of these changes. Indeed, the recent pandemic has made contactless services flourishing, and robots have aided this process. Throughout times of social distancing, robots have been used for cleaning and sanitizing surfaces, scanning for fevers, and delivering food and medical supplies. However, using robots to fight pathogens is nothing new, and they have been utilized in the past to combat viruses. For example, the robot Saul was used to help mitigate the Ebola outbreak. This particular robot was utilized to sanitize hospital rooms after operations by emitting pulses of high-intensity, high-energy ultraviolet rays to neutralize pathogens (Air Force Magazine, 2014). Robots have also proven useful on the front lines of the COVID-19 pandemic in hospitals. They have helped healthcare workers to focus on crucial tasks and to save on personal protective equipment (Graham, 2021). Robots can support rehabilitation therapy, and are especially useful for this purpose during pandemics (Céspedes et al., 2021). Robots have also been useful for stocking

store shelves and delivering essential items to people who are quarantined at home. Tally, a robot developed by Simbe Robotics, has been used in grocery stores to scan shelves and identify out-of-stock, misplaced, or mispriced goods. Tally is able to check 15,000 to 30,000 products per hour (Bandoim, 2020). There are also examples of the successful implementation of robots for grocery delivery (Marr, 2020). In China, Meituan Dianping, a delivery app, expanded its 'contactless delivery' options through the use of autonomous vehicles and robots. Another China-based start-up, Pudu Technology from Shenzhen, implemented the home delivery of drugs and meals via robots.

We are at the early stages of this journey, and most people do not yet have exposure to social robots. As this may soon change, psychological studies looking into this topic are gaining in importance. How people react to social robots varies, and is currently being researched, including what conditions can trigger or enhance these social interactions. But as noted by Wykowska et al. (2014), robot actions can elicit perceptual effects of the same kind as human actions. Therefore, psychology has a lot to contribute to the field of Human-Robot Interactions.

Designing Human-Robot Interaction

Human-robot interaction (HRI) design is rooted in Human-Computer Interaction (HCI) and follows design and evaluation techniques coming from current User Experience (UX) practice (Dautenhahn, 2007).

HCI is a field of research that followed the massive growth of computer science in the 20th century. The main concern of the field is adapting technology to human needs. Why has it become so important? Today, the number of digital products and services, including robots, is growing, and these technologies have become ubiquitous. Humans are not fully equipped to face this reality. Evolution has prepared us to thrive in a physical and social environment, and we had

to adapt these lenses to deal with technology. We treat our interactions with machines and digital products in the same way that we think about our interactions with the environment and fellow humans (Reeves & Nass, 1996). Research on computers as social actors has shown that people often treat computer interfaces as human (Reeves & Nass, 1996). For example, Fogg and Nass (1997) observed the rule of reciprocity when participants were interacting with a computer. Machines that helped with performing a task received more help in return from participants in a subsequent task.

Traditionally, the HCI research field is connected to computer science and ergonomics, usability engineering, and information systems (Hassenzahl, 2008). The discipline has a rich research body and a grounding in academia, it has been practised in the industry and has earned wide recognition. The field of computer-human interactions (CHI) refers to a narrower area of research that is even more closely connected to computer science than HCI. The main platforms for this field are the Association for Computing Machinery Special Interest Group on Computer-Human Interaction (ACM SIGCHI) and the annual CHI conference. Since 2006, an event has been organized dedicated explicitly to interactions with robots (HRI, 2006). The psychological perspective has been present and growing during those events, in 2022 all three keynote speakers had a background in psychology (e.g., Eyssel, 2022).

The term User Experience (UX) was adopted by the creative and technology industries and is mostly related to researching and designing digital and interactive technological products. UX encompasses all aspects of the interactions of a company user, its products, and services, write Norman and Nielsen (*What Is the Secret of Don Norman's Success?*, n.d.), to whom we can attribute laying the foundations for this new discipline. An ISO norm defines UX as the whole spectrum of reactions and perceptions of the usage or the anticipated usage of a product, service,

or system (International Organization for Standardization [ISO], 2014). Mościchowska and Rogoś-Turek (2015) point to three attributes of a product that are crucial for UX: usability (functionality, ergonomics, ease of use), attractiveness (brand image and visual aspects), and evoking positive emotions.

UX started as a job title for the designer Don Norman (*What Is the Secret of Don Norman's Success?*, n.d.), who worked at Apple in the late 1990s as a UX architect. This was at a time when the company started designing the experience of owning a piece of technology more holistically. This holistic approach embraces UX as a discipline of understanding and positively influencing human life experiences. The specialists practising UX have focused on an individual user perspective to solve problems and create value. It is a valuable approach but not the only one that should be taken. The perspective of the natural or social environment are another two layers that should be considered to centre our design thinking for a sustainable future. For the presented topic, the perspective of the social environment of a piece of technology is crucial.

Design thinking is an approach that is tightly connected to innovations and designing interactions for digital, interactive products. Both UX and design thinking are rooted in academia but have been transferred to industry, where they were developed and popularized. The symbolic start of UX can be traced to Apple in the 1990s. Since then, the discipline has been dynamically developing in Silicon Valley, whereas Design thinking as a framework was developed a bit later. From the beginning, design thinking as a concept was comprehended more broadly than UX and the audience was larger, spanning outside of start-ups and the technology domain. In this case, commercial use cases pushed the development of design thinking further (Knemeyer, 2015). The design thinking method originally comes from a 1969 text by Herbert A. Simon, entitled *The Sciences of the Artificial* (Costa, 2019). Since then, the model has been transformed many times

and has been improved. Current trends and inspirations can be found at The Hasso Plattner Institute of Design at Stanford. Design thinking is considered both the idea and the process of solving problems in a human-centred fashion. There is no doubt that design thinking shares its roots and approaches to building products and services with UX. Even if the design process is not following either of these approaches, it is still a human-centred design as long as people, communities, and environmental needs are kept at the centre. Design thinking can be seen as a framework that can be adopted and used by the UX specialist to solve broad, undefined, and complex problems. Product design in robotics has been facing exactly the type of challenges mentioned above.

Promoting design techniques in the field of robotics is critical, as there are not many examples in the literature following established human-centred frameworks. However, one has been provided by McGinn et al. (2019), which aims to develop a socially assistive service robot that empowers vulnerable members of society in the spirit of 'design thinking philosophy.' A high-resolution prototype was designed, fabricated, and evaluated by research using a mixed-methods approach. User sessions were conducted with four distinct groups of residents and care staff at a retirement community centre. The team evaluated the first impressions of the robotic system, its anticipated usefulness, and its acceptance. The response was generally positive, and the more detailed findings were incorporated into the future product design.

One way to look at HRI design is to distinguish 'biologically inspired' robots from functional ones. In the first approach, designers try to mimic social intelligence based on living creatures. This approach is rooted in the natural or social sciences, such as anthropology, cognitive science, developmental psychology, ethology, sociology, and theory of mind. In this case, the design task is to translate current scientific knowledge to robotic systems, their

behaviours, motivations, and cognition. ‘Functional design’, on the other hand, does not aim to draw from scientific theories and is focused on achieving intelligent impressions. This approach makes it sufficient to follow folk understandings of some processes without deeply comprehending them. Functional design in robotics is where authors place HCI design and other techniques and frameworks known from today’s UX practices (Fong et al., 2003).

Another view on designing HRI is to follow two main approaches. The bottom-up approach aims to achieve a believable agent, and the effect is less important than the appearance details. This type of design relies on utilizing a combination of elements from the human face and body and behaviours (e.g., gaze pattern, tone of voice, facial expressions, gestures) that are crucial for human social interactions, e.g. people show preference for robots who follow our gaze (Willemse et al., 2018). Examples of robots following this design pattern are Probo, Kismet, and MDS. The top-down approach is aimed at producing an autonomous replica of a human, visible in robots like Repliée Q2 and Actroid DER. Replicating human interaction is the end in itself, whereas the bottom-up approach is more a utilitarian way to enhance HRI (Giger et al., 2019).

This thesis aims at evaluating robot’s perception based on its appearance, thus the style of the body is crucial. The literature on robot embodiment is vast. Fong et al. (2003) describe four main design directions that robot creators can take. These are anthropomorphic, zoomorphic, caricatured, and functional designs. Anthropomorphic designs display recognizable human-like features like eyes, ears, and similar body shapes. Humanoids fall into this category. Zoomorphic agents have animal-like features in their appearance, such as four legs or tails. Caricatured-designed robots do not have realistic, bio-inspired features. They follow stereotypical representations, and their appearance can distract or attract attention to certain machine features. Finally, functional robots are fully task-dependent.

Putting a familiar shape on a robot makes it easier to start interacting with it. For example, when we see a robot dog, we have ready scripts to try to interact with it. A warmly welcomed robot toy was a baby dinosaur known as Pleo (IEEE, 2018). From the UX perspective, it was a fascinating design, as no one exactly knows how to interact with a small dinosaur; thus, we approach the robot with few expectations and no ready scripts for joint activities. It makes the interaction more interesting and allows for teaching users new patterns, but it is also risky, as the interaction is more difficult to start without knowing what to expect. Mental models are responsible for forming our expectations. These models are based on beliefs regarding what users know or think they know about a system. This allows us to predict a system's behaviuor and plan future actions (Nielsen, 2010). Users form a mental model of an agent based on its appearance, its behaviours, and their preferences. Returning to the example of a dog-like robot, users would intuitively interact with it differently than, for example, a cat-like robot. This is important as the first impression is based on design elements, like the face, height, gender, or speech expressions, and users then make inferences about a robot's intelligence, competence, warmth, or friendliness, and, most importantly, what kind of tasks the robot can perform (Komatsu et al., 2011). Matsumoto et al. (2005) introduced a 'minimal design policy' as a framework for designing robotic agents. The policy stated that anthropomorphic robots should be designed to reduce possible expectation gaps and ensure that users do not amplify or underestimate the robot's competency level. The design also gives cues to treat a robot as a social actor, attributing human-like states to it, having a mind and intentions.

The degree of human likeness of robots can play a vital role in our willingness to adopt this technology. One of the most well-known effects related to our anthropomorphism of robots has been termed the uncanny valley (Mori, 1970). This phenomenon refers to the observation that the

more human-like the robot's appearance, the more positive emotion it elicits in humans. Interestingly, Mori (1970) hypothesized that a person's reaction to a human-like robot would abruptly shift from empathy to revulsion if the robot's appearance too closely resembles that of a human. These unsettling emotions are thought to have an evolutionary origin, namely that they are a defence mechanism used to avoid people not holding up to our norms (Moosa & Ud-Dean, 2010). While there are no clear data regarding the psychological basis of this phenomenon (e.g., Rosenthal-von der Putten & Kramer, 2014, as cited in Rosenthal-von der Pütten et al., 2019), primates also show this aversion. Steckenfinger and Ghazanfar (2009) presented monkeys with unrealistic and realistic artificial monkey faces, as well as real monkey faces, and examined preference by measuring the time spent looking at each picture. The longest time was spent looking at the real and artificial unrealistic faces, with the least amount of time spent looking at the realistic synthetic faces. The authors attributed these results to the uncanny valley effect.

Future scenarios

One of the most important aspects to consider is the humanization of social machines, which is a topic that induces controversy and attention and benefits the market. Giger et al. (2019) see three ways forward. One would aim to replicate our intelligence and social skills, creating machines as versatile as humans. However, this will not occur in the foreseeable future, and is still in the realm of science fiction fantasies. The second way would be to only add certain anthropomorphic characteristics for the benefit of interactions, but not to approach human likeness. The third way is to aim for full human-like features but to stop short of the uncanny valley threshold.

We can also look at this topic from a different angle. There are undoubtedly advantages to designing machines inspired by ourselves. This may not only benefit our interactions with

robots, but can also contribute to their functionality. If we want social robots to thrive in our environment, such as in cities or homes, making them similar to us makes a lot of sense. On the other hand, what current technology is best at is different from our strengths. For example, a simple calculator seems unbeatable at calculus for any human. Machines can work without rest, lift heavy loads, and identify explosive materials or patterns in data that are challenging for us to see. In contrast, our strengths will remain unbeatable by machines for many years to come. A simple conversation can derail the smartest algorithm, although it has become increasingly difficult with recently developed LLM applications, such us ChatGPT. Therefore, technology is supplemental and should be designed this way. 'Our goal shouldn't be to recreate human intelligence and skill in the far-away future. Why recreate what we already have, when we can purposely create something new?' (Darling, 2020).

There is also the view that humanizing social machines is a very natural way to go. Hanson (2021) argues that human nature is the most beautiful development in the natural world, 'Humans are brilliant, beautiful, compassionate, loveable, and capable of love, so why shouldn't we aspire to make robots humanlike in these ways? Don't we want robots to have such marvellous capabilities as love, compassion, and genius?' He argues for striving to give social robots the qualities we value in ourselves. One benefit that could result from this is good science because we can learn a lot and develop in engineering, cognitive sciences, and even biology, as robotics can serve as a meta-level of it. It will also push the fields of AI, speech recognition, and navigating tasks like grasping and manipulating fragile objects. In Hanson's (2021) words, social robots 'cross-pollinate among the sciences, and represent a subject of scientific inquiry themselves'. The other point is that humanizing robots will produce good companions that are just as useful as us in many human-specific situations. Lastly, it may be vital to build creatures

that we can understand, that we are able to influence, and that are willing to adopt our values and have compassion towards us, becoming our friends in the far-fetched future when they become conscious or have general adaptive intelligence.

Another important topic of the future with robots is the stereotypes that machines can reinforce, as they are designed by people who are unconsciously biased or fed with data based on our biased behaviours. Our perception of their intentionality can also leverage the effects of reinforced biases. As reported by Wilson et al. (2019), state-of-the-art object detection systems that might be used, for example, in autonomous cars, exhibit poorer performance in detecting pedestrians with darker skin tones. The authors of this study argue that neither the time of the day nor occlusion can explain this behaviour. A racial bias has also been reported for commercial facial recognition application programming interfaces (APIs; Wang et al., 2019). Not only do social robots 'inherit' existing biases from humans. We can also be biased towards robots, conveying prejudiced characteristics for human social life and affecting HRIs, which is the topic examined in this thesis. Robot design aspects, their appearances, behaviour, and even the surrounding marketing communications could affect society. Given the increasing popularity of social robots, now is the time to prepare for these potential effects.

What will profoundly affect the user experience of social machines is whether we will perceive robots as intentional in their behaviours. The following chapter discusses how we attribute intentions to fellow humans and whether this is possible in interactions with advanced technology as well.

Intentional interpretation

People distinguish between intentional and unintentional actions when explaining behaviours of others and are substantially consistent with these judgments (Malle & Knobe,

1997). Adopting an intentional stance is a strategy taken in order to predict the behaviour of a system or a human (Dennett, 1987). This strategy is different from the theory of mind, which is defined as the capacity to understand the mental states of other people in a given situation, and is based on explaining and predicting their behaviour. Theory of mind is needed to adopt an intentional stance, but it is not the same construct, is used in different contexts, and is operationalized in a different way (Perez-Osorio & Wykowska, 2020).

Developmental psychology states that a child develops an understanding of another's behaviour that overrides default interpretations. According to this approach, this ability becomes more sophisticated as a child grows up (Rosset & Rottman, 2014). Children are able to pass explicit theory of mind tests around the age of four years, but research has indicated that they show signs of intentional stance adoption much earlier. It is likely that infants do not fully adopt an intentional stance, but can comprehend mental operations and use this strategy to predict others' behaviour. According to Tomasello et al. (2005), children's skills of shared intentionality develop gradually during the first 14 months of life. Research on primates has also confirmed that it is possible for animals to predict the consequences of human actions, without having a mental representation of them (Perez-Osorio & Wykowska, 2020).

As we grow up immersed and are constantly trained in understanding and assessing the mental states of others and our states, the intentional stance becomes our default option. We use this default option when starting a new interaction and continue to use this prediction strategy if it turns out to be effective in a given situation. Rosset and Rottman (2014) compiled a convincing body of research showing that actions are understood as intentional at a very early stage of human life. They proposed a framework suggesting that the moment we acquire this ability, it becomes our default mode of interpretation. What we shape during the course of our

development are the skills used to detect accidental or non-voluntarybbehaviours. For example, a child below the age of five may have difficulty understanding that sneezing is not an intended action. This development is likely due to the strengthening of executive processing – mainly in inhibitory control – which allows for the overriding of initial impulses. Adults are capable of using this properly but can underuse it in situations of high cognitive load (Rosset & Rottman, 2014).

Intentionality attribution triggers

Movement can be a strong cue for intentional interpretation. In a classic experiment by Heider and Simmel (1944), the participants were presented with a series of short videos. In each video, three geometrical shapes were moving around an empty rectangle. Observers attributed personality traits to the shapes and explained the movements in terms of intentions and emotions. It was obvious to them which shape was bad, which was good, and what the goal of each shape was. The large rectangle was referred to as a home or a shelter, and the shapes were described with emotionally loaded words, such as aggressive or compassionate.

A good example of reducing other cues but movement comes from a series of experiments by Johansson (1973, as cited in Blakemore & Decety, 2001). In this study, the stimulus was a recording of an actor moving in a dark environment with lights attached to the main joints of the body. The participants could easily recognize a human and could determine gender. Other research on biological motion confirms that people can recognize not only the biological nature of movement based on moving dots but also the mover's personality traits and emotions (Blakemore & Decety, 2001). According to the researchers, the attribution to movement is a lower level of the theory of mind skills that is a prerequisite to developing the full potential of this ability.

In social interactions, the phenomenon of motor interference (MI) also plays an important role. MI refers to the finding that observing a different (incongruent) movement by another individual leads to a higher variance in one's movement trajectory (Kupferberg et al., 2011). Therefore, MI is a consequence of the tendency to imitate the movement of other individuals in order to enhance mutual rapport, a sense of togetherness, and sympathy. Although MI occurs while observing a human agent, it does not happen when observing an industrial robot moving with a constant velocity. Kupferberg et al. (2011) compared the effects of different types of movement (biological vs. mechanical) and agent types (industrial robots, humanoid robots, or humans), and it was observed that the humanoid robot produced MI, but not the industrial arm. However, when the industrial robot arm closely reproduced biological motion qualities, it triggered MI. Thus, the authors of the latter study concluded that an artificial biological-like movement velocity profile is sufficient to facilitate the perception of anthropomorphic robots as interaction partners (Kupferberg et al., 2011).

Currently, we know that not only movement, but the look of the eyes, mouth, or hands, can also trigger intentionality attribution (Perez-Osorio & Wykowska, 2020). The cerebral cortex in and near the superior temporal sulcus (STS) region is an important component of the perceptual system involved in analysing biological motion. The STS region is activated by eye, mouth, hand, and body movements in humans and monkeys. This structure is also activated by static images of the face and body, suggesting sensitivity to implied motion or signals from the actions of others (Allison et al., 2000).

Considering people's ability to identify personal qualities and intentions from minimal movement cues, it is crucial to design and continuously evaluate a robot's physical appearance and movement and gestures to enhance interactions with human partners (Giger et al., 2019).

Intentionality bias

Intention attribution is one of the essential elements of human interactions. In ambiguous situations, we tend to interpret the situation as intentional. This intentionality bias can be explained by the dual processing model (Rosset, 2008), which states that intention attribution is our default interpretation mode. Cognitive effort is needed to overwrite the default mode to perceive human behaviour as accidental.

The influence of cognitive load on the use of intentional interpretation has been demonstrated in a series of studies where time pressure (Rosset, 2008) or alcohol consumption (Bègue et al., 2010) was shown to increase the bias. Underlying the concept of intentionality bias is the assumption that a mature understanding of behaviour – which we acquire throughout development from a child to an adult – depends not on differences in intentional interference but on differences in intentional inhibition (Rosset, 2008). According to this claim, if asked whether the act of 'setting a house on fire' is deliberate or accidental, our initial reaction is that it is intentional. This automatic intention attribution, however, may subsequently be inhibited by additional information or knowledge that we have accumulated about humans and our environments, such as human nature (fallibility, forgetting something), social norms (it is an undesired behaviour), and environmental properties (flammability of the materials). According to the intentionality bias, every action is judged to be intentional until proven otherwise (Rosset, 2008).

To validate the intentionality bias, Rosset (2008) designed three studies. In each of these, the participants read sentences describing behaviours that could be done either on purpose or by accident. Both the control and experimental groups judged whether the behaviours were done on purpose or by accident. The experimental conditions introduced situations enhancing the

likelihood of biases (see Kahneman, 2011). Findings across the three studies suggested that adults have a bias to infer intention in behaviours usually assessed as accidental.

If our perceptions of anthropomorphic technology are similar to our perceptions of other humans, then biases may appear. As the impact of biases and cognitive heuristics on our social interactions with fellow humans is tremendous (see Kahneman, 2011), these are also likely to influence our interactions with social machines and AI.

Intentional robots

Although some neuro research has not found evidence that the intentionality perception is being triggered during HRIs (Chaminade et al., 2012), it is agreed that intentionality is a feature that can be attributed by us to nonhuman agents (Gray et al., 2007; Marchesi et al., 2019). Research has shown that the behaviour of robots can invoke a perceptual effect similar to that triggered by the behaviour of fellow humans (Thellman et al., 2017; Wykowska et al., 2014b). Individual biases toward treating robots as either intentional agents or mechanistic artifacts can be observed at the neural level in a resting state EEG signal. The human brain exhibits resting-state activity patterns that allow for the prediction of biases in attitudes toward robots (Bossi et al., 2020). However, even if we do not believe that robots have a mind, desires, or motivations, we adopt the strategy as if it were true in order to make our relationships with them effective. However, even if we do not believe that robots have a mind, desires, or motivation, we adopt the strategy as if it were true in order to make our relationships with them effective.

When we interact with agents for which we lack specific knowledge, we choose the ‘human’ model to predict their behaviour, as this is by training our default mode (Rosset, 2008). Once the model is activated, we use it to infer what was behind the observed action. We assess whether this model makes the interaction more effective and either keeps it or readjusts it by choosing

one of the other modes of understanding and predicting our environment. According to Rosset (2008), the default mode is intention attribution; however, it may be inhibited by additional information, and the behaviour judged as accidental or happening because something was designed that way.

According to Dennett (1987), we can perceive technology, as well as other beings and systems, using different strategies for predicting its behaviour. There are three main strategies, including the physical stance, design stance and intentional stance. The physical stance allows us to produce predictions based on our knowledge of physics and the basic stable rules governing our world. For example, we know we will accelerate on a bicycle going down a hill. However, the physical strategy is less effective in the case of a complex system. Thinking about a bicycle again, without knowing exactly how the brakes work or what the differences are between disc brakes and drum brakes, we know that we must pull a lever to slow down. This is the design stance. The design stance is often applicable to explaining and predicting a robot's behaviour (Perez-Osorio & Wykowska, 2020). We anticipate a robot's reactions based on our knowledge of how it was built. The intentional stance, on the other hand, comes to us spontaneously and effortlessly in a social context, as this skill has been developed, extensively practised, and has proven useful in day-to-day life with other humans. Perez-Osorio and Wykowska (2020) concluded that people seem to have a natural tendency to explain and describe others' behaviour in terms of mental states. Although we tend to adopt a mentalistic stance, there may be levels or degrees to which this occurs. The first level can be observed in our use of language, in which we use metaphors to make communication more efficient. For example, we portray inanimate objects and phenomena as having emotions and intentions (e.g., the sea is angry today). The second level applies to animate agents that do not possess human-like characteristics, but we do

interact with them and predict their behaviour as if this was the case. A good example would be cats and dogs. We could easily explain their behaviour from the design stance perspective, so shaped by evolution, but it is more comforting and attractive to attribute human-like states to them. The authors also refer to a level in between that involves our reactions to characters in movies, plays, books, and other media. We are well aware that fictional characters do not experience the story we witness, but following it from an intentional stance makes it enjoyable for us. The last level is a fully-fledged one, where we attribute intentions and, based on that, predict what is going to happen. As an example, we can think of city traffic and realize that it takes us a fraction of time, often without focusing attention on it, to go through this process.

We can also view technology from the intentional stance, thus, assuming that the agent or system has a mind in order to predict its behaviour. When using this strategy in interactions with technology, we anticipate as if it had a mind, but the belief that the mind is really there is of less importance. This instrumentalist approach is similar to the Turing test, where we do not assess whether the machine is truly intelligent, but rather evaluate whether and when people attribute intelligence to its behaviour (Turing, 1950, as cited in Gray et al., 2007).

We use the intentional strategy because it makes our interactions effective and also because evolution did not equip us to deal with new kinds of agents (Reeves & Nass, 1996). It seems more plausible if we think about robots with human-like bodies and faces, but it does even happen for personal computers that are nowhere near a humanoid shape. A car seems more humanoid than a computer, as it has headlights for eyes, a hood line for a mouth, and turn signals for facial expressions. Computers do not express emotions or refer to themselves as ‘I’. However, Reeves and Nass (1996) found that we tend to overuse human social categories, such as gender and ethnicity, and attribute them to computers. They also provided evidence that

people engage in politeness and reciprocity toward computers, which are learned social patterns, and expect better performance from a piece of technology, such as a TV, if it is specialized for specific content. This occurred even though not a single participant in the experiments declared that computers should be understood in a social manner (Nass & Moon, 2000). Based on these studies, the authors proposed a theory called the media equation, which encapsulates our automatic social responses to personal computers and other media that we are not naturally equipped to deal with. The computers are social actors (CASA) framework, derived from media equation theory, explains how people interact with media with this social potential. CASA suggests that humans mindlessly apply social scripts to interactions with media, although this specific explanation has been challenged (Gambino et al., 2020).

Due to the characteristics of the personal computer, anthropomorphism could be ruled out as a mechanism playing a vital role in this interaction. One of the plausible reasons is that computers are filling roles traditionally filled by humans. Hence, the cues that encourage social responses and the treatment of computers and other assistive technologies as social are ubiquitous (Nass & Moon, 2000). The notion that a small set of cues can automatically elicit social interaction scripts has been impactful, for both HCI and HRI research. It has paved the way for an interest in humanizing the virtual and physical robotic agents that we start to experience today.

Intentionality in human-robot interactions

When we take an intentional stance towards others, we refer to their mental states in order to predict their behaviours (Dennett, 1987). Mentalistic perception has many advantages in HRIs, including activating social brain areas connected to mentalizing, enhancing bonding and empathy, facilitating shared tasks, and producing more emotional responses (Wiese et al., 2017).

A human-like appearance and behaviour impact the quality of interactions and UX; thus, making robots more likely to be accepted (Duffy, 2003), perceived as more pleasant (Axelrod & Hone, 2005), and that people empathize more strongly with them (Riek et al., 2009). A human-like appearance also makes functional sense, as a robot's human size and movement abilities simply fit the environment created for us. It is easier for engineers to design versatile and flexible use-case machines that resemble us and are able to perform tasks in the urban landscape adjusted to our needs, as well as inside our homes.

An intentional strategy seems a rational and effective way to deal with intelligent systems. And any system that can be effectively predicted based on this approach can be viewed as truly intentional (Dennett, 2009, as cited in Perez-Osorio & Wykowska, 2020). However, there are also negative aspects of attributing intentions to robots. For example, perceiving an artificial agent in this way raises expectations that the machine cannot currently fulfil, which can negatively affect the UX. Sometimes the design stance can be a more effective way of predicting the behaviours of a complex system. To illustrate this, Wykowska et al. (2016) give the example of autonomous cars. In this case, the driver anticipates that the car will break when encountering an obstacle because that is the way it was programmed. The mentalistic view, which ascribes intention to the performed action of braking, would be less accurate and effective in this situation.

Robots do have the potential to trigger mind attributions. This intentional perspective assumes there is a mind behind the behaviour, or at least we predict the behaviour as if a mind was there. There are some visible clues that are needed to trigger this strategy, such as a robot's behaviour or anthropomorphic embodiment.

Related concepts

Anthropomorphism

Our perception of animals and inanimate creatures, including social robots, is affected by our tendency to anthropomorphize. Anthropomorphism is the attribution of human-like characteristics, such as appearance, emotions, and inner mental states (e.g., motivations and intentions), to animals, non-living things, and natural phenomena. It can be applied to either real or imagined nonhuman agents (Epley et al., 2008). Anthropomorphism is a common tendency with a high level of individual differences. When we think for a moment about the diversity of beings surrounding us in nature and culture (e.g., religion), all the forces associated with these beings, and the rapidly growing digital layer of our reality, it is overwhelming. We have had to learn how to deal with the environment and make it, so to speak, more familiar. This is why we picture God to have grey hair, we talk to animals, and we stroke a device to make it work. Measuring this tendency can help with understanding people's reactions to humanoid robots (Epley et al., 2007).

Anthropomorphism and intention attribution

As discussed earlier, the attribution of intentions, conscious experience, and metacognition are central to anthropomorphism (Epley et al., 2007; Gray et al., 2007), but are not exhaustive factors. Anthropomorphism also entails attributing human-like emotional states, behavioural traits, or human-like forms to abstract agents (Waytz et al., 2010). Anthropomorphism can therefore be operationalized on a research level as a particular form of mental state attribution.

Anthropomorphism is a process of inference about the unobservable characteristics of a nonhuman agent rather than their descriptive reports (Epley et al., 2007). The process is an example of induction where people reason about an unknown property based on a well-known

related representation – in this case themselves and other humans. For this to occur, we need to acquire knowledge, retrieve it, and apply it at the time of judgment (Higgins, 1996, as cited in Epley et al., 2008). We focus on the time of judgment, as anthropomorphism is primarily a tool that allows us to interact with the environment, not a description of the world (Damiano & Dumouchel, 2018).

Motivation to anthropomorphize

Although anthropomorphism has an obvious folk psychology element, there have not been many attempts to scientifically explain why, and under what conditions, people follow this perceptual tendency. One of the theories to explain this tendency is the SEEK model developed by Epley et al. (2007), which involves three determinants: sociality, effectance, and elicited agent knowledge.

Sociality encompasses the need to establish connections and social bonds with others. Anthropomorphism fosters the satisfaction of this need by turning a nonhuman agent into an agent having the perceived ability to connect with us in a human-like manner. This mechanism predicts that anthropomorphism as a tendency will increase when people lack social connections to other humans and will decrease when people experience a strong sense of social connection. Effectance refers to the need to interact effectively with one's environment (White, 1959, as cited in Epley et al., 2008). The concept of one's own egocentric experience is likely to serve as a useful knowledge structure when reasoning about our environment, including for nonhuman agents. In other words, our egocentric experience works as a heuristic. The use of this heuristic depends on our motivation to understand, control, and effectively interact with the environment. Effectance motivation is boosted by higher incentives for competence, such as the desire for control or predictability and the likelihood of future interactions. Anthropomorphism, therefore,

provides a sense of understanding and control of a nonhuman agent. Waytz, Morewedge, et al. (2010) explored the hypothesis that anthropomorphism occurs in part to satisfy the effectance motivation. The results showed that participants were more likely to attribute a mind to gadgets described as unpredictable than to those described as predictable, which is in line with 'the mastery' of the environment motivation. The third one, elicited agent knowledge, is related to the information available to us at the time of the interaction. We cover the lack of data about an agent with inferences from our human-to-human experience. On the other hand, anthropomorphism should play a lesser role if we are equipped with rich information about an agent. The importance of accessible knowledge relates to priming as a method of experimental manipulation, which is examined in this research line.

Psychologically, anthropomorphism has been considered an invariant and automatic process that is simply a persistent feature of human judgment (Guthrie, 1993, as cited in Epley et al., 2007). At the very least, it provides a source of testable hypotheses to guide our behaviour towards an unknown agent or stimulus in our environment. Having induction at its core, anthropomorphism works through a process of starting with highly accessible knowledge structures as an anchor point that can be an inductive base that may be applied to a nonhuman target and corrected, if necessary (Epley et al., 2007).

Drawing from the work on anthropomorphism, how people perceive social robots depends on the same mechanisms involved when people think about other people. Using one's own mental states and characteristics as a guide for reasoning about other humans is known as egocentrism. Using one's own mental states and characteristics as a guide for reasoning about nonhuman agents is anthropomorphism (Epley et al., 2008).

Anthropomorphism and mental state attribution to robots

One might ask a question, is the anthropomorphisation of robots connected to how we attribute mind to them? While testing the effectance motivation (the need to interact effectively with one's environment, being a master of the environment), on our perception of unpredictable gadgets, Waytz et al. (2010) examined the neural correlates of anthropomorphism. The results showed that evaluating the mental capacity of unpredictable gadgets was associated with relative increases in activity in the ventral medial prefrontal cortex (vmPFC) and anterior cingulate cortex measured with fMRI. These regions are known to be involved in socio-cognitive processes, such as thinking about similar others. Cullen et al. (2013) also examined brain structures related to answers on a self-report questionnaire measuring the level of anthropomorphic attributions about nonhuman animals and nonanimal stimuli using magnetic resonance imaging in 83 young adults. It was found that individual differences in anthropomorphism for nonhuman animals correlated with grey matter volume in the left temporoparietal junction, a brain area involved in mentalizing. These data support a link between areas of the brain involved in anthropomorphism and attributing mental states to other humans. Further evidence for locating anthropomorphic processing within the mentalizing network comes from autistic individuals. These individuals have difficulties attributing mental states to other humans and display decreased activation in areas of the brain known to be involved in the theory of mind processes (Castelli, 2002). These results suggest that autistic individuals have an impaired ability to anthropomorphize as well. Indeed, this has been shown in many studies. For example, autistic individuals gave less anthropomorphic descriptions of forms presented in a movie (Heider & Simmel, 1944; see also Cullen et al., 2013).

Epley et al. (2007) refer to strong and weak anthropomorphism. The strong form involves behaving as if the object has the attributed human-like traits and an explicit endorsement of these beliefs. The weak form entails 'as if' reasoning, more similar to metaphorical thinking. The authors suggest that even the weak form of anthropomorphism may have a profound influence on our behaviour. This prediction is similar to Dennett's (1987) intentional stance, where the true belief that a nonhuman agent is conscious is not crucial. It is perfectly enough to behave 'as if' to make our relationship with the environment effective and successful.

What makes robots anthropomorphic?

The literature focuses on two factors that trigger the anthropomorphizing of robots: appearance and movement (or behaviour; Damiano & Dumouchel, 2018; Levillain & Zibetti, 2017). Empirical evidence suggests (Urquiza-Haas & Kotrschal, 2015) that realistic behaviour dominates human likeness in activating anthropomorphic projections, even when the appearance is missing. When an object displays autonomous coordination with human movement, we perceive it as intentional almost instantly. Can a robot with a human-like appearance promote anthropomorphism even when behaviour is not presented? With strong realism, even one of these vectors, the appearance or behaviour, can pass a threshold of social perception, after which people experience the robot as a social agent. Passive viewing of a static stimulus can induce the activation of brain regions associated with social cognition (Wheatley et al., 2011).

Empathy

Empathy is one of the most important mechanisms that allow us to navigate the world of social interactions. The ability to empathize is common in all cultural backgrounds, but people differ in their tendency to empathise, and this tendency is a relatively stable personality trait (Leiberg & Anders, 2006).

The term empathy has been challenging to define and operationalize in a consistent way that encompasses the multidimensionality of this process. A working definition should exclude behavioural aspects, such as sympathy. In line with Reniers et al. (2011) we adopt the distinction between cognitive and emotional empathy. Cognitive empathy requires that visual, auditory, or situational cues about others are held in mind and manipulated to represent another person's cognitive and emotional states (Reniers et al., 2011). This process of representation can occur at the explicit level, but it can also appear as a meta-representation. In contrast, emotional empathy concerns vicariously experiencing the emotional experiences of others. The key here is sensitivity to and experience of the other person's feelings, and not awareness or understanding.

Cognitive empathy involves an understanding of other people's experiences, but also an understanding of others' emotions, which is different from the theory of mind. Theory of mind is commonly described as 'the ability to attribute mental states to others' (Vollmet et al., 2006, p. 90, as cited in Reniers et al., 2011). Theory of mind-related skills most likely overlap with those related to cognitive empathy, and they can impair empathy if not functioning properly. However, both are conceptualized as distinct phenomena.

Research shows that empathy is related to anthropomorphism (Harrison & Hall, 2010), which is particularly interesting in the current research. Anthropomorphism seems to be an effect of the ability to draw from one's own intentions, emotions and convictions and apply knowledge of those experiences to understanding the mental states of others (Gallup, 1985, as cited in Harrison & Hall, 2010). Therefore, anthropomorphism can be an integral part of this general ability to comprehend the inner states of other beings. Also, cognitive, but not emotional empathy, allows for significant predictions of the intentional interpretation of others' behaviours

in situations prototypically perceived as accidental or ambiguous (Slavny & Moore, 2018). It seems justified to examine these processes with regard to social robots.

Attitudes towards robots

An attitude is a fairly stable tendency to evaluate an object or behaviour in a positive or negative manner. Attitudes are intensively researched in psychology, as they allow, to some degree, for predictions of human behaviour (Wojciszke, 2004). Allport (1935) characterizes them as a mental and neural state of readiness. Attitudes are among the most distinctive and indispensable concepts in social psychology. Attitudes have cognitive components, such as knowledge or beliefs, affective components, including emotions and feelings, as well as behavioural components, which we can see as dispositions to act in a certain way. The connection with behaviour is one of the reasons there have been so many studies on attitudes; however, the strength of the relationship to behaviour varies and depends on many individual and situational factors (Marcinkowski & Reid, 2019).

As technology becomes more prevalent and touches almost every aspect of our lives, it is important and insightful to study attitudes towards new solutions. For example, negative attributes can prevent people from adopting useful technological solutions. Nomura et al. (2005) were some of the first researchers to use the concept of an attitude to understand human interactions with robots. For example, in experiments using the robot Robovie, participants with higher negative attitudes toward interaction with robots tend to take longer to talk to robots, also they tend to avoid self-expression to robots (Nomura et al., 2005).

Attitudes toward robots are not immune to cross-cultural differences (Bartneck et al., 2005). For example, visible differences can be seen between European and Asian societies, especially in Japanese society where exposure to social robotics in real life and in pop culture has occurred

over a longer period of time and is much more prevalent (Mozaryn et al., 2016). Exposure to robots and using them, as well as having a degree in technology or engineering, predicts higher trust toward robots and IA (Oksanen et al., 2020). Attitudes toward robots can help to predict human behaviour towards them, such as the willingness to collaborate, which is the main use of these machines.

Nomura et al., (2005) created a tool to measure this concept, the Negative Attitudes towards Robots Scale (NARS). This scale focuses on potential reactions to robots and allows for the identification of characteristics that facilitate reluctance. Various other scales measure different aspects of human-robot interaction, for example the Godspeed Questionnaire measures five key concepts in HRI: anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety (Bartneck et al., 2008), or the Robot Anxiety Scale assesses predominantly anxiety (Nomura et al., 2006). But it is NARS that has been the most used scale for measuring general attitudes (Koverola et al., 2022).

Belief in the uniqueness of human nature

Medin and Ortony (1989) propose the term `psychological essentialism' to refer to the belief that many categories have essences and share fundamental similarities. This system of beliefs, that there are natural kinds sharing a common essence, which defines what they are, is called 'essentialism' (Haslam & Whelan, 2008). Essentialism is connected with prejudices and stereotypes, justifications for social inequalities and emphasizes between-group differences (Demoulin et al., 2006). Although it should not be monolithically associated with prejudice, but as an emphasis on certain aspects of social categorization (Haslam et al., 2000). Essentialism also plays a role in understanding of what it means to be a human and the level of humanness we are ready to attribute to others (Giger et al., 2017).

The system of beliefs suggesting that human nature is unique and that people differ in qualities from other beings can also influence attitudes toward robots. People who believe in the uniqueness of human nature are more likely to hold stronger negative attitudes toward anthropomorphic robots and are less likely to accept them as a part of our society (Mozaryn et al., 2016). It is also likely to influence attributions of intentions to robots.

Influencing the perception of robots

Priming is a method of using one stimulus to alter the perception of subsequently presented stimuli. This prior exposure activates knowledge structures, causing the entire net of associations with the first stimulus to become more available. For example, the activation of feeling thirsty makes associations with water more available. It occurs when the first stimulus is related to the latter one semantically or emotionally or as an association, even when the stimuli are presented in different modalities (McNamara, 2005). In general, except for instinctive behaviour, something must be learned about a prime to have an effect on the subsequent stimuli (Ramscar, 2016).

In a famous priming-related paper by Vohs et al. (2006), the authors conclude after presenting nine experiments, that money encourages self-sufficiency. The manipulations used were small environmental changes or minor tasks to be completed by the participants, and the main task was to work on puzzles. In comparison to participants exposed to neutral concepts, participants primed with money preferred to work and play alone and kept more physical distance between themselves and the new people they met.

Another famous example of priming comes from Bargh et al. (1996). The authors reported three experiments, one on how priming with rudeness influenced the unpolite behaviour of participants, another on how subliminal priming with the African American stereotype

encouraged hostile reaction, and the third, arguably the most famous one, on priming with elderly stereotypes influencing walking speed. The participants in this experiment were asked to work on a scrambled-sentence task: one with neutral words and the second version with elderly prime, containing words like grey, retired, and Florida. After completing the task, the experiment was wrapped up, and participants could go. Experimenters measured the time it took for the participants to walk the hallway. The results were clear, those primed with the stereotype of the elderly walked more slowly. The general learning about unconscious cues that influence behaviour coming from priming studies has been repeatedly challenged (e.g., Abbott, 2013; Yong, 2012). Concerning the famous priming examples Rohrer et al. (2019) found no evidence of money priming, trying to replicate the first experiment from the series with puzzles (Vohs et al., 2006). Doyen et al. (2012) worked on replicating the experiment with walking speed being affected by a prime with the elderly stereotype (Bargh et al., 1996). Despite the use of a larger sample the authors were not able to replicate the outcome. Investigating possible reasons they found that the priming does replicate when they manipulated the experimenter's expectations. If people running the experiment expected participants to walk slower, the priming effect appeared. It is crucial to mention that the primed participants did not walk faster if the person running the study believed they would walk faster, so the effect could not be attributed to a self-fulfilling prophecy only. The main conclusion points to the importance of context to the behavioural expression of the prime. Another point is about participants being to large extent aware of the attempts to prime them, which could influence the outcomes in different ways, most likely encouraging more self-control to impair the effect.

The effects from straightforward lexical priming paradigms have turned out to be relatively robust, in contrast to the outcomes of more abstract and specific ways of priming. Although

social priming studies have recently failed to replicate, one must be mindful of the difficulties in replicating priming studies in time (Ramscar, 2016). Researchers have been finding that social priming is more nuanced, the effects are sensitive to individual differences and are smaller than first thought (Chivers, 2019). It still seems worth studying whether it is possible to impact people's behaviours with low-cost interventions. And there are examples of exposure to a robot altering the interaction with the next robot of different levels of anthropomorphic features.

The anthropomorphic robot priming effect was described by Zanatto et al. (2016). The study examined whether interactions with more human-like robots would alter subsequent interactions with a less human-like robot. The dependent measure was the adjustment of the initial valuation of an object to sell, in response to the robot's valuation. A significant change was reported for iCub (a humanoid robot) and was absent for Scitos G5 (a less anthropomorphic robot). Moreover, the difference for iCub was increased when the robot engaged in gaze behaviour mimicking social interactions compared with a fixed gaze scenario. This increase was not observed for the machine-like robot. After priming with iCub, people were more likely to adjust their price suggestions for Scitos G5, and this difference increased in the social gaze condition. The researchers concluded that the credibility of anthropomorphic robots increases due to the automatic activation of social stereotypes that are absent when interacting with less-anthropomorphic technology.

Rea and Young (2018) obtained a similar effect with teleoperation, where participants were presented with an opportunity to remotely 'drive' a robot. Both physical and tangible priming was employed, as well as descriptive priming. In one of the two studies, the researchers manipulated the joystick haptic experience to alter the sense of control and safety. In the second study, a description of the machine's capabilities and visuals were used to prime the participants.

In both cases, teleoperation effectiveness was operationalized and measured in the same way. It was reported that the priming methods shaped the perceptions of the robot and convinced participants that the machines were different. In reality, they were operating the same machine for 30 minutes. The researchers concluded that designing robots to feel a certain way – in this case, safer – can affect the perception of quality, adoption, or popularity, and can heavily influence usage patterns (Rea & Young, 2018). Even employing subtle shifts in language while talking about robots can influence how personally people view or treat a robot (Coeckelbergh, 2011).

If an object is deemed anthropomorphic, we connect it with it in a naturalistic manner that acts as a scaffold for learning new interaction patterns and employing social processes (Zanatto et al., 2016). As outlined above, anthropomorphism can be primed if not initially present. In the current line of research, the phenomenon of anthropomorphic priming is further explored to replicate the outcome and determine if primed human likeness affects intentionality bias towards robots.

Literature review summary and an introduction to the current research

Today, people have more experiences with technology or virtual agents than with biological ones (Nowak, 2001, as cited in Epley et al., 2008). It is a great challenge as well as an opportunity for psychology to contribute to the field and enhance the user-centred design practice, with basic knowledge about perceptual mechanisms playing a role in Human-Rrobot Interactions. After reviewing scientific and popular publications concerning biases in HRI, one can conclude that the main motivation for the research is to determine how we are going to be perceived by AI and how our own biases will be ingrained into the new systems by designers and engineers. There is much less attention paid to how our biases influence our relationship with

technology, especially with social robots. Specifically, the topic of intentionality perception is fascinating and seems to be crucial for the social layer of human-technology interactions. There have been a growing number of publications on this subject, but there are still only a limited number of projects looking at the differences between interpretations of behaviours by humans vs. technological agents (Thellman et al., 2017). Examining the intentionality bias in the perception of robots has the goal of filling those gaps. This leads to a research question of whether intentionality bias is present in the perception of robots, similar to the perception of human behaviour.

To assess intentionality bias, Rosset (2008) used a paradigm based on one-sentence behavioural descriptions. The test behaviours were either prototypically accidental (for less than 40% of pilot participants) or prototypically intentional (for more than 60% of pilot participants). The control sentences were those that almost all people categorize as either done on purpose or by accident (e.g., wrote an email - intentional, fell down the stairs - accidental). The control groups in the experiments judged whether the behaviour was done intentionally or not. The test conditions made biases and mental shortcuts more pronounced, promoting the use of the intuitive first system in Kahneman's (2011) terminology. The participants in the first experimental group judged the behaviours under time pressure and were asked whether they were done intentionally or by accident. The second experiment did not include time pressure, but the instructions for the participants did not explicitly mention the accidental option, removing the reminder that it was a plausible option. The hypothesis stated that our default interpretation is the intentional one and that without a reminder the bias has a better chance to appear. The third experiment investigated how well the sentences were remembered after the exercise. According to the dual processing model, the intentional interpretation is the default mode and has to be overwritten in case of

accidental behaviour, thus requiring more cognitive processing and making it more memorable. The results of the three studies provide support for the intentionality bias mechanism as a lack of inhibition to the default intentional interpretation (Rosset, 2008). The time constraint conditions made participants judge the ambiguous actions as more intentional. Not reminding participants about the potential accidental explanation of the ambiguous actions also caused more intentional interpretations. Finally, a recall of the sentences revealed that people needed additional processing to judge the behaviours as accidental, even for actions that are always perceived to be done unintentionally.

Considering that our perceptions of anthropomorphic artificial agents are similar to the perceptions of fellow humans, the same type of biases are plausible. Eisenkoeck and Moore (2017) adopted the Rosset's (2008) sentences paradigm for HRI research. In this study, the participants had to judge the actions described in the sentences as intentional or accidental when carried by a robot or, for a separate group, by a human. The robot condition mentioned the robot as the agent performing actions in the main instructions, and the human condition mentioned Robert as the performer of the actions. Conclusions were drawn based on a comparison between the two groups. Because the author did not introduce any bias-enhancing conditions it is not straightforward to conclude on the intentionality bias for robots, as the number of intention attributions could be, for example, the result of the specificity of a behaviour used in the context of judging robots. This leads to the second research question, whether intentionality bias operationalized as the difference between normal and bias-enhancing conditions, like the time pressure, occurs at the same rate for robots as for humans.

It is plausible that we explain and rate the behaviour of robots differently than human actions, not because of our perception of a robot, but because of the way we view the behaviour

itself. To determine the differences between human and robot behavioural explanations, we need to establish whether people judge the basic properties of this behaviour similarly for both agents. De Graaf and Malle (2018) advise that HRI studies should include behaviours that have a baseline, general perception that is similar for people and robots in three dimensions: intentional, surprising and desired. Researchers have collected behaviours used in HRI studies from the literature for perception evaluation. Seventy-eight action descriptions were included in the online questionnaire, and participants (n = 239) were asked to rate each on the three dimensions mentioned above either for a human or a robot as the agent. Each participant rated half of the behaviours for one agent for one of the dimensions. This study identified 29 robust behaviours for future comparative HRI studies. In light of this, the Rosset (2008) paradigm is not adapted to research in the human-robot interaction context, and this thesis aims at filling this gap, adding a pilot phase that evaluates Rosset's behaviour in the fashion suggested by De Graaf and Malle (2018).

Considering that our interactions with technology are entering the social level, the same type of biases are plausible as in human-to-human situations. As mentioned above, Eisenkoeck and Moore (2017) used the Rosset's (2008) sentences paradigm in HRI research. In the study, participants judged the actions described in the sentences as being done on purpose or by accident by a robot or a human. Eisenkoeck and Moore (2017), mentioned a several aspects worth improving in future studies. One of them is to expand the research procedure by adding a robot picture, instead of just mentioning a robot in the instructions. This is an important element of the experimental design that allows for manipulation of the level of anthropomorphic characteristics of a robot stimulus. Other researchers (Thellman et al., 2017) suggest that experimental stimulation should be in at least the visual modality – with preferably more

engaging modalities utilized. The authors also suggest designing studies that follow a stimulation level increase, starting from pictures, continuing to storyboards, videos, and interactions in virtual reality, and ending with real interactions. It is of importance to control the mental representations of robots carried by the participants entering the experiment, and using visual stimuli has a greater chance of doing so. They may differ greatly, being rarely taken from a personal experience rather than pop culture patterns. They are most often organized around the concept of a tool, the future, and subsidiarity (Piçarra et al., 2016). Intuitively, the more the stimuli approach real interactions with a robot, the more pronounced the experience. Mental state attribution is also an 'online' process that occurs throughout the interaction with stimuli (Takahashi et al., 2013). This is in contrast to a one-time categorization for the object class, as in the case of written instructions as the only stimulation. In the proposed continuation of the research by Eisenkoeck and Moore (2017), the goal is to expand the research procedure by adding visual stimuli instead of just mentioning a robot in the instructions. The research focuses on the appearance of the robots, presenting pictures at the beginning of the experiment and a miniature of it on every subsequent screen with a question. This procedure reflects the 'online' process of attributing mental states. It also allows to introduce different levels of anthropomorphic features and leads to the research question of whether the potential intentionality bias towards robots depends on the level of anthropomorphism.

HRI studies confirm that robots rated as more anthropomorphic are more likeable, and people express a greater willingness to interact with them in the future (Damiano & Dumouchel, 2018). The extent to which a robot is perceived as anthropomorphic is due not only to its characteristics, but also to how strong this perceptual tendency is for the individual human. Research shows that empathy is related to anthropomorphism (Harrison & Hall, 2010), as

anthropomorphism can be an integral part of this general ability to comprehend the inner states of other beings. In addition, as shown by Slavny and Moore (2018), cognitive, but not emotional empathy, allows for predictions of the intentional attributions to others' behaviour in situations prototypically perceived as accidental or ambiguous, making it another interesting topic of individual differences. Research participants can also differ in terms of the valence of their attitudes toward robots. They can help predict human behaviour towards them, such as the willingness to collaborate, which is the main use of these machines (e.g Nomura et al., 2005). Also, people who believe in the uniqueness of human nature are more likely to hold more negative attitudes toward anthropomorphic robots (Mozaryn et al., 2016). The degree of uniqueness that a participant attributes to human nature is likely to influence the evaluation of an anthropomorphic robot, including intentionality. It is justified to evaluate these traits of participants exposed to robots, which leads to investigating questions whether the tendency to anthropomorphize, empathy, attitudes towards robots and the belief in human nature uniqueness is related to the number of intention attributions to robots.

The goal of many robot designers is to make them anthropomorphic 'enough' to give cues on how to interact with them. Many design decisions are being taken intuitively. Phillips et al. (2018) attempted to systematize robots' appearance characteristics, organizing them by those that make them look more anthropomorphic. They collected more than 200 pictures of humanoid robots and included them in two studies. The results showed four sets of characteristics with the strongest prediction powers towards the anthropomorphizing of a robot (body-manipulators, surface-look, facial features, mechanical locomotion). The team then created a database covering commercial social robots and rated them in terms of the predicted anthropomorphism they invoke. The database was used to select the stimuli for the current study. This allows giving a

number to a level of anthropomorphism of studied robots, which can be helpful in using learnings from the study, especially if these robots are not well-known products. Choosing stimuli in a systematic way makes the study more useful to the field of human-robot interaction design, so important, as robots' appearance is its main interface.

As discussed earlier, according to Dennett (1987) we can perceive technology using three main strategies for predicting its behaviour, the physical stance, design stance and intentional stance. We can assume we use mainly design stance, the knowledge of how the machine was built, to interact with the technological tools. Robots may be perceived from both a design and intentional stance as the intentional strategy seems a rational and effective way to deal with intelligent systems (Perez-Osorio & Wykowska, 2020). Wiese et al. (2017) suggest that we should design robotic agents in a way that can trigger us to switch between the perceptual models or stances, when appropriate. Thus, as we enter the AI era, it seems even more important to understand how and under what conditions these stances are being activated.

Knowing how to influence the perception of a robot with a small intervention could be a powerful tool for the UX practice dealing with human-robot interaction design. Priming with a different level of anthropomorphism could potentially fulfil this role, allowing to switch the ‘perception mode’ of a robot user. Salem et al. (2011) showed a huge influence of gestures on the anthropomorphizing of artificial beings, as well as changes in sympathy and a general willingness to interact with them. As movement is one of the strongest cues for anthropomorphism, therefore the current research uses video stimuli as a priming manipulation. The goal is to replicate earlier results (e.g., Rea & Young, 2018) that priming can influence robots’ perception, and to investigate the research question of whether priming with anthropomorphism level can influence intentionality bias towards robots.

The existing research on intentionality perception can utilize paradigms introduced in studies on people. The work of Eisenkoeck and Moore (2017) is an interesting extrapolation, but it lacks the aspects of enhancing biases in the experimental design, as well as a validation of the behaviours themselves for relevance in the human-robot context. In addition, the stimuli used in this study were text-based, and others (e.g., Thellman et al., 2017) have suggested that the experimental stimulation should be in at least the visual modality. Thanks to presenting the robot's appearance, it is possible to test intentionality attributions to robots of different levels of anthropomorphic features. The current project is aimed at filling these gaps. Therefore, the main research question is whether intentionality bias occurs for a social robot at the same level as that for a human. To answer this question, the current research validates the original paradigm used to study intentionality bias and examines its applicability in HRI studies.

Method

The general goal of research

The main goal of this research is to answer whether intentionality bias is specific to the perceptions of humans or whether it is present in the perceptions of social robot behaviours. In addition, this research will determine whether the level of anthropomorphic features of a robot affects the level of intentionality attribution. This project will also explore whether the tendency to anthropomorphize, empathy, attitudes towards robots, and a belief in human nature uniqueness have effects on intentionality perceptions. Moreover, this project examines the possibility of influencing a robot's intentionality perception by employing priming to steer the perception of the anthropomorphic features of a robot.

General outline

Study 0 - Pilot

This pilot study aimed to prepare a tool to measure intentionality attribution. The Rosset's (2008) paradigm was adapted. The item sentences from the original battery were translated into Polish and then back-translated by an independent translator. The final version was negotiated with a doctoral advisor and tested in a pilot study, which was aimed at checking whether the sentences would be assessed as intentional by Polish participants at a level similar to that obtained in Rosset's (2008) pilot. In addition, as advised by De Graaf and Malle (2018), the pilot consisted of two groups of participants: one assessing whether the behaviour described in the sentence is intentional for a human and the other group for a robot, on a 7-point Likert scale. The goal was to establish whether people judge the basic properties of behaviour similarly, regardless of whether the actor is a human or a robot.

Study I

The goal of this study was to compare the intentionality bias towards humans and robots with low or high levels of anthropomorphic features. In addition, a time pressure condition was introduced as a between-subjects factor, as it enhances the prevalence of cognitive biases (Rosset, 2008) and, thus, the intentionality bias. The tendency to anthropomorphize (Waytz et al., 2010), empathy (Reniers et al., 2011), negative attitude towards robots and belief in human nature uniqueness (Pochwatko et al., 2015) were included as covariates.

This study aimed to address the following research questions:

1. Does time pressure increase the number of intention attributions?

2. Does intentionality bias (operationalized as the number of intention attributions to prototypically accidental behaviours under time pressure) occur at the same rate for robots as for humans?
3. Does the level of anthropomorphism (operationalized as the general human-likeness score from the ABOT database) influence the number of intention attributions to robots?
4. Is the tendency to anthropomorphize related to the number of intention attributions?
5. Is empathy related to the number of intention attributions?
6. Is the negative attitude towards robots related to the number of intention attributions?
7. Is the belief in human nature uniqueness related to the number of intention attributions?

Study II

The goal of this study was to evaluate whether priming can alter the perception of the robots and influence intentionality bias. The stimulus included two robots with a low or high level of anthropomorphic features. Priming in the opposite direction of the level of the anthropomorphic features of the robot (a mechanical robot was primed with a humanoid and vice versa) was introduced as a between-subjects factor. A time pressure condition was introduced as a between-subjects factor, as a bias enhancer (Rosset, 2008). The tendency to anthropomorphize (Waytz et al., 2010), empathy (Reniers et al., 2011), the negative attitude towards robots and the belief in human nature uniqueness (Pochwatko et al., 2015) were included as covariates.
This study aimed to address the following research questions:

1. Does priming influence the intentionality bias (operationalized as the number of intention attributions to prototypically accidental behaviours) towards social robots?
2. Is the tendency to anthropomorphize related to the number of intention attributions to robots?

3. Is empathy related to the number of intention attributions to robots?
4. Is the negative attitude towards robots related to the number of intention attributions to robots?
5. Is the belief in human nature uniqueness related to the number of intention attributions to robots?

Summary

Table 1 shows a summary of the experimental manipulations in both studies. The negative attitude towards robots, the belief in human nature uniqueness, the tendency to anthropomorphize, and empathy were measured.

Table 1.

Experimental conditions in the proposed studies

	Human	Robot high anthropomorphism	Robot low anthropomorphism	Time pressure	Priming
Study I	❖	❖	❖	❖	
Study II		❖	❖	❖	❖

Note. Priming with the opposite level of the robot's anthropomorphism (a mechanical robot was primed with a humanoid and vice versa).

Material and apparatus

Agent type. For the current line of research static visual stimuli were used. Previous research has shown that intentional agents can be differentiated from non-intentional agents within a fraction of a second, and that passive viewing of stimuli can induce the activation of

brain regions associated with social cognition (Wheatley et al., 2011). In addition, attributing ‘humanness’ follows a categorical pattern of either a yes or a no answer, and ambiguity appears in around 63% of physical humanness (Looser & Wheatley, 2010). Therefore, for the following research line, extreme-looking robots were chosen, either very mechanical or with a high level of anthropomorphic features.

The manipulation of the anthropomorphic features of a robot’s appearance was based on the ABOT database developed by Phillips et al. (2018), which contains examples of various robots and their human-likeness assessments. The ABOT database provides a framework for assessing a robot’s appearance, giving it a rating on a scale between 0 and 100, where 0 means no anthropomorphic features at all, and 100 means fully resembling a human being. The predicted human-likeness score and other dimensions are based on a multiple regression model obtained from data from 1,240 participants. Two robots were included as stimuli in the study, one with a low level of anthropomorphic features and the other with a high level. The features of these robots were chosen such that an ABOT assessment placed the low score below 20 and the high score above 80.

Figure 1.

Robot 1

ABOT Database Estimator Results for the mechanical robot, shown in Figure 1, used as a stimulus in the studies is 16.1 for the overall human-likeness score. The robot's 1 dimensions scores are: Body-Manipulators: 0.6, Surface-Look: 0, Facial Features: 0, Mechanical Locomotion: 0.

Figure 2.

Robot 2

ABOT Database Estimator Results for the anthropomorphic robot, shown in Figure 2, used as a stimulus in the studies is 83.75 for the overall human-likeness score. The robot's 2 dimensions scores are: Body-Manipulators: 1, Surface-Look: 0.43, Facial Features: 1, Mechanical Locomotion: 0.

Figure 3.

Human 'Robert'

The picture for the human condition was chosen with the help of independent judges and aimed at a neutral male human, shown in Figure 3. The human was described as 'Robert' in the instructions for the participants, the same name used by Eisenkoeck and Moore (2017). The rationale for such a choice is that 'Robert' and 'Robot' are words similar in length and sound, which minimises the influence of the name itself on the experimental conditions.

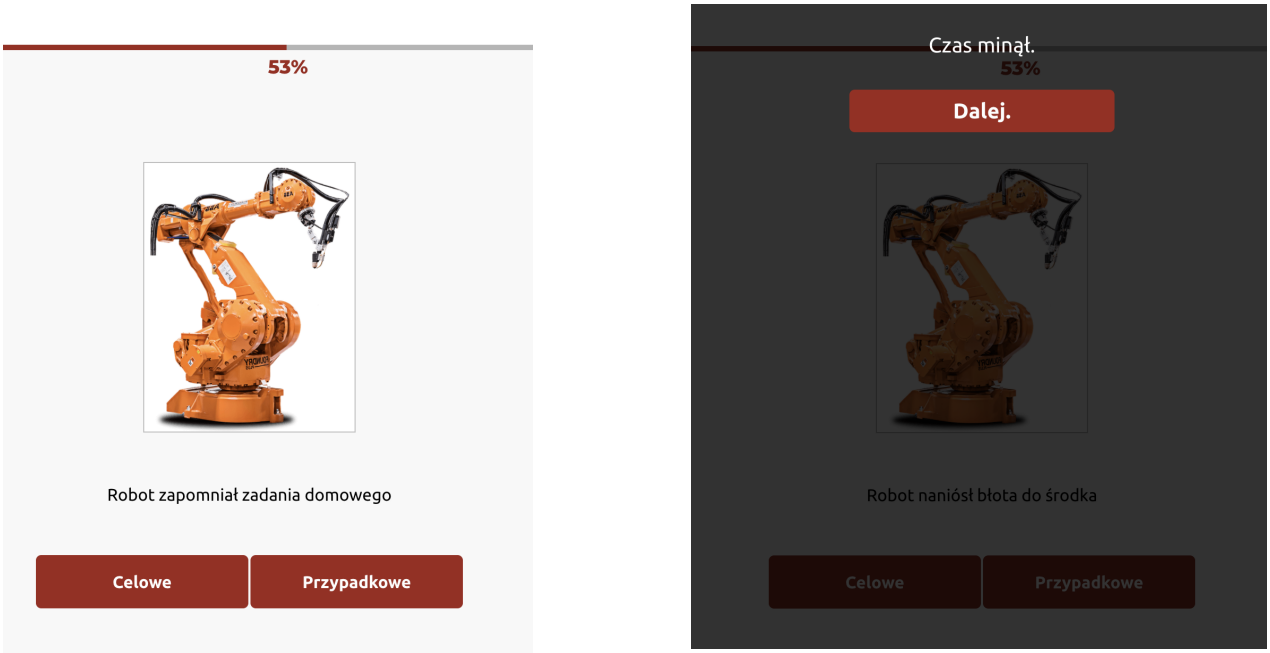
Intentionality. Intentionality attribution was measured by an adaptation of the sentences paradigm used by Rosset (2008), which was translated into the Polish language and evaluated in pilot Study 0 according to the procedures of De Graaf and Malle (2018). There were four types

of behaviours in the original paradigm that formed four scales. The two test scales include statements regarding ambiguous behaviours that are usually perceived as done on purpose or by accident: Prototypically Accidental (PA; e.g., ‘He burnt the meal’) and Prototypically Intentional (PI; e.g., ‘He deleted the email’). The two control scales include unambiguous behaviours that most people would judge as done on purpose or by accident: Control Accidental (CA; e.g., ‘He fell down the stairs’) and Control Intentional (CI; e.g., ‘He typed the email’). All the labels of the scales are kept the same as in the ordinal paradigm (Rosset, 2008), to make it consistent with the literature.

Manipulation enhancing biases. In the original sentence paradigm for researching intentionality bias, one of the experimental manipulations used by Rosset (2008) was time pressure, namely allowing 2.4 s to determine if the described behaviour in a sentence was intentional or accidental. Although the time for an answer is relatively long and likely to influence the effects (e.g., Sharma & McKenna, 2001), the current research adopted this manipulation, to keep it consistent with the literature (e.g., Hughes et al., 2012). The screen design is pictured in Figure 4 on the left, and the feedback after the timeout is shown on the right.

Figure 4.

The design of screens in the experiments



Priming. Priming is a method of using one stimulus to cause an effect on the perception of a subsequent stimulus. Research has shown that priming can affect the perception of a robot when it concerns the level of anthropomorphic features (Zanatto et al., 2016). In the study, researchers introduced priming using a written description or a picture. Exposing participants to a robot's behaviour can have an even more substantial effect. In the current project, 23 s videos were used. For priming with a low level of anthropomorphic features, the video showed an industrial arm carving a piece of wood. For priming with a high level of anthropomorphic features, the video showed the iCube robot walking and making gentle gestures, such as waving with one hand. The look of the robots matched those used in the still picture stimuli, as in Figures 1 and 2.

Tendency to anthropomorphize. The tendency to anthropomorphize is common, but there are significant individual differences. Measuring this variable can potentially explain, in part, the

level of intention attribution to social robots. The tendency to anthropomorphize was measured using the Individual Differences in Anthropomorphism Questionnaire (IDAQ; Waytz et al., 2010). This scale was translated into the Polish language and received a back-translation from a professional translator. The scale consists of 30 items in two subscales: the IDAQ (which assesses anthropomorphism) and the IDAQ-NA (which assesses non-anthropomorphic attribution). The IDAQ assesses the tendency to anthropomorphize (e.g., 'To what extent does the average fish have free will?', 'To what extent does the average insect have a mind of its own?'), and the IDAQ-NA assesses non-anthropomorphic attribution (e.g., 'To what extent is the average cloud good-looking?', 'To what extent is the forest durable?'). The participants are asked to rate the extent to which the object possesses the feature on a 0–10 scale, where zero represents 'not at all,' and ten represents 'very much'.

Empathy. Cognitive empathy allows individuals to predict intention attribution in situations stereotypically perceived as accidental or ambiguous (Slavny & Moore, 2018). Empathy as a trait was measured with the Questionnaire of Cognitive and Affective Empathy (QCAE; Reniers et al., 2011) in the Polish adaptation by Agnieszka Lasota and Katarzyna Tomaszek (Lasota et al., 2020), obtained with the consent of the authors. The QCAE has 31 items divided into five subscales. Two of them measure cognitive empathy, i.e. Perspective Taking (e.g., 'I can easily tell if someone else wants to enter a conversation'), and Online Simulation (e.g., 'Before criticizing somebody, I try to imagine how I would feel if I was in their place.'). There are also three subscales measuring affective empathy: Emotional Contagion (e.g., 'People I am with have a strong influence on my mood'), Proximal Responsivity (e.g., 'I often get emotionally involved with my friends' problems'), and Peripheral Responsivity (e.g., 'It is hard for me to see why

some things upset people so much.'). The answers are given on a 4-point scale, with one representing 'strongly disagree' and four representing 'strongly agree'.

Negative Attitudes Towards Robots Scale (Polish version). The scale measuring attitudes towards robots was originally created by Nomura et al., (2005) and was adapted to the Polish language and culture as the Negative Attitudes Towards Robots Scale Polish version (NARS-PL; Pochwatko et al., 2015). This scale measures psychological reactions to robots, both to humanoids and those not resembling people. The tool focuses on the willingness to interact with the presented robot in the future. The Polish adaptation of the scale has 12 items and forms two subscales: the Negative Attitudes Toward Interactions with Robots (NATIR; e.g., 'I would feel uneasy if I was given a job where I had to use robots') and the Negative Attitudes toward Robots that display Human Traits (NARHT; e.g., 'I would feel uneasy if robots really had emotions'). The answers are given on a 7-point scale, where one represents 'totally disagree' and seven represents 'totally agree'.

Belief in Human Nature Uniqueness. The belief that human nature is unique can influence attitudes toward social robots (Pochwatko et al., 2015). The conviction is measured by the Belief in Human Nature Uniqueness scale (BHNU; Pochwatko et al., 2015), which consists of six items (e.g., 'Even if ultra-sophisticated, a robot will never feel the same emotions as a human being'. Responses are given on a 7-point Likert scale, where one represents 'totally disagree', and seven - 'totally agree'.

Statistical analysis

All statistical analyses were conducted in JASP (JASP Team, 2022), a statistical open-source program, and the open-source RStudio (RStudio Team, 2020) with R language (R Core Team, 2022), utilizing the package psycho (Makowski, 2018), among others. The

visualizations and tables were constructed using both programs. All figures are reported with a 95% confidence interval. All the data are available on the Open Science Framework, a data management program for open science (Zerka, 2022).

Research reports

Study 0 - Pilot

Participants

Participants for the pilot study were recruited via private social media channels. The pilot of the test scales utilized two groups of participants, $N = 78$ for the robot version and $N = 86$ for the human version. The pilot of the control scales used $N = 36$ for the robot version and $N = 33$ for the human version.

Procedure

This pilot study aimed to prepare a tool to measure intentionality attribution. The Rosset (2008) paradigm was adapted. The item sentences from the original battery were translated into Polish, and then back-translated by an independent translator. The final version was negotiated with a doctoral advisor and tested in a pilot study conducted to validate the translation and identify potential culture-based differences, as well as to assess the behaviours in the context of human-robot interactions. The validation goal was to check whether the sentences from a given scale would obtain the same level of intentional interpretation after translation to another language and to collect data from a different population. According to the original paradigm, PA test sentences should have an intentionality likelihood score below 40% (judged to be accidental by at least 60% of participants), and PI sentences should have an intentionality likelihood score above 60% (judged to be intentional by at least 60% of participants). The control scales consisted

of behaviours that are almost always interpreted as intentional or accidental. In Rosset's (2008) study, these items obtained intentionality likelihood scores of at least 98% or lower than 2%.

To validate the behaviours for researching human-robot interactions, the procedure of De Graaf and Malle (2018) was followed for the intentionality dimension. According to this approach, there should be no differences in the intentionality scores for each behaviour between the human and robot versions before the experimental manipulation. This analysis allows us to assess if people judge the basic properties of behaviour similarly regardless of whether it is performed by a human or a robot and to mitigate uncontrolled variability in the data for the main studies.

Rosset's (2008) original paradigm was designed to study the perception of human behaviours; therefore, each sentence was presented with a gender pronoun. As this pilot aimed to compare a human to a robot, gender pronouns were replaced by Robert for the groups judging human behaviour and by Robot in the groups judging the robot's behaviour, as was done by Eisenkoeck and Moore (2017).

Results

Test scales. Participants made intentionality judgments regarding each sentence describing a behaviour (intentional or done by accident) by filling out an online survey and checking the appropriate boxes. As outlined above, the pilot study consisted of two groups, one judging the actions of a robot (n = 78; e.g., 'Robot hit the man with his car') and the other judging the actions of a human (n = 86; e.g., 'Robert hit the man with his car').

The original paradigm consists of two test scales, the PA scale, and the PI scale. The PA test sentences were used by Rosset (2008) as the main scale to assess the intentionality bias. The

percentage of people who checked ‘on purpose’ gives us the 'intentionality likelihood rating’ for each sentence.

The prototypically accidental scale (PA). The same threshold as that utilized by Rosset (2008) was used in the current study. PA test sentences should have an intentionality likelihood score below 40% (judged accidental by at least 60% of the participants). This level was used for studies on human behaviour; therefore, the analysis starts by evaluating the human condition in our pilot. The following sentences were above the threshold of 40% intentionality and were subsequently excluded: 'Robert kicked the dog' (54.65%), 'Robert set the house on fire' (41.86%), and ‘Robert popped the balloon’ (48.84%). In addition, sentences with significant differences in intentionality ratings between the Robert vs. Robot questionnaire versions were excluded. This included: ‘Robert dripped paint on the canvas’, $\chi^2(1, N = 164) = 5.06, p = .02$. The results for the PA behaviours are shown in Table 2.

‘Robot left the water running’ was removed because of the complex grasping operation that Polish language translation implies, which might be difficult for robots to perform today.

Table 2.

Intentionality likelihood ratings for the test sentences (prototypically accidental sentences)

Sentence in Polish	Sentence in English	% Robert	% Robot	χ^2	*p*
R. potrącił człowieka samochodem	R. hit the man with his car	29.1	28.2	0.01	.903
R. przypalił posiłek	R. burnt the meal	4.7	7.7	0.66	.416
R. stłukł wazę	R. broke the vase	17.4	16.7	0.02	.895
R. naniósł błota do środka	R. tracked mud inside	10.47	10.26	1.92e-3	.965
R. zapomniał zadania domowego	R. forgot the homework	10.47	16.67	1.37	.245
R. spóźnił się 5 minut	R. arrived 5 minutes late	11.63	11.54	3.19e-4	.986
R. wybił okno	R. broke the window	26.74	26.92	6.67e-4	.979
R. obudził dziecko	R. woke the baby up	32.56	38.46	0.62	.430
R. wdepnął w kałużę	R. stepped in the puddle	19.77	24.37	0.50	.478
R. uruchomił alarm	R. set off the alarm	32.56	44.87	2.62	.105
R. poplamił farbą płótno	R. dripped paint on the canvas	13.95	28.21	5.06	.025
R. kopnął psa	R. kicked the dog	54.65	19.23	21.82	< .001
R. zostawił odkręcony kran	R. left the water running	10.47	19.23	2.52	.113
R. podpalił dom	R. set the house on fire	41.86	20.51	8.61	.003
R. przebił balon	R. popped the balloon	48.84	39.74	1.37	.242

The PA scale was formed by the remaining ten sentences. In order to assess the scale's reliability, Cronbach's alpha[1] was used. 'R. hit the man with his car' correlated negatively with the rest of the scale (r = -0.192) and was therefore excluded. The final PA scale consisted of nine items and was fairly reliable ($\alpha = 0.734$).

The prototypically intentional scale (PI). The same procedure was carried out for the PI sentences, which created the supporting test scale, consistent with Rosset's (2008) approach. The aim was to obtain sentences judged to be intentional by at least 60% of the participants for the human condition. In this case, the bar was lowered to 55% to form the scale, as an insufficient number of items passed the threshold. Five sentences having a less than 55% intentionality likelihood score were removed: 'Robert cut him off driving' (39.53%), 'Robert drove over the speed limit' (46.51%), 'Robert took an illegal left turn' (40.70%), 'Robert walked by without saying hello' (29.07%), and 'Robert left without leaving a tip' (45.35%). All the remaining sentences were significantly different in terms of the intentionality ratings for a human and a robot, and the behaviours were judged to be less intentional when carried out by a robot. The results for the PI behaviours are listed in Table 3.

[1] KR20 was considered, but for the type of analysed data, it is mathematically equivalent to the formula for coefficient alpha. Multiple problems related to using Cronbach's alpha were considered (see Sijtsma, 2008). Following the advice of Trizano-Hermosilla and Alvarado (2016), GLB was considered, as some of the scales were skewed, but as reported by Malkewitz et al. (2023) in small samples GLB overestimated strongly and the performances of alpha and omega were similar, Crombach's alpha was used as a reliability measure.

Table 3.

Intentionality likelihood ratings for the test sentences (prototypically intentional sentences)

Sentence	Sentence in English	% Robert	% Robot	χ^2	*p*
R. rozerwał kartkę papieru	R. ripped the piece of paper	51.16	42.31	1.29	.256
R. zajechał mu drogę	R. cut him off driving	39.53	24.36	4.31	.038
R. usunął email	R. deleted the email	62.79	47.44	3.90	.048
R. przekroczył prędkość	R. drove over the speed limit	46.51	24.36	8.71	.003
R. zignorował pytanie	R. ignored the question.	56.98	25.64	16.48	< .001
R. zburzył zamek z piasku	R. knocked over the sand castle	58.14	30.77	12.37	< .001
R. zostawił znak na kartce papieru	R. made a mark on the paper	56.98	35.90	7.30	.007
R. spryskał go wodą	R. sprayed him with water	55.81	32.05	9.35	.002
R. skręcił w lewo na zakazie	R. took an illegal left turn	40.70	16.61	11.41	< .001
R. przeszedł obok, nie witając się	R. walked by without saying hello	29.07	30.77	0.06	.812
R. wyszedł nie zostawiając napiwku	R. left without leaving a tip	45.35	24.36	7.88	.004

In order to assess the reliability of the scale, Cronbach's alpha was used. The final scale for the PI sentences consisted of five items, and was fairly reliable ($\alpha = 0.764$).

Control scales. A verification procedure was performed for the translated control sentences as well, which aimed to confirm that they were at least 98% judged as intentional or accidental. The original study presented 40 control sentences, 20 for intentional behaviours and 20 for accidental behaviours. Eighte sentences were included in the pilot study (nine in each category

chosen randomly) to limit the number of items in the questionnaire. We asked two groups of participants to indicate the intentionality of a given action (intentional or done by accident), one group for a robot (n = 36) and the other for a human (n = 33). The responses from one participant were excluded as the answer was yes for all of the items in both scales.

For the human condition, we included in the control scales items that were rated as either intentional or accidental by at least 97% of the participants. The threshold was lowered to 90% for the robot condition, as the rates were less consistent. The results for the control behaviours are listed in Table 4.

Table 4.

Intentionality likelihood ratings for the control sentences (unambiguously intentional and accidental)

Sentence in Polish	Original sentence in English	% Robert	% Robot	χ^2	*p*
R. nie zdał egzaminu na prawo jazdy	R. failed the driving test	3.03	16.67	3.51	.061
R. spadł ze schodów	R. fell down the stairs	3.03	8.33	0.89	.346
R. spadł z deskorolki	R. fell off the skateboard	3.03	8.33	0.89	.346
R. nie trafił piłką do kosza	R. missed the hoop with the ball	3.03	16.67	3.51	.061
R. przytrzasnął palce drzwiami	R. pinched his fingers in the door	3.03	8.33	0.89	.346
R. potknął się o krawężnik	R. tripped on the curb	0.0	5.56	1.89	.169
R. zgubił klucze	R. lost her keys	0.0	13.89	4.94	.026
R. popsuł telefon	R. broke her cell phone	12.12	36.11	5.34	.021
R. poślizgnął się na lodzie	R. slipped on the ice	0.00	8.33	2.88	.090
R. narysował widok z plażą	R. drew a picture of the beach	93.94	83.33	1.89	.169
R. napisał email	R. typed the email	96.97	91.67	0.89	.346
R. odkurzył dywan	R. vacuumed the carpet	96.97	94.44	0.26	.607
R. zaadresował list	R. addressed the letter	100.0	88.89	3.89	.049
R. upiekł ciasto	R. baked a cake	100.0	94.44	1.89	.169
R. zmienił przebitą oponę	R. changed the flat tire	96.97	91.67	1.89	.346
R. zrobił korektę artykułu	R. proofread her paper	100.0	94.44	1.89	.169
R. nawlekł igłę	R. threaded the needle	96.97	91.67	0.89	.346
R. zapalił świeczkę	R. lit the candle	100.0	91.67	2.88	.090

The control accidental scale (CA). From the six items that passed the threshold of being judged as accidental 97% for the human and 90% for the robot ('R. fell down the stairs', 'R. fell off the skateboard', 'R. pinched his fingers in the door', 'R. tripped on the curb', 'R. lost her keys', and 'R. slipped on the ice'), two items were excluded to achieve the best possible reliability ('R. lost his keys' and 'R pinched his fingers in the door'). The final scale consisted of 4 items and $\alpha = 0.88$.

The control intentional scale (CI). Four items from the CI scale that passed the same thresholds of being judged as intentional were chosen to match the number of items in the other control scale. These included: 'R. typed the email', 'R. baked a cake', 'R. changed the flat tire', and 'R. lit the candle'. For this scale $\alpha = 0.91$.

Summary

The final items for the four scales are shown in Table 5.

Table 5.

The final scales used in the current study

Prototypically Accidental (PA)	*Prototypically Intentional (PI)*	*Control Accidental (CA)*	*Control Intentional (CI)*
R. burnt the meal	R. deleted the email	R. fell down the stairs	R. typed the email
R. broke the vase	R. drove over the speed limit	R. fell off the skateboard	R. baked a cake
R. tracked mud inside	R. ignored the question.	R. tripped on the curb	R. changed the flat tire
R. forgot the homework	R. knocked over the sand castle	R. slipped on the ice	R. lit the candle
R. arrived 5 minutes late	R. made a mark on the paper		
R. broke the window	R. sprayed him with water		
R. woke the baby up			
R. stepped in the puddle			
R. set off the alarm			

Study 1

Hypotheses

1. Time pressure will increase the level of intentionality attribution compared to the condition without time pressure.
2. The intentionality bias (operationalized as the number of intention attributions to prototypically accidental behaviours under time pressure) will occur for both robots and humans, and ambiguous actions will be rated more often as intentional under time pressure.
3. The intentionality bias will be stronger for a robot with a high level of anthropomorphic features than for a robot with a low level of anthropomorphic features.

4. The tendency to anthropomorphize is related to the number of intention attributions.
5. The level of cognitive empathy is related to the number of intention attributions.
6. The negative attitude toward robots is related to the number of intention attributions.
7. The belief in human nature uniqueness is related to the number of intention attributions.

Participants

Study participants ($N = 288$; detailed characteristics in Table 6) were recruited from the general panel pool and randomly divided into six groups. The study was conducted using the online panel Ariadna, a Polish research panel with 300,000 registered individuals aged 15 years and above. The sociodemographic profile of the individuals registered on the Ariadna panel corresponds with the profile of Polish Internet users. Exposure to technology was controlled by questions about working or studying in related fields. In the overall sample, only 12 participants were related professionally to technology, as higher technology exposure might lead to more complex or different cognitive representations of technological products (Pochwatko et al., 2015).

Table 6.

Demographic characteristics of the participants

			Education					
	N	Age Mean (SD)	Primary	Lower secondary	Upper secondary	Post-secondary	Bachelor's	Master's
Female	127	44 (14)	3.2%	7.0%	23.6%	15.7%	6.2%	44.1%
Male	161	49 (16)	4.3%	7.5%	27.3%	11.2%	11.8%	37.9%

Design

This study was a between-subjects 2 (time pressure vs. no time pressure,) x 3 (human vs. highly anthropomorphic robot vs. a low anthropomorphic robot) experimental design. A summary is presented in Table 7.

Table 7.

Experimental conditions for Study I

Condition	Stimuli	Time pressure	N after data cleaning
1	Robot high anthropomorphism level	Yes	34
2	Robot high anthropomorphism level	No	42
3	Robot low anthropomorphism level	Yes	37
4	Robot low anthropomorphism level	No	27
5	Human	Yes	32
6	Human	No	34

Procedure

This study was conducted online. The participants received a link from an online research panel that they were subscribed to. First, they were informed about the goal and rules of the study, asked to consent to participate in the study and about basic demographics. Next, the participants saw a screen with the instructions and five mock questions, which was especially important for groups with the 2.4 s time limitation to practise the procedure. The time limit was the same length as in the study that introduced the paradigm (Rosset, 2008). Next, the participants were exposed to one picture (out of three used in the study, a human, a highly anthropomorphic robot, and a low anthropomorphic robot) for 5 s. The 22 behaviours were presented from the PA, PI, CA, and CI scales in random order, followed by a question asking whether the described behaviour was intentional or accidental. Four questionnaires were administered in the following order: NARS-PL, BHNU, QCAE, and IDAQ. The last questions asked about the current activities of the participants, such as learning or working in a specific domain. The intention was to control the participant's exposure to technology, such as the number of engineers in each group.

Results

An 'intentionality endorsement score' (Rosset, 2008), defined as the number of 'on purpose' responses divided by the number of possible responses, was computed for each participant for each item in the test and control scales (PA, PI, CI, CA). The proportion represents each person's tendency to choose an intentional interpretation.

Of the 288 participants, the final sample after data cleaning was $N = 206$. Data from 58 participants were not included in the analysis because of missing more than 25% of the responses on one or more of the scales (PA ≤ 2 missing answers, PI ≤ 1, CA ≤ 1, CI ≤ 1). Missing values

were due to the experimental conditions, as the time was limited to 2.4 s. The missing values per question in the experimental condition ranged between 5–12%. The proportion of missing values did not differ between the scales, χ^2 (3, N = 288) = 5.99, p = 0.112, as checked with the Friedman test. Another set of data from 24 participants was removed because of answering the control scales of accidental and intentional behaviours the same, which might suggest that they did not understand the instructions or did not try to follow them. Outliers were investigated in the next step (defined as the average score for the participant for a given scale being more than three standard deviations away from the group mean; standard deviations for scales were between 0.2 and 0.3), but no cases were found.

The reliability of the test and control scales was assessed using Cronbach's alpha. The reliability was poor, but acceptable (Ngulube, 2022), PA, 9 items, α = .65, PI, 5 items, α = .57, CA, 4 items, α = .55, CI, 4 items, α = .65. This is most likely due to a large amount of missing data (Zhang & Yuan, 2016), as the experimental condition restricted the time for answers to 2.4 s. The item-test correlations for all items fell between the range of .02–.04; therefore, the scales were accepted (Piedmont, 2014). All the scales measuring covariates were fairly reliable (α > .71).

An ANCOVA[2] analysis was conducted to check the influences of agent type and time pressure on intentionality scores, with the tendency to anthropomorphize (IDAQ and IDAQ-NA subscales), empathy (Cognitive Empathy and Affective Empathy subscales), a negative attitude towards robots (NATIR and NARHT subscales) and the belief of human nature uniqueness (BHNU scale) as covariates for each of the four behaviour types. A non-normal distribution was expected, especially for the control scales (CA and CI). The ANCOVA was considered robust, as

[2] The ANCOVA reporting style is based on Field (2013).

according to Schmider et al. (2010), empirical type I and type II errors remain constant when the assumption is violated.

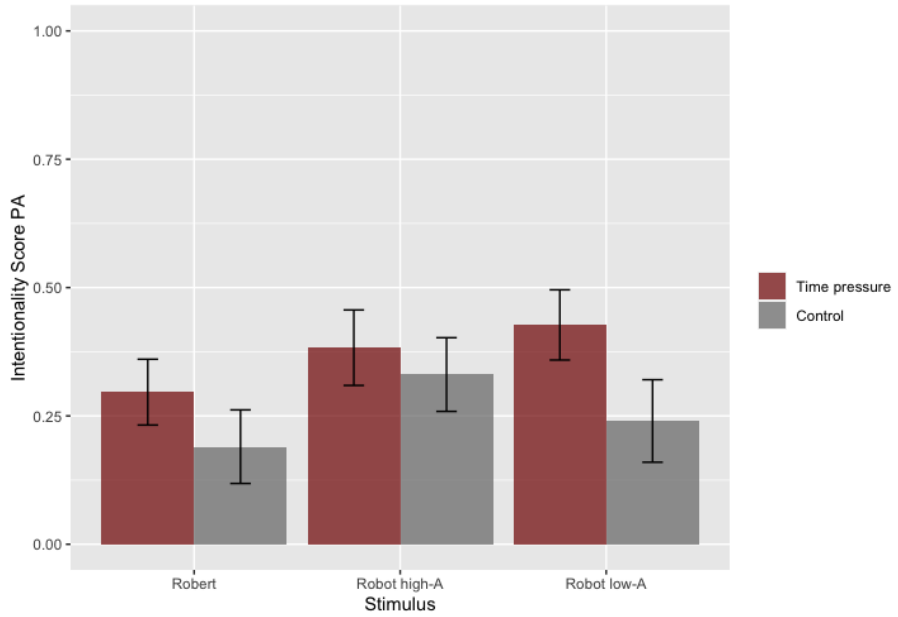
For the PA scale, the outcome shown in Figure 5, the homogeneity of regression slopes assumption check revealed no significant interaction between the covariates and independent variables.

There was no significant interaction between the time pressure and the agent type conditions in the main model, $F(2, 193) = 1.652$, $p = .194$, $\eta^2 = .014$. However, there were significant main effects of time pressure, $F(1,193) = 14.454$, $p < .001$, $\eta^2 = .059$, and agent type, $F(1,193) = 5.553$, $p = .005$, $\eta^2 = .045$. Post hoc testing using the Bonferroni correction revealed that the difference between the two robot agents was not significant, but it was between the human and both robots, (a difference between Robert and Robot with a low level of anthropomorphic features, and a difference between Robert and Robot with a high level of anthropomorphic features, $t = -2.458$, $p = .045$ and $t = -3.145$, $p = .006$, respectively). The intentionality scores were higher in the time pressure groups ($M = .356$) than in the control groups without time pressure ($M = .254$), as shown in Figure 5.

The negative attitude towards robots NATIR subscale was significantly related to the intentionality score of the PA scale, $F(1,193) = 11.360$, $p < .001$, $\eta^2 = .046$, as well as Affective Empathy, $F(1,193) = 4.421$, $p = .037$, $\eta^2 = .018$.

Figure 5.

The intentionality scores for prototypically accidental (PA) behaviours



Note. The figure presents the intentionality scores for prototypically accidental (PA) behaviours under the time pressure and control conditions for the three agent types; A: anthropomorphic features.

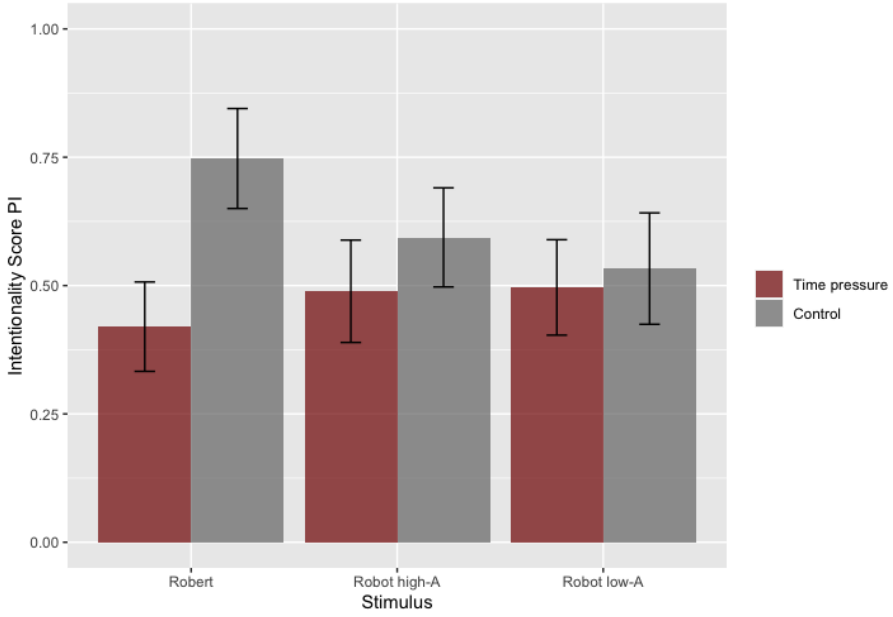
For the PI scale, the assumptions check revealed a significant interaction between the NATIR subscale and the type of agent. Therefore, NATIR was not included in the model.

There was a significant interaction effect between time pressure and agent type, $F(2, 196) = 5.339$, $p = .006$, $\eta^2 = .047$. In addition, there was a significant effect of time pressure in the main model, $F(1,193) = 12.486$, $p < .001$, $\eta^2 = .055$. The effect of the agent type was not significant, $F(1,193) = .958$, $p = .385$, $\eta^2 = .008$. None of the covariates were significantly related to the intentionality score. The intentionality scores were higher in the control groups ($M = .690$) than in the time pressure groups ($M = .520$), as shown in Figure 6. As to the interaction, the simple

main effect of time pressure was significant for the human, $F(1,196) = 25.300$, $p < .001$, $\eta^2 = .115$.

Figure 6.

The intentionality scores for prototypically intentional (PI) behaviours



Note. The figure presents the intentionality scores for prototypically intentional (PI) behaviours under the time pressure and control conditions for the three agent types; A: anthropomorphic features.

For the CA scale, the assumption check revealed a significant interaction between the subscales for anthropomorphism (IDAQ and IDAQ-NA) and time pressure; therefore, these covariates were not included in the model.

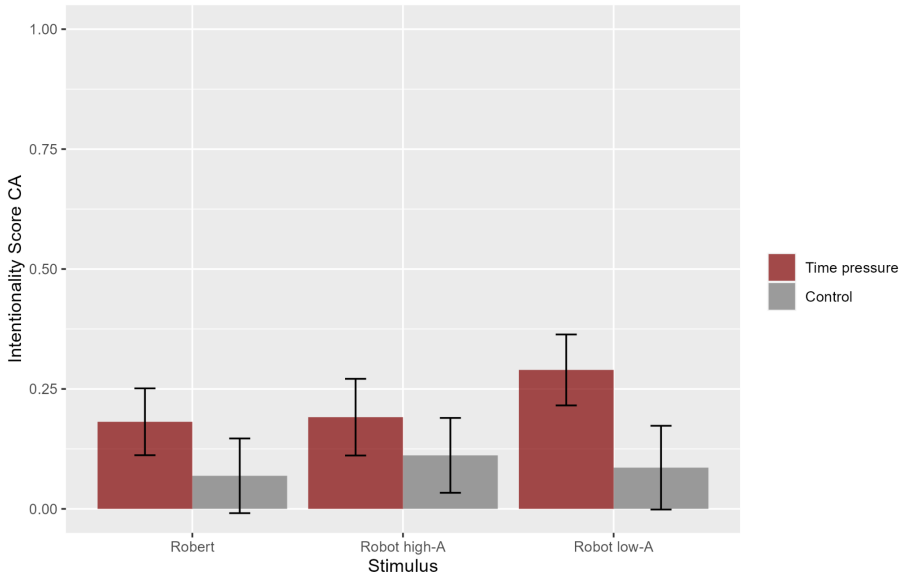
There was no significant interaction effect between the time pressure and agent type conditions, $F(2, 195) = 1.260$, $p = .286$, $\eta^2 = .010$. However, there was a significant effect of time pressure, $F(1,195) = 16.107$, $p < .001$, $\eta^2 = .067$. The effect of agent type was not significant

$F(1,195) = 1.230$, p = .295, $\eta^2 = .010$. The intentionality scores were higher in the time pressure groups (M = .221) than in the control groups (M = .088), as shown in Figure 7.

The negative attitude towards robots subscales (NATIR and NARHT) were significantly related to the intentionality scores on the CA scale, $F(1,195) = 16.817$, $p < .001$, $\eta^2 = .070$ and $F(1,195) = 6.476$, $p < .012$, $\eta^2 = .027$, respectively.

Figure 7.

The intentionality score for control accidental (CA) behaviours



Note. The figure presents the intentionality score for control accidental (CA) behaviours under the time pressure and control conditions for the three agent types. A: anthropomorphic features.

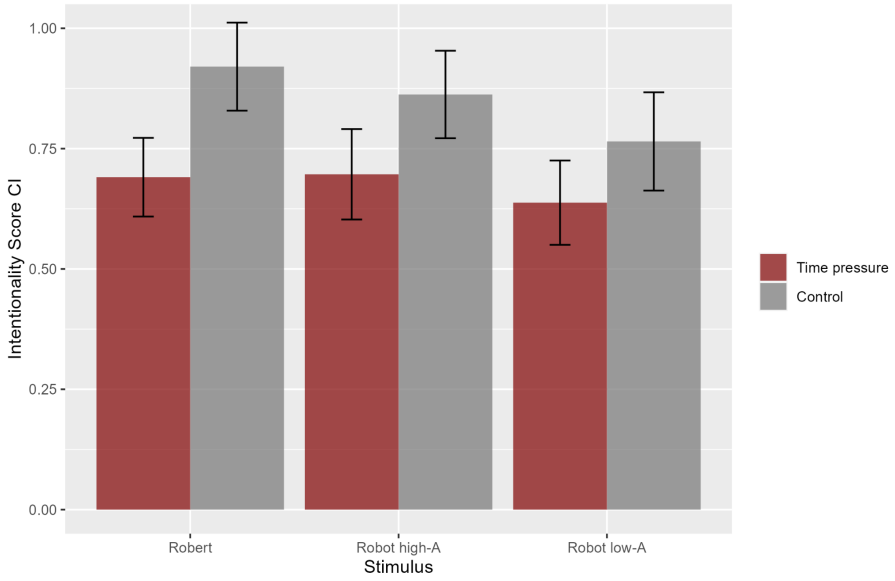
For the CI scale, the assumption check revealed a significant interaction between NATIR and the type of agent, as well as for the BHNU scale and condition. Therefore, these covariates were not included in the model.

There was no significant interaction effect between time pressure and agent type, $F(2, 194) = .632$, $p = .533$, $\eta^2 = .005$, and the main effect of agent type was not significant as well, $F(1,194) = 2.644$, $p = .074$, $\eta^2 = .020$. However, there was a significant effect of time pressure, $F(1,194) = 20.621$, $p < .001$, $\eta^2 = .077$. The intentionality scores were higher in the control groups ($M = .855$) than in the time pressure groups ($M = .675$), as shown in Figure 8.

Both of the anthropomorphism scales (IDAQ, IDAQ-NA), as well as NARHT and Cognitive Empathy, were significantly related to the intentionality scores on the CI scale, $F(1,194) = 24.877$, $p <. 001$, $\eta^2 = .093$, $F(1,194) = 8.855$, $p < .003$, $\eta^2 = .033$, $F(1,194) = 7.085$, $p = .008$, $\eta^2 = .026$, and $F(1,194) = 4.597$, $p = .033$, $\eta^2 = .017$, respectively.

Figure 8.

The intentionality scores for control intentional (CI) behaviours

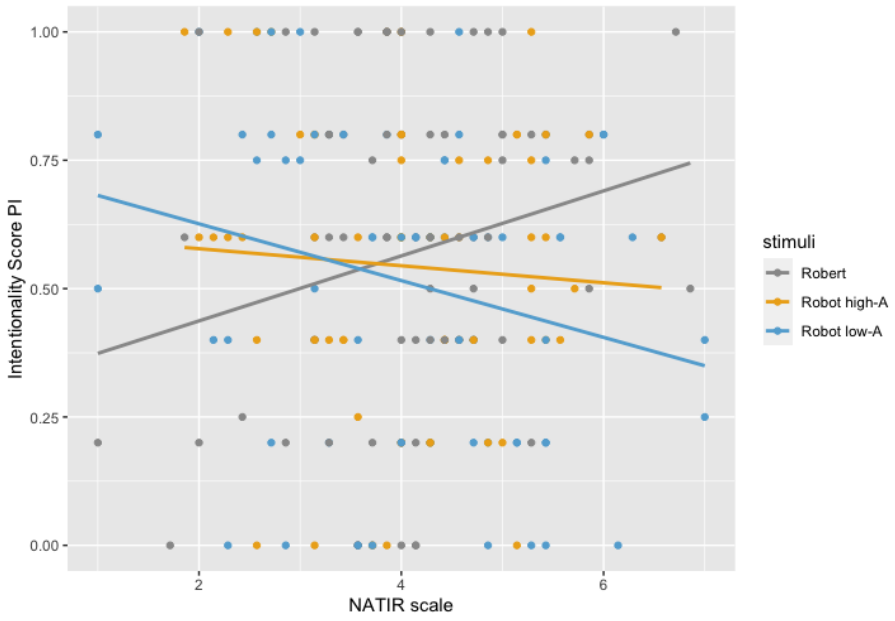


Note. The figure presents the intentionality scores for control intentional (CI) behaviours under the time pressure and control conditions for the three agent types; A: anthropomorphic features.

NATIR scale. The NATIR subscale of the NARS-PL was consistently significant in models with accidental behaviours. For both scales (PA and CA), the relationship was positive (i.e., the intentionality score was higher with higher NATIR scores; see the charts in Appendix A1 for PA and Appendix A2 for CA). As to the intentional behaviours, there was an interaction between NATIR and the type of agent on the PI and CI scales. As shown in Figure 9, for PI behaviours, the intentionality was rated higher for robots by participants with a low score on the NATIR scale, and lower for participants with a high score. The pattern for the human (i.e., Robert) was reversed.

Figure 9.

The scores for the NATIR scale for the three agent types for PI scale



Note. The figure presents the intentionality scores for prototypically intentional (PI) behaviours plotted against the scores for the NATIR scale for the three agent types; A: anthropomorphic features.

Post hoc analysis. To make sure that the results of this study are not random and significantly differ from a 50% occurrence, especially for the condition with the time limited to only 2.4 s, additional analyses were done. A one-sample t-test was used to separately check the time pressure and control conditions. Each ambiguous scale (PA and PI) mean was compared to a test value of 0.5. The control scales (CA, CI) were compared to test values 0 or 1, as the control behaviours are supposed to be always (or almost always) judged as accidental or intentional. Only the PI scale did not differ from the 0.5 test value in the no-time pressure condition. This is not surprising, as the behaviours for this scale that were chosen in the pilot study had to have at least 60% intentional attributions. However, as there were not enough items to form a scale, the bar was lowered to 55%. In the time pressure condition, all scale means differed from the respective test values, as shown in Table 8.

Table 8.

One sample t-tests for the conditions with and without time pressure

	No time pressure			Time pressure		
	t	**df**	**p**	**t**	**df**	**p**
PA Scale	-15.592	143	< .001	-6.172	127	< .001
PI Scale	1.464	143	0.145	-2.174	127	0.032
CA Scale	5.979	143	< .001	9.994	127	< .001
CI Scale	-7.849	143	< .001	-11.795	127	< .001

Note. For the Student's t-tests for the PA and PI scales, the alternative hypothesis specified that the mean is different from 0.5. For the CA scale, the alternative hypothesis specified that the mean is different from 0, and, for the CI scale, that the mean is different from 1.

Study 1 Discussion

The goal of this study was to compare the intentionality bias towards humans and robots with a low or high anthropomorphism level. The time pressure condition was introduced as a between-subject factor, as it was shown (Rosset, 2008) to enhance the prevalence of intentionality bias. The tendency to anthropomorphize, empathy, the belief in human nature uniqueness, and the negative attitude toward robots were included as covariates. There were four types of behaviours judged: PA and PI (ambiguous behaviour but usually perceived as done by accident or on purpose) and CA and CI (not ambiguous, most people judge them as done by accident or on purpose). The types of behaviours that formed the scales were analyzed together and separately, only the latter, detailed analyses provided interesting insights.

The main hypothesis concerned the topic of intentionality bias prevalence in judging human and robot behaviour. According to Rosset (2008), the PA scale, so accidental behaviour judgments, should be the strongest signal. The author argues that if PA scenarios are interpreted more often as intentional under time pressure, they provide stronger evidence for the intentionality bias. As the behaviours were selected to be PA, one would expect a low rate of intentional interpretation in any condition if the bias was not there. If we take into consideration the accidental behaviours from the PA scale, the presented results confirm the hypothesis that an intentionality bias is present in the perception of robots' behaviours. An analysis of the results showed that the difference between robots and the human picture was not significant. Even the results for the CA scale, unambiguously accidental behaviours, from the present study, also showed an increased bias. There was a significant time pressure effect for the CA behaviours, almost always judged as accidental, which received more intentional attributions under time pressure for all agents. We could conclude that the intentionality bias is equal to or higher for the perception of robot behaviours than for human behaviours.

In the results reported above for the PA behaviours, it was the robot whose behaviour was interpreted as intentional more often than the human, both in the conditions with and without time pressure. Eisenkoeck and Moore (2017) made a human-robot comparison using Rosset's (2008) paradigm, and found a significant difference between PA intentionality scores between the human and robot conditions, where the robot scored higher. In addition, in Study I, the CA sentences were judged as more intentional for both robot conditions, but only robot 1 (i.e., the mechanical one) was significantly different from the human condition. Are we more prone to attributing intentions to robots when observing accidents or mistakes?

Possible explanations are available in the literature. As Eisenkoeck and Moore (2017) hypothesize, all PA sentences describe an action with a negative outcome, done by accident, possible to avoid, or simply a mistake. Other studies from the field of human-robot interactions suggest that we perceive unpredictability (Duffy, 2003) or making mistakes as a human-like quality. Salem et al., (2015) reported that participants perceived the robot's incongruent gestures as errors or 'imperfections', and it made the robot appear more humanlike and generally more likeable. This work concluded that a certain level of unpredictability or 'imperfection' can lead to better human-robot interactions, which is not an intuitive finding. It is plausible that an 'error-free' perception of robots and artificial intelligence played a role in the study presented above, based on the fact that participants did not expect mistakes from robots and attributed deliberation to the behaviours. It is also plausible that this deliberation means judging the behaviour from the design stance (Dennett, 1987), because of the way the machine has been built and programmed.

This might shed some light on the failed prediction that the robot with a high level of anthropomorphic features would receive more intentional attributions than the robot with a low level of anthropomorphic features. This hypothesis was not confirmed in the current study, as the robot score patterns were similar (except for the CA scale, where mechanical robots scored higher). It is possible that the general 'error-free' perception of robots and the deliberation behind mistakes played a role for both mechanical and humanoid agents, as they are perceived similarly, as a robot, a machine. An alternative explanation is that the experimental manipulation was too weak to differentiate them. Thellman et al. (2017) suggested that the experimental manipulation should be in at least the visual modality, and preferably more engaging (e.g., videos or real interactions with robots), which is a direction for further research.

According to the obtained results, if a humanoid robot's behaviour judgment follows the judgment of human behaviour, we would expect an even higher intentionality bias for the mechanistic robot for accidental behaviours, and a lower intentionality bias for intentional behaviours. This pattern can be seen in the current results; however, it was not significant. Therefore, future research on this matter would be of interest.

According to Rosset's (2008) dual-processing model, intentionality judgements made under time pressure should result in a higher intentionality bias, as these are the conditions that enhance biases and promote the use of heuristics. In the present study, the effect of time pressure was statistically significant for all the behaviour types, but surprisingly, for the PI and CI behaviours the relationship was opposite to the one expected based on the model. In particular, intentionality scores were lower in the time pressure condition than in the no time pressure (control) condition. This pattern of results does not confirm the first hypothesis that time pressure, in general, increases the likelihood of intentionality attributions. However, it does so for accidental behaviours. This hypothesis is tested again in Study II.

As for the covariates, the current study did not confirm a consistent relationship with judging intentionality based on Rosset's (2008) paradigm, with the exception of NATIR. Eisenkoeck and Moore (2017) also did not find the expected relationship between judging ambiguous behaviour and the IDAQ and QCAE scales.

A high NATIR score was positively related to the intentionality perceptions for accidental behaviours, but not for intentional ones. This boost in intentionality attributions appears to be in line with the perception of 'error-free' technology. The more positive our attitude towards interacting with robots, the more we do not assume mistakes and attribute deliberation, being 'error-free' to them.

Limitations

The intentionality measure is not a psychometrically derived instrument, and there may be more dimensions on which to evaluate the behaviours. One inspiration comes from a recent study of the intentionality bias in schizotypy (Roodenrys et al., 2021), where the researchers separated items that describe an action of one person in relation to another to create a 'social scale'. The results showed that the social scale correlated more strongly with the schizotypy measure than the nonsocial scale.

There may also be more evaluation needed to adjust this paradigm to human-robot interaction studies. De Graaf and Malle (2018) suggest that, in order to determine the differences between human and robot behavioural explanations, we need to establish whether people judge the basic properties of this behaviour similarly for both agents. The authors advise that human-robot interaction studies should include behaviours that have a baseline, general perception that is similar for people and robots in three dimensions, if they are intentional, surprising, and desired. The pilot study evaluated the baseline for the intentional aspect only, as this was the main concern of the project.

The manipulation of the level of anthropomorphic features of the robots may not have been strong enough. It would be beneficial to examine the same relationships for a stronger modality of manipulation (e.g., a video, an interaction with a robot in VR, or a real interaction), as advised by Thellman et al. (2017).

The test scales for measuring intentionality bias did have poor reliability, which was lower than in the pilot study, as the time pressure introduced in the experimental condition resulted in a substantial amount of missing data.

Study 2

Hypotheses

1. The intentionality bias (operationalized as the number of intention attributions to prototypically accidental behaviours under time pressure) will be stronger for a robot with a high level of anthropomorphic features than for a robot with a low level of anthropomorphic features.
2. Priming is related to the level of intentionality bias.
    a. Priming with robots with a high level of anthropomorphic features will positively influence the number of intention attributions to a mechanical robot.
    b. Priming with robots with a low level of anthropomorphic features will negatively influence the number of intention attributions to a human-like robot.
3. The level of cognitive empathy is related to the level of intention attribution to robots.
4. The tendency to anthropomorphize is related to the level of intention attribution to robots.
5. The negative attitude towards robots is related to the level of intention attribution to robots.
6. The belief in human nature uniqueness is related to the level of intention attribution to robots.

Participants

This study was conducted using the online panel Ariadna, a nationwide research panel. The sociodemographic profile of the individuals registered on Ariadna corresponds with the profile of Polish Internet users. Participants ($N = 235$; detailed characteristics in Table 9) were recruited from the general pool and randomly divided into groups corresponding to the experimental

conditions (conditions 1–6 from Table 10). Two groups ($N = 39$; conditions 7 and 8) were pooled from the previous study following the practice of integrative data analysis (Curran & Hussong, 2009). The procedures and outlook of both studies were exactly the same. Both used a random sample from the same online panel and were carried out two weeks apart.

Exposure to technology was controlled by questions about working or studying in related fields. In the sample, only 17 participants were related professionally to technology.

Table 9.

Descriptive statistics for the participants

			Education					
	n	Age Mean (SD)	Primary	Lower secondary	Upper secondary	Post-secondary	Bachelor's	Master's
Female	146	41 (14)	3.4%	3.4%	27.4%	17.1%	9.6%	39.0%
Male	126	44 (15)	0.8%	7.9%	38.1%	6.4%	8.7%	38.1%

Procedure

The study was a between-subject 2 (time pressure vs. control) by 2 (a highly anthropomorphic robot vs. a low anthropomorphic robot) by 2 (priming vs. no priming) experimental design. Table 10 summarises the conditions in the study.

Table 10.

Experimental conditions for Study II

Condition	Stimuli - Robot	Time pressure	Priming	N after data cleaning
1	High anthropomorphism level	Yes	Yes	28
2	Low anthropomorphism level	Yes	Yes	31
3	High anthropomorphism level	No	Yes	35
4	Low anthropomorphism level	No	Yes	38
5	High anthropomorphism level	No	No	38
6	Low anthropomorphism level	No	No	33
7	High anthropomorphism level	Yes	No	32
8	Low anthropomorphism level	Yes	No	37

This study was conducted online. The participants received a link from an online research panel to which they were subscribed. First, they were informed about the goal and rules of the study, asked to consent to participate in the study and about basic demographics. Next, the participants were shown a screen with the instructions and five practice test questions, which was especially important for groups with the 2.4 s time limit. The participants were then exposed to a screen indicating that the main study was loading, which took around 30 s. The participants were asked to watch a 23 s video during this time. The videos showed either a robot with a high level of anthropomorphic features or a low level of anthropomorphic features. As there was no autoplay feature, the log data indicating if the video was played or not were captured, and only participants who watched the video were included in the study. Next, the participants saw a

picture of one of the two stimuli (a robot with a high level of anthropomorphic features or a low level of anthropomorphic features) for 5 s. The 22 behaviours from the PA, PI, CA and CI scales were presented in random order, followed by a question asking whether the behaviour was done intentionally or by accident. This was followed by the administration of four questionnaires (NARS-PL, BHNU, QCAE, and IDAQ). The last questions concerned the current activities of the participants, such as learning or working in a specific domain, as the intention was to control exposure to technology in study groups, such as working as engineers.

Results

An 'intentionality endorsement score' (Rosset, 2008), defined as the number of 'on purpose' responses divided by the number of possible responses, was computed for each participant for each item on the test and control scales (PA, PI, CI, and CA).

Of the 304 participants, data from 24 participants were not included in the analysis because of missing more than 25% of the responses on one or more of the scales (PA ≤ 2 missing answers, PI ≤ 1, CA ≤ 1, CI ≤ 1). The missing values were due to the experimental conditions, as the time was limited to 2.4 s in some groups. An investigation for outliers (defined as the average score for the participant for a given scale being more than three standard deviations away from the group mean; standard deviations for scales were between 0.2 and 0.3) found 8 cases. The final sample was $N = 272$.

The reliability of the test and control scales was assessed using Cronbach's alpha. The reliability was poor or questionable (Ngulube, 2022), for PA, 9 items, $\alpha = .65$, PI, 5 items, $\alpha = .56$, CI, 4 items, $\alpha = .72$., and unacceptable for CA, 4 items, $\alpha = .45$. Thus, the interpretation of the CA outcome can be questioned because of the low reliability. The low reliability is most likely due to a large amount of missing data (Zhang & Yuan, 2016), as the experimental

condition restricted the time for answers to 2.4 s. The item-rest correlations for all items were higher than .02; therefore, the scales were accepted (Piedmont, 2014). All the scales measuring covariates were fairly reliable ($\alpha > .78$).

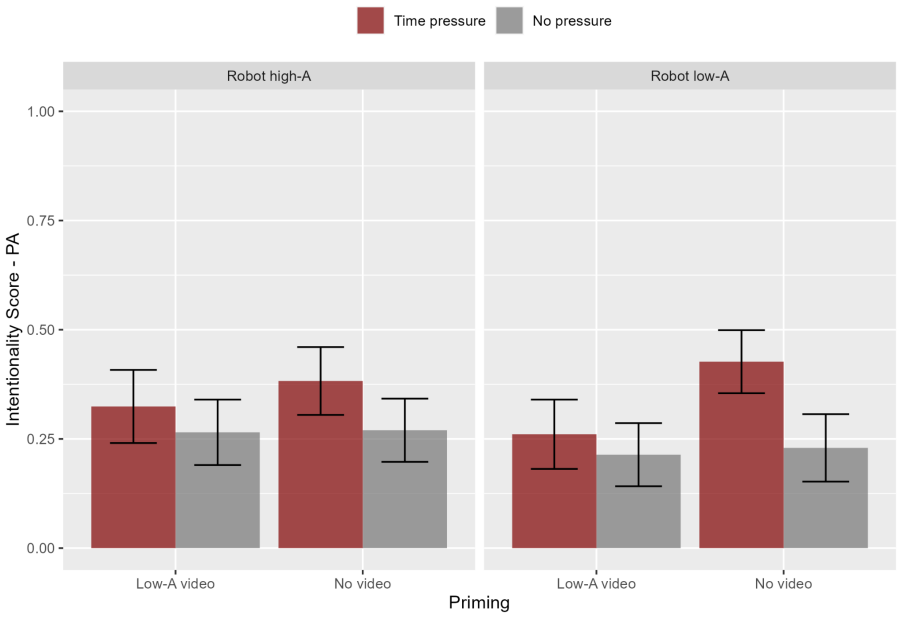
An ANCOVA analysis was conducted to examine the influence of agent type, time pressure, and priming on the intentionality scores for each of the four behaviour types (PA, PI, CA, CI), with a negative attitude towards robots (NATIR and NARHT subscales), a belief in human nature uniqueness (BHNU scale), empathy (Cognitive Empathy and Affective Empathy subscales) and the tendency to anthropomorphize (IDAQ and IDAQ-NA subscales) as covariates.

For the PA scale, the homogeneity of regression slopes assumption check revealed a significant interaction between the Affective Empathy subscale and the type of agent. Thus, this scale was not included as a covariate.

There were no significant interaction effects between the conditions in the main model, although the interaction between time pressure and priming was close to significance, $F(1, 258) = 3.394$, $p = .067$, $\eta^2 = .012$. There were significant main effects of time pressure, $F(1,258) = 13.677$, $p < .001$, $\eta^2 = .047$, and priming, $F(1,258) = 4.949$, $p = .027$, $\eta^2 = .017$. The intentionality scores were higher in the time pressure groups ($M = .353$) than in the control groups ($M = .244$), and the intentionality scores were lower in the priming groups ($M = .262$) than in the groups without priming ($M = .328$), as shown in Figure 10. None of the covariates were significantly related to the intentionality scores.

Figure 10.

The intentionality scores for prototypically accidental (PA) behaviours



Note. The figure presents the intentionality scores for prototypically accidental (PA) behaviours under the time pressure and control conditions without time pressure and with and without priming, for the two agent types; A: anthropomorphic features.

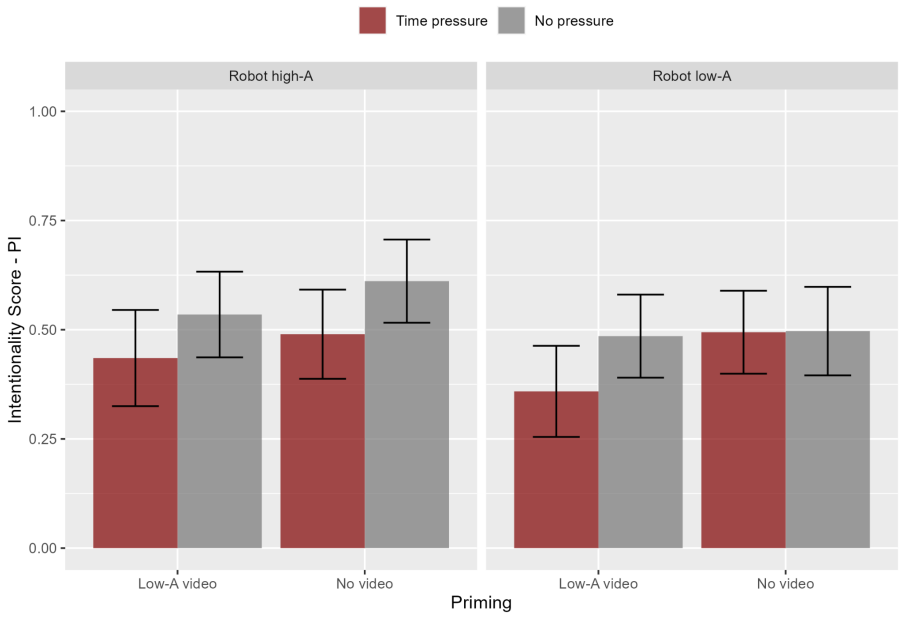
For the PI scale, the homogeneity of regression slopes assumption check revealed significant interactions between the Affective Empathy subscale and the BHNU with time pressure. Thus, these scales were not included as covariates.

There were no significant interaction effects between the conditions in the main model. However, there was a significant main effect of time pressure, $F(1,259) = 5.594$, $p = .019$, $\eta^2 = .020$. The priming effect was close to significance, $F(1,259) = 3.681$, $p = .057$, $\eta^2 = .013$. The intentionality scores were lower in the time pressure groups ($M = .447$) than in the control groups

(M = .533), as shown in Figure 11. None of the covariates were significantly related to the intentionality scores.

Figure 11.

The intentionality scores for prototypically intentional (PI) behaviours



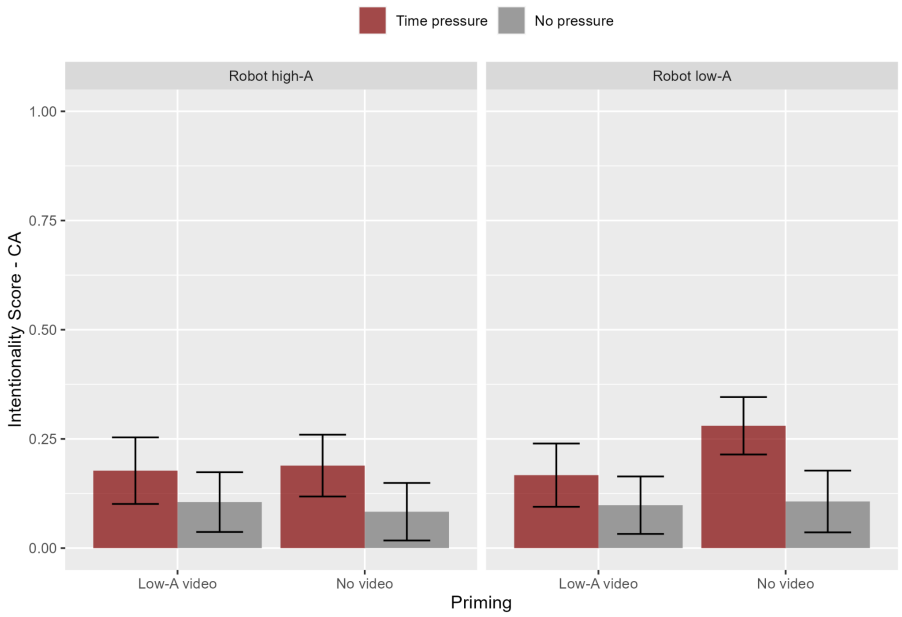
Note. The figure presents the intentionality scores for prototypically intentional (PI) behaviours under the time pressure and control conditions without time pressure and with and without priming, for the two agent types; A: anthropomorphic features.

For the CA scale, there were no significant interaction effects between the conditions in the main model. However, there was a significant main effect of time pressure, $F(1,257) = 16.468$, $p < .001$, $\eta^2 = .054$. The intentionality scores were higher in the time pressure groups ($M = .207$) than in the control groups ($M = .097$), as shown in Figure 12.

The NATIR subscale was significantly related to the intentionality scores on the CA scale, $F(1,257) = 12.150$, $p < .001$, $\eta^2 = .040$. The tendency to anthropomorphize (IDAQ) was also significantly related to the intentionality scores on the CA scale, $F(1,257) = 8.146$, $p = .005$, $\eta^2 = .027$.

Figure 12.

The intentionality scores for control accidental (CA) behaviours



Note. The figure presents the intentionality scores for control accidental (CA) behaviours under the time pressure and control conditions without time pressure and with and without priming, for the two agent types; A: anthropomorphic features.

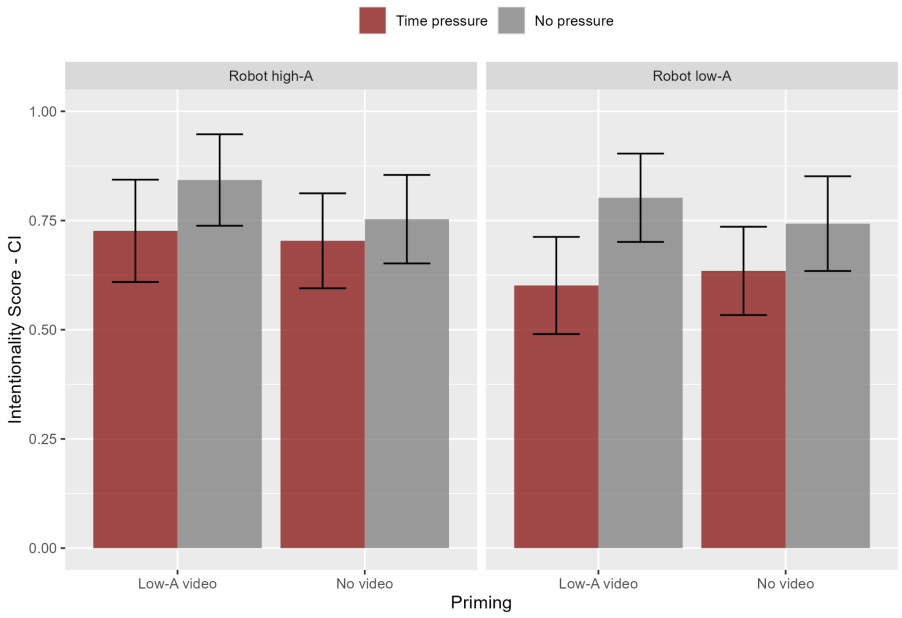
For the CI scale, the homogeneity of regression slopes assumption check revealed a significant interaction between the BHNU and time pressure. Thus, this scale was not included as a covariate.

There were no significant interaction effects between the conditions in the main model. However, there was a significant main effect of time pressure, $F(1,258) = 8.918$, $p = .003$, $\eta^2 = .030$. The intentionality scores were lower in the time pressure groups ($M = .664$) than in the control groups ($M = .785$), as shown in Figure 13.

The tendency to anthropomorphize (IDAQ) was significantly related to the intentionality scores on the CI scale, $F(1,258) = 9.033$, $p = .003$, $\eta^2 = .031$, as well as the IDAQ-NA measuring non-anthropomorphic attributions, $F(1,258) = 7.832$, $p = .006$, $\eta^2 = .027$.

Figure 13.

The intentionality scores for control intentional (CI) behaviours



Note. The figure presents the intentionality scores for control intentional (CI) behaviours under the time pressure and control conditions without time pressure and with and without priming, for the two agent types; A: anthropomorphic features.

Study 2 Discussion

The goal of this study was to evaluate whether priming can alter the perception of the robots and influence intentionality bias. The stimulus included two robots with a low or high level of anthropomorphic features. Priming in the opposite direction of the level of the anthropomorphic features of the robot (a mechanical robot was primed with a humanoid and vice versa) was introduced as a between-subjects factor. A time pressure condition was introduced as a between-subjects factor, as a bias enhancer (Rosset, 2008). The tendency to anthropomorphize,

empathy, belief in human nature uniqueness, and the negative attitude towards robots were included as covariates.

The pattern for time pressure was replicated, as in Study I, where this factor increased the intentionality bias for accidental behaviours and decreased it for intentional behaviours.

There was no significant effect of the type of robot; thus, the level of anthropomorphic features of the robot did not make a difference. Therefore, there is no evidence to confirm the first hypothesis. It is possible that the experimental manipulation was too weak to differentiate these effects, or that the level of anthropomorphic features did not matter.

The prediction based on the second hypothesis concerning the priming effect on intentionality bias appears for the PA scale. It shows a similar pattern to the results of Study I, where the PA scale is the one where the bias effect has the most 'room' to appear. Also in the literature (e.g., Rosset, 2008), PA behaviours are treated as the main bias measure.

The sub-hypotheses of the second hypothesis regarding the priming effects were not confirmed. Priming with an opposite level of anthropomorphic features decreased the intentionality scores for both robots, but especially for the robot with a low level of anthropomorphic features primed with the robot with a high level of anthropomorphic features. As observed in Study I mechanistic robots can receive more intentional attributions, potentially because of the perception of being error or mistake-free. In this context, the priming with anthropomorphic robots could affect this perception.

General discussion

The main goal of this research line was to determine whether the intentionality bias is specific to the perception of other humans or whether it is also present in the perception of robots' behaviours. In addition, this line of studies aimed to determine whether a robot's

anthropomorphic features affect intentionality attributions. The tendency to anthropomorphize, empathy, negative attitude towards robots, and the belief in human nature uniqueness were included as covariates. Moreover, this project looked at the possibility of influencing a robot's intentionality perception by employing priming to steer the perception of a robot.

Considering the intentionality bias enhancement by time pressure, the results did not fully follow the predictions, although an interesting pattern emerged. Accidental behaviour judgments were aligned with the model predictions, and the intentionality score was higher in the time pressure conditions. However, the effect was reversed for intentional behaviours, for both the test and control scales. The same pattern was obtained in both of the reported studies. The literature provides evidence that the type of behaviour changes the time pressure effect, specifically that the effect is not present or is reversed for intentional scales, as discussed in Study I.

In the original studies (Rosset, 2008), the difference between time and no pressure conditions for the PI scale was insignificant. Rather, the paper focused on 'general intentionality', which was a proportion of all the intentional attributions for the test scales (i.e., PA and PI) divided by the number of all possible answers for those scales, which was significant between conditions. In the results presented in both studies for the PI scale, the expected effect was reversed: under time pressure, the intentionality attribution was lower. As shown by the interaction effect, this effect was significant for the human condition. The PI scale score in the original Rosset study did not differ between the time pressure and control conditions, and other studies are reporting similar results.

For example, Hughes et al. (2012) aimed to replicate Rosset's (2008) results and, while the results were replicated, this was only when looking at the global intentionality score for all of the scales collapsed. A more detailed examination showed that the intentionality score did not differ

between the time pressure and control conditions for PI and PA behaviors. The CI scale, on the other hand, had a greater intentionality score in the conditions without time pressure, similar to the results of the presented studies. Also, the pattern for the CA scale is in line with the current results, as the intentionality score was higher in the time pressure condition. Researchers also measured the reaction time when judging the intentionality of behaviours from Rosset's paradigm. Taking into consideration the dual-processing model, it should take longer to respond to unintentional than intentional sentences. This should be the case under both conditions, with and without time pressure, and for both the test and control scales. In line with this prediction, participants responded faster to the control intentional scale than to the accidental one. However, they were slower to respond to the test intentional scale, PI, than to the accidental one, PA.

There are other examples from the literature where the PA and PI scales behaved differently. For example, Slavny and Moore (2018) investigated whether individual differences in the intentionality bias are related to emotional and cognitive empathy using Rosset's paradigm (2008). They found an association between cognitive empathy, specifically the perspective-taking ability, and prototypically accidental behaviours, the PA scale, but not PI. The reason for this is unclear, but researchers debate that the ambiguous but mostly accidental scenarios are a good context for the intentional bias to reveal. To rephrase this, an accidental setting seems to be more sensitive to detecting bias.

Similarly, Subra (2021) investigated a relationship between anger and intentionality bias measured with Rooset's paradigm. The results show that anger increases intentional attributions for accidental test sentences, PA scale, but not for intentional test sentences, PI. The results were also significant for the CA scale, where behaviours are almost always judged as accidental. Both accidental scales behaved consistently, which is similar to the studies presented in this thesis.

To summarize, in the literature, the pattern of results for prototypically intentional behaviours in the PI scale is either not significant or lower under time pressure, and, thus, the presented studies add to the previous evidence, contradictory to the prediction made based on the dual-processing model (Rosset, 2008). How should we understand this discrepancy between the different behaviour types? One possible explanation is that these findings may suggest a different mechanism responsible for the bias.

The core of Rosset's proposal (2008) is that the intentional interpretation of actions is automatically activated. It is plausible to interpret events as accidental, but this requires additional effort. Rosset proposes that this is based not on intentional inference but on intentional inhibition. As many studies have focused on the developmental aspect of intention attributions, Rosset and Rottman (2014) reviewed developmental stages, from infancy and an early sensitivity to intentions to adulthood and over-attributing intentions to the behaviour of others. The conclusions support the statement that people become increasingly skilled at understanding different causes of behaviour not by mastering how to attribute intentions but rather by becoming increasingly skilled at inhibiting the default interpretation. Other studies (e.g., Donovan & Kelemen, 2011) confirm that young children display a strong bias toward intentional explanation, and Rosset provides evidence of this bias in adults when their inhibitory skills might be impaired, either by time pressure (2008) or alcohol consumption (2010). Another example of data supporting the model comes from research on intentionality bias in schizophrenia (Peyroux et al., 2014). The authors report increased intentionality bias in researched population. The overlap in responses between people with schizophrenia, adults under time pressure, and young children supports that impaired inhibitory abilities may be the mechanism for intentionality bias in all three populations.

Alternatively, the incoherent results can be caused by way of testing. The paradigm can result in different outcomes for specific behaviours. One of the hypotheses is that negative behaviours are judged differently because of the potential consequences for others. Indeed, many ambiguous behaviours used in the paradigm could result in negative outcomes for others. Subra (2021) tested this hypothesis but found no differences in intentionality attributions. In addition, the effect of anger on intentionality bias was not significantly different depending on those types of test behaviours (different consequences for others). Another counterargument to this explanation comes from Monroe et al. (2015), who tested the potential influence of negative outcomes for others in used behaviours, calling it a moral valance, and their results did not confirm this hypothesis.

Another possible explanation could be the semantic layer of the used procedure. As reported by Strickland et al., (2011), under time pressure, people have an enhanced bias to attribute more intentions to grammatical subjects (in the ‘avoir’ verbs in French). Such a bias enhancement was not observed for grammatical objects moved to the subject position in the structure (in the ‘etre’ verbs). The intentional bias was measured using both declaration and time reactions. This might suggest that the language used could be a reason for the PI scale to behave inconsistently with the model. It is worth further investigation, nevertheless, there are studies confirming intentionality bias using a different stimulus modality. For example, Moore and Pope (2014), in their study on people with schizophrenia, reported intentionality bias using video stimuli depicting the movement of a hand on a keyboard.

Yet another angle to view the results of the presented studies could be inspired by the work of Monroe et al. (2015). It distinguishes between low-level behaviours, guided by simple cues, and high-level behaviours, more abstract, influenced by judgment on motives and social

information. According to researchers, low-level behaviours should presumably be less or not affected by cognitive load. In contrast, high-level behaviours judgment should be affected by the load as it relies more on social cues, as well as the linguistic aspect of the description. Although most of the behaviours used in Rosset's paradigm fit the description of low-level behaviours, their reliance on social perception and cognitive load should be further investigated.

Another interesting line of thought comes from Moore and Pope (2014), who explore the potential negative outcomes of ambiguous behaviours. Due to that, we might be equipped to tolerate false positives more than false negatives as an evolutionary development. Abstracting from the context of a potential outcome to us or others, assessing accidental behaviour as intentional might be a smart protecting strategy, but it loses its importance for mostly intentional events.

The above examples from the literature are in line with those presented in this thesis, providing a possible future direction for investigating the results obtained using the PI scale. Interestingly, this pattern was replicated in Study II, where only robots were presented. Whatever the cause of this pattern, it applies to humans and robots alike. Despite this discrepancy, in general, the intentionality bias prevalence has been replicated in numerous studies with different stimulus modalities.

Interestingly, for the PA scale in Study I, both robots' behaviour were scored as more intentional than human behaviours. As this scale shows the strongest signal of intentionality bias, robot behaviour judgments are prone to this bias, even more so than human behaviour. A similar pattern of results to those reported in this thesis, where the intentionality bias was higher for robots for the main PA scale but not for the PI sentences, was obtained in the human-robot-interaction context by Eisenkoeck and Moore (2017). Researchers did not employ

time pressure but rather compared intentionality attributions for different scales. Also, the level of robot's anthropomorphism did not play a role, as the stimulus was a simple vignette mentioning a robot.

Study I's discussion provided several possible explanations for this effect. Eisenkoeck and Moore (2017) hypothesize that all PA sentences describe an action with a negative outcome, done by accident, possible to avoid, or simply a mistake. The robots may be perceived as error-free; therefore, mistakes are not expected. It is plausible that an 'error-free' perception of robots and artificial intelligence played a role in the results presented above, based on the fact that participants did not expect mistakes from robots and attributed deliberation to the behaviours. It is also plausible that this deliberation might also mean judging the behaviour from the design stance (Dennett, 1987), because of how the machine has been built and programmed.

There were no differences between robots with a high or low level of anthropomorphic features in Study I, which may suggest that the experimental manipulation was too weak or that the general 'error-free' perception of robots and the deliberation behind mistakes played a role for both mechanical and humanoid agents. In Study II, where priming with a robot with the opposite level of anthropomorphic features was used, the effect of priming was significant for the main PA scale and close to significance for the other test scale, PI. Priming with an opposite level of anthropomorphic features decreased the intentionality scores for both robots, but especially for the robot with a low level of anthropomorphic features, as the interaction between priming and the type of robot was close to significance. These results follow the potential interpretation from Study I, that the 'error-free' perception of a robot makes participants attribute more intentions to its behaviours, or, rather, deliberate, programmed behaviours. This is potentially why exposure to a more human-like robot as a priming stimulus effects in a drop in

intentional attributions to a robot with a low level of anthropomorphic features, as it appears less ‘programmed’ in contrast to the more complex machine. It is also plausible that the video stimuli decreased the intentionality bias, as exposing participants to moving, operating, and performing actions robots made them appear less like an error-free machine in general. To summarise, the hypothesis concerning priming stated before the study did not find confirmation, but the results show that the video stimuli had a stronger effect on intentionality perception than just the pictures.

To summarize, considering the PA scale only, the data showed that the intentionality bias is present for robots at the same level or higher than humans and that priming should be further investigated to influence it. Further research is needed to evaluate the influence of a mental model of a robot as a deliberate machine and its influence on the user experience of the new class of social robots.

Further research

One of the directions for further investigation concerns the level of anthropomorphism or how engaging the stimuli are. Thellman et al. (2017) suggest that the experimental manipulations in human-robot interaction studies should be in at least the visual modality, and preferably engaging like real interactions with robots, which provides a direction for future research.

The presented results showed a difference between predictions from the Rosset’s model (2008) and judging intentional behaviours. As there is no psychometrically derived instrument available for research on this topic, further investigation should evaluate the different properties of the behaviours, such as the semantic layer, the level of behaviour, or social aspects of it (Roodenrys et al., 2021).

The results showed that robot behaviours get at least an equal number of intentional attributions as humans. Whether our judgment of robot behaviour is an automatic bias or a deliberate process, the effect it has on our interactions with these machines can be substantial. Assuming that a robot makes a mistake deliberately can break the interaction quality and prevent the adoption of this technology. Research on intentionality perception seems vital for the sake of a good experience when utilizing social robots.

Literature

Abbott, A. (2013). Disputed results a fresh blow for social psychology. *Nature*, *497*(7447), 16–16. https://doi.org/10.1038/497016a

Air Force Magazine. (2014, October 27). *Meet saul, the ebola-zapping robot*. Air Force Magazine. Retrieved January 18, 2022, from https://www.airforcemag.com/meet-saul-the-ebola-zapping-robot/

Allison, T., Puce, A., & McCarthy, G. (2000). Social perception from visual cues: role of the STS region. In Trends in Cognitive Sciences (Vol. 4, Issue 7, pp. 267–278). Elsevier BV. https://doi.org/10.1016/s1364-6613(00)01501-1

Allport, G. W. (1935). Attitudes. In A Handbook of Social Psychology (pp. 798–844). Clark University Press.

Axelrod, L., & Hone, K. (2005). Uncharted Passions: User Displays of Positive Affect with an Adaptive Affective System. *Affective Computing and Intelligent Interaction*, 890–897. https://doi.org/10.1007/11573548_114

Bandoim, L. (2020, March 31). *How robots are helping grocery stores during the coronavirus outbreak*. Forbes. Retrieved January 18, 2022, from https://www.forbes.com/sites/lanabandoim/2020/03/30/how-robots-are-helping-grocery-stores-during-the-coronavirus-outbreak/?sh=5ca73df2242a

Bargh, J. A., Chen, M., & Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology*, *71*(2), 230–244. https://doi.org/10.1037/0022-3514.71.2.230

Bartneck, C., Kulić, D., Croft, E., & Zoghbi, S. (2008). Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of

Robots. *International Journal of Social Robotics*, *1*(1), 71–81. https://doi.org/10.1007/s12369-008-0001-3

Bartneck, C., Nomura, T., Kanda, T., Suzuki, T., & Kennsuke, K. (2005). A cross-cultural study on attitudes towards robots.

Bègue, L., Bushman, B.J., Giancola, P.R., Subraand, B., Rosset, E. (2010). 'There is no such thing as an accident,' especially when people are drunk. Personality and social psychology bulletin, 36(10), pp.1301-1304. https://doi.org/10.1177/0146167210383044

Belpaeme, T., Kennedy, J., Ramachandran, A., Scassellati, B., & Tanaka, F. (2018). Social robots for education: A review. In Science Robotics (Vol. 3, Issue 21). American Association for the Advancement of Science (AAAS). https://doi.org/10.1126/scirobotics.aat5954

Malle, B. F., & Hodges, S. D. (2006). Other minds: how humans bridge the divide between self and others. *Choice Reviews Online*, *43*(05), 43–3084. https://doi.org/10.5860/choice.43-3084

Birks, M., Bodak, M., Barlas, J., Harwood, J., & Pether, M. (2016). Robotic Seals as Therapeutic Tools in an Aged Care Facility: A Qualitative Study. In Journal of Aging Research (Vol. 2016, pp. 1–7). Hindawi Limited. https://doi.org/10.1155/2016/8569602

Blakemore, S.-J., & Decety, J. (2001). From the perception of action to the understanding of intention. In Nature Reviews Neuroscience (Vol. 2, Issue 8, pp. 561–567). Springer Science and Business Media LLC. https://doi.org/10.1038/35086023

Bossi, F., Willemse, C., Cavazza, J., Marchesi, S., Murino, V., & Wykowska, A. (2020). The human brain reveals resting state activity patterns that are predictive of biases in attitudes toward robots. In Science Robotics (Vol. 5, Issue 46). American Association for the Advancement of Science (AAAS). https://doi.org/10.1126/scirobotics.abb6652

Breazeal, C., Dautenhahn, K., & Kanda, T. (2016). Social robotics. *Springer handbook of robotics*, 1935-1972.

Cabibihan, J.-J., Javed, H., Ang, M., Jr., & Aljunied, S. M. (2013). Why Robots? A Survey on the Roles and Benefits of Social Robots in the Therapy of Children with Autism. In International Journal of Social Robotics (Vol. 5, Issue 4, pp. 593–618). Springer Science and Business Media LLC. https://doi.org/10.1007/s12369-013-0202-2

Castelli, F. (2002). Autism, Asperger syndrome and brain mechanisms for the attribution of mental states to animated shapes. *Brain*, *125*(8), 1839–1849. https://doi.org/10.1093/brain/awf189

Céspedes, N., Raigoso, D., Múnera, M., & Cifuentes, C. A. (2021). Long-Term Social Human-Robot Interaction for Neurorehabilitation: Robots as a Tool to Support Gait Therapy in the Pandemic. In Frontiers in Neurorobotics (Vol. 15). Frontiers Media SA. https://doi.org/10.3389/fnbot.2021.612034

Chaminade, T., Rosset, D., Da Fonseca, D., Nazarian, B., Lutcher, E., Cheng, G., & Deruelle, C. (2012). How do we think machines think? An fMRI study of alleged competition with an artificial intelligence. *Frontiers in Human Neuroscience*, *6*. https://doi.org/10.3389/fnhum.2012.00103

Chivers, T. (2019). What's next for psychology's embattled field of social priming. *Nature*, *576*(7786), 200–202. https://doi.org/10.1038/d41586-019-03755-2

Costa, R. (2019). The design thinking process for better UX design. Retrieved September 22, 2021 from https://www.justinmind.com/blog/design-thinking-process-ux-design/.

Cullen, H., Kanai, R., Bahrami, B., & Rees, G. (2013). Individual differences in anthropomorphic attributions and human brain structure. Social cognitive and affective neuroscience, 9(9), 1276-80. https://doi.org/10.1093/scan/nst109

Curran, P. J., & Hussong, A. M. (2009). Integrative data analysis: the simultaneous analysis of multiple data sets. *Psychological Methods*, *14*(2), 81–100. https://doi.org/10.1037/a0015914

Damiano, L., & Dumouchel, P. (2018). Anthropomorphism in Human-Robot Co-evolution. Frontiers in psychology, 9, 468. https://doi.org/10.3389/fpsyg.2018.00468

Darling, K. (2020, October 21). *Problemas de Ia E robótica, de kate darling*. Iberdrola. Retrieved January 18, 2022, from https://www.iberdrola.com/shapes-en/kate-darling-robotics-artificial-intelligence-problems

Darling, K. (2012). Extending legal rights to social robots. *SSRN Electronic Journal*. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2044797

Dautenhahn, K. (2007). Methodology & themes of human-robot interaction: A growing research field. *International Journal of Advanced Robotic Systems*, 4(10), doi: 10.5572/5702.

Davern, Michael & Shaft, Teresa & Te'eni, Dov. (2012). Cognition Matters: Enduring Questions in Cognitive IS Research. Journal of the Association for Information Systems. 13. https://doi.org/10.17705/1jais.00290

De Graaf, M. M. A., & Malle, B. F. (2018). People's Judgments of Human and Robot Behaviors: A Robust Set of Behaviors and Some Discrepancies. *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 97–98. Presented at the Chicago, IL, USA. doi:10.1145/3173386.3177051

Demoulin, S., Leyens, J. P., & Yzerbyt, V. (2006). Lay Theories of Essentialism. *Group Processes &Amp; Intergroup Relations*, *9*(1), 25–42. https://doi.org/10.1177/1368430206059856

Dennett, D. (1987). The Intentional Stance. Cambridge, MA: MIT Press.

Donovan, E., & Kelemen, D. (2011). Just Rewards: Children and Adults Equate Accidental Inequity with Intentional Unfairness. *Journal of Cognition and Culture*, *11*(1–2), 137–150. https://doi.org/10.1163/156853711x568725

Doyen, S., Klein, O., Pichon, C. L., & Cleeremans, A. (2012). Behavioral Priming: It's All in the Mind, but Whose Mind? *PLoS ONE*, *7*(1), e29081. https://doi.org/10.1371/journal.pone.0029081

Duffy, B. R. (2003). Anthropomorphism and the social robot. *Robotics and Autonomous Systems*, *42*(3–4), 177–190. https://doi.org/10.1016/s0921-8890(02)00374-3

Eisenkoeck, A., & Moore, J. (2017). Differences in the Intentionality Bias when Judging Human and Robotic Action.

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three–factor theory of anthropomorphism. Psychological Review, 114, 864–886. https://doi.org/10.1037/0033-295X.114.4.864

Epley, N., Waytz, A., Akalis, S., & Cacioppo, J. T. (2008). When we need a human: Motivational determinants of anthropomorphism. Social Cognition, 26(2), 143-155. https://doi.org/10.1521/soco.2008.26.2.143

Eyssel, F. (2022, March 7-10). What's Social about Social Robots? A Psychological Perspective. [Keynote address]. ACM/IEEE International Conference on Human-Robot Interaction (HRI) 2022

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

Fogg, B., & Nass, C. (1997). How users reciprocate to computers. *CHI '97 Extended Abstracts on Human Factors in Computing Systems Looking to the Future - CHI '97*. https://doi.org/10.1145/1120212.1120419

Fong, T., Nourbakhsh, I., & Dautenhahn, K. (2003). A survey of socially interactive robots. Robotics and Autonomous Systems, 42(3-4), 143–166. https://doi.org/10.1016/s0921-8890(02)00372-x

Forbes Magazine. (n.d.). *The 2020 world's most valuable brands*. Forbes. Retrieved January 17, 2022, from https://www.forbes.com/powerful-brands/list/

Fujita, M. (2001). AIBO: Toward the Era of Digital Creatures. In The International Journal of Robotics Research (Vol. 20, Issue 10, pp. 781–794). SAGE Publications. https://doi.org/10.1177/02783640122068092

Gambino, A., Fox, J., & Ratan, R. (2020). Building a Stronger CASA: Extending the Computers Are Social Actors Paradigm. In Human-Machine Communication (Vol. 1, pp. 71–86). Nicholson School of Communication, UCF. https://doi.org/10.30658/hmc.1.5

Giger, J.-C., Piçarra, N., Alves, Oliveira, P., Oliveira, R., & Arriaga, P. (2019). Humanization of robots: Is it really such a good idea? In Human Behavior and Emerging Technologies (Vol. 1, Issue 2, pp. 111–123). Wiley. https://doi.org/10.1002/hbe2.147

Giger, J.-C., Moura, D., Almeida, N., & Piçarra, N. (2017). *Attitudes towards Social Robots: The Role of Belief in Human Nature Uniqueness, Religiousness and Taste for Science Fiction*.

Graham, C. (2021, March 30). *Ready for duty: Health care robots get good prognosis for next pandemic*. The Hub. Retrieved January 18, 2022, from https://hub.jhu.edu/2021/03/30/robots-in-health-care-settings-challenges-advantages/

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of Mind Perception. Science, 315(5812), 619–619. 10.1126/science.1134475

Hanson, D. (2021, June 24). *Why we should build humanlike robots*. IEEE Spectrum. Retrieved January 18, 2022, from https://spectrum.ieee.org/why-we-should-build-humanlike-robots

Harrison, M. A., & Hall, A. E. (2010). Anthropomorphism, empathy, and perceived communicative ability vary with phylogenetic relatedness to humans. *Journal of Social, Evolutionary, and Cultural Psychology*, *4*(1), 34–48. https://doi.org/10.1037/h0099303

Haslam, N., & Whelan, J. (2008). Human Natures: Psychological Essentialism in Thinking about Differences between People. *Social and Personality Psychology Compass*, *2*(3), 1297–1312. https://doi.org/10.1111/j.1751-9004.2008.00112.x

Haslam, N., Rothschild, L., & Ernst, D. (2000). Essentialist beliefs about social categories. *British Journal of Social Psychology*, *39*(1), 113–127. https://doi.org/10.1348/014466600164363

Hassenzahl, M. (2008). User experience (UX): Towards an experiential perspective on product quality. IHM '08: Proceedings of the 20th Conference on l'Interaction Homme-Machine.

Heider, F., & Simmel, M. L. (1944). An Experimental Study of Apparent Behavior. *American Journal of Psychology*, *57*(2), 243. https://doi.org/10.2307/1416950

Hughes, J. S., Sandry, J., & Trafimow, D. (2012). Intentional inferences are not more likely than unintentional ones: some evidence against the intentionality bias hypothesis. The Journal of Social Psychology, 152(1), 1–4. https://doi.org/10.1080/00224545.2011.565383

IEEE. (2018, May 18). *Pleo*. ROBOTS. Retrieved January 19, 2022, from https://robots.ieee.org/robots/pleo/

International Federation of Robotics Frankfurt (2021), Robots in Daily Life – the positive impact of robots on wellbeing.

International Organization for Standardization (2010). Ergonomics of human-system interaction – Part 210: Human-centered design for interactive systems (formerly known as 13407). ISO 9241-210:2010. Retrieved September 22, 2021 from https://www.iso.org/standard/52075.html.

International Organization for Standardization (2018). Ergonomics of human-system interaction – Part 11: Usability: Definitions and concepts. ISO DIS 9241-11:2015. Retrieved September 22, 2021 from https://www.iso.org/standard/63500.html (22.09.2021).

International Organization for Standardization [ISO] (2014). Robots and Robotic Devices – Safety Requirements for Personal care Robots. Avaliable at: https://www.iso.org/obp/ui/#iso:std:iso:13482:ed-1:v1:en (accessed November 28, 2021).

JASP Team (2022). JASP (Version 0.16.3)[Computer software].

Kahneman, D. (2011). Thinking, fast and slow. New York, NY, US: Farrar, Straus and Giroux.

Knemeyer, D. (2015). Design Thinking and UX: Two sides of the same coin. Retrieved September 21, 2021 from interactions.acm.org.

Komatsu, T., Kurosawa, R., & Yamada, S. (2011). How Does the Difference Between Users' Expectations and Perceptions About a Robotic Agent Affect Their Behavior? In International Journal of Social Robotics (Vol. 4, Issue 2, pp. 109–116). Springer Science and Business Media LLC. https://doi.org/10.1007/s12369-011-0122-y

Koverola, M., Kunnari, A., Sundvall, J., & Laakasuo, M. (2022). General Attitudes Towards Robots Scale (GAToRS): A New Instrument for Social Surveys. *International Journal of Social Robotics*, *14*(7), 1559–1581. https://doi.org/10.1007/s12369-022-00880-3

KPMG. (2016). *Social robots.* https://assets.kpmg/content/dam/kpmg/pdf/2016/06/social-robots.pdf

Kupferberg, A., Glasauer, S., Huber, M., Rickert, M., Knoll, A., & Brandt, T. (2011). Biological movement increases acceptance of humanoid robots as human partners in motor interaction. In AI & SOCIETY (Vol. 26, Issue 4, pp. 339–345). Springer Science and Business Media LLC. https://doi.org/10.1007/s00146-010-0314-2

Lasota, A., Tomaszek, K., & Bosacki, S. (2020). How to become more grateful? The mediating role of resilience between empathy and gratitude. *Current Psychology*, *41*(10), 6848–6857. https://doi.org/10.1007/s12144-020-01178-1

Lee, H. R., Cheon, E., de Graaf, M., Alves-Oliveira, P., Zaga, C., & Young, J. (2019). Robots for Social Good: Exploring Critical Design for HRI. 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI). IEEE. https://doi.org/10.1109/hri.2019.8673130

Leiberg, S., & Anders, S. (2006). The multiple facets of empathy: a survey of theory and evidence. *Understanding Emotions*, 419–440. https://doi.org/10.1016/s0079-6123(06)56023-6

Levillain, F., & Zibetti, E. (2017). Behavioral Objects: The Rise of the Evocative Machines. *Journal of Human-Robot Interaction*, *6*(1), 4. https://doi.org/10.5898/jhri.6.1.levillain

Leyzberg, D., Spaulding, S., Toneva, M., & Scassellati, B. (2012). The Physical Presence of a Robot Tutor Increases Cognitive Learning Gains. *Cognitive Science*, *34*(34). http://scazlab.yale.edu/sites/default/files/files/Leyzberg-Cogsci-12.pdf

Looser, C. E., & Wheatley, T. (2010). The Tipping Point of Animacy. *Psychological Science*, *21*(12), 1854–1862. https://doi.org/10.1177/0956797610388044

Makowski, D., (2018). 'The Psycho Package: An Efficient and Publishing-Oriented Workflow for Psychological Science.' *Journal of Open Source Software*, **3**(22), 470. doi:10.21105/joss.00470, R package, https://github.com/neuropsychology/psycho.R.

Malle, B. F., & Knobe, J. (1997). The Folk Concept of Intentionality. *Journal of Experimental Social Psychology*, *33*(2), 101–121. https://doi.org/10.1006/jesp.1996.1314

Marchesi, S., Spatola, N., Perez-Osorio, J., & Wykowska, A. (2021). Human vs Humanoid. A Behavioral Investigation of the Individual Tendency to Adopt the Intentional Stance. *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*. https://doi.org/10.1145/3434073.3444663

Marchesi, S., Ghiglino, D., Ciardo, F., Perez-Osorio, J., Baykara, E., & Wykowska, A. (2019). Do We Adopt the Intentional Stance Toward Humanoid Robots? *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.00450

Marcinkowski, T., & Reid, A. (2019). Reviews of research on the attitude–behavior relationship and their implications for future environmental education research. In Environmental Education Research (Vol. 25, Issue 4, pp. 459–471). Informa UK Limited. https://doi.org/10.1080/13504622.2019.1634237

Coeckelbergh, M. (2011). Talking to Robots: On the Linguistic Construction of Personal Human-Robot Relations. *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, 126–129. https://doi.org/10.1007/978-3-642-19385-9_16

Malkewitz, C. P., Schwall, P., Meesters, C., & Hardt, J. (2023). Estimating reliability: A comparison of Cronbach's α, McDonald's ωt and the greatest lower bound. *Social Sciences &Amp; Humanities Open*, *7*(1), 100368. https://doi.org/10.1016/j.ssaho.2022.100368

Marr, B. (2020, March 18). *Robots and drones are now used to fight COVID-19*. Forbes. Retrieved January 18, 2022, from https://www.forbes.com/sites/bernardmarr/2020/03/18/how-robots-and-drones-are-helping-to-fight-coronavirus/?sh=338c547c2a12

Matsumoto, N., Fujii, H., Goan, M., & Okada, M. (2005). Minimal design strategy for embodied communication agents. *Robot and Human Interactive Communication*. https://doi.org/10.1109/roman.2005.1513801

McGinn, C., Bourke, E., Murtagh, A., Donovan, C., Lynch, P., Cullinan, M. F., & Kelly, K. (2019). Meet Stevie: a Socially Assistive Robot Developed Through Application of a 'Design-Thinking' Approach. *Journal of Intelligent &Amp; Robotic Systems*, *98*(1), 39–58. https://doi.org/10.1007/s10846-019-01051-9

McNamara, T.P. (2005). Semantic Priming: Perspectives from Memory and Word Recognition. Semantic Priming: Perspectives from Memory and Word Recognition. https://doi.org/10.4324/9780203338001

Medin, D., Ortony, A. (1989). Psychological essentialism. Similarity and analogical reasoning. 179-196.

Monroe, A. E., Reeder, G. D., & James, L. (2015). Perceptions of Intentionality for Goal-Related Action: Behavioral Description Matters. *PLOS ONE*, *10*(3), e0119841. https://doi.org/10.1371/journal.pone.0119841

Moore, J. W., & Pope, A. (2014). The intentionality bias and schizotypy. *Quarterly Journal of Experimental Psychology*, *67*(11), 2218–2224. https://doi.org/10.1080/17470218.2014.911332

Moosa, M. M., & Ud-Dean, S. M. M. (2010). Danger Avoidance: An Evolutionary Explanation of Uncanny Valley. *Biological Theory*, *5*(1), 12–14. https://doi.org/10.1162/biot_a_00016

Mori, M. (1970). Bukimi no tani [the uncanny valley]. Energy, 7, 33-35.

Mościchowska, B., Rogoś-Turek, B. (2015). *Badania jako podstawa projektowania user experience*, Warszawa: PWN.

Mozaryn, J., Różańska-Walczuk, M., Świdrak, J., Kukiełka, K., Pochwatko, G. (2016). Wybrane predyktory postawy wobec robotów społecznych. *Prace Naukowe Politechniki Warszawskiej. Elektronika.* 195. 15-24.

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. Journal of Social Issues, 56(1), 81–103. https://doi.org/10.1111/0022-4537.00153

Ngulube, P. (Ed.). (2022). Handbook of Research on Mixed Methods Research in Information Science. *Advances in Knowledge Acquisition, Transfer, and Management*. https://doi.org/10.4018/978-1-7998-8844-4

Nielsen, J. (2010, October 17). *Mental models and User Experience Design*. Nielsen Norman Group. Retrieved January 19, 2022, from https://www.nngroup.com/articles/mental-models/

Nomura, T., Kanda, T., & Suzuki, T. (2005). Experimental investigation into influence of negative attitudes toward robots on human–robot interaction. *AI &Amp; SOCIETY*, *20*(2), 138–150. https://doi.org/10.1007/s00146-005-0012-7

Nomura, T., Suzuki, T., Kanda, T., & Kato, K. (2006). Measurement of Anxiety toward Robots. *ROMAN 2006 - the 15th IEEE International Symposium on Robot and Human Interactive Communication*. https://doi.org/10.1109/roman.2006.314462

Ogunyale, Tobi & Bryant, De'Aira & Howard, Ayanna. (2018). Does Removing Stereotype Priming Remove Bias? A Pilot Human-Robot Interaction Study. arXiv:1807.00948v1

Oksanen, A., Savela, N., Latikka, R., & Koivula, A. (2020). Trust Toward Robots and Artificial Intelligence: An Experimental Approach to Human–Technology Interactions Online. In

Frontiers in Psychology (Vol. 11). Frontiers Media SA. https://doi.org/10.3389/fpsyg.2020.568256

Onnasch, L., & Roesler, E. (2021). A Taxonomy to Structure and Analyze Human–Robot Interaction. *International Journal of Social Robotics*, *13*(4), 833–849. https://doi.org/10.1007/s12369-020-00666-5

Owen-Hill, A. (n.d.). *What's the Difference Between Robotics and Artificial Intelligence?* Robotiq. Retrieved November 20, 2022, from https://blog.robotiq.com/whats-the-difference-between-robotics-and-artificial-intelligence

Perez-Osorio, J., & Wykowska, A. (2020). Adopting the intentional stance toward natural and artificial agents. *Philosophical Psychology*, *33*(3), 369–395. https://doi.org/10.1080/09515089.2019.1688778

Pérez-Osorio, J., Wykowska, A., Kopp, S., Kahl, S., Nirenburg, S., Pöppel, J., Dagioglou, M., & Spatola, N. (2020). The role and relationship of mindreading and social attunement in HRI: position statements of interdisciplinary researchers. Open Science Framework. https://doi.org/10.17605/OSF.IO/MXGFK

Peyroux, E., Strickland, B., Tapiero, I., & Franck, N. (2014). The intentionality bias in schizophrenia. *Psychiatry Research*, *219*(3), 426–430. https://doi.org/10.1016/j.psychres.2014.06.034

Phillips, E., Zhao, X., Ullman, D., & Malle, B. F. (2018). What is Human-like? *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. https://doi.org/10.1145/3171221.3171268

Piedmont, R. L. (2014). Inter-item Correlations. In Encyclopedia of Quality of Life and Well-Being Research (pp. 3303–3304). Springer Netherlands.

Piçarra, N., Giger, J. C., Pochwatko, G., & Gonçalves, G. (2016). Making sense of social robots: A structural analysis of the layperson's social representation of robots. *European Review of Applied Psychology*, *66*(6), 277–289. https://doi.org/10.1016/j.erap.2016.07.001

Pochwatko, G., Giger, J., Różańska-Walczuk, M., Świdrak, J., Kukielka, K., Możaryn, J., & Piçarra, N. (2015). Polish Version of the Negative Attitude Toward Robots Scale (NARS-PL). *Journal of Automation, Mobile Robotics & Intelligent Systems*, *9*(3), 65–72. https://doi.org/10.14313/jamris_3-2015/25

HRI '06: Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction. (2006). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/1121241

Ramscar, M. (2016). Learning and the replicability of priming effects. *Current Opinion in Psychology*, *12*, 80–84. https://doi.org/10.1016/j.copsyc.2016.07.001

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available online at https://www.R-project.org/.

Rea, D. J., & Young, J. E. (2018). It's All in Your Head. *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. https://doi.org/10.1145/3171221.3171259

Reeves, B., & Nass, C. I. (1996). The media equation: How people treat computers, television, and new media like real people and places. Stanford, Calif: CSLI Publications.

Reniers, Renate & Corcoran, Rhiannon & Drake, Richard & Shryane, Nick & Völlm, Birgit. (2011). The QCAE: a Questionnaire of Cognitive and Affective Empathy. Journal of personality assessment. 93. 84-95. https://doi.org/10.1080/00223891.2010.528484

Riek, L. D., Rabinowitch, T. C., Chakrabarti, B., & Robinson, P. (2009). How anthropomorphism affects empathy toward robots. *Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction*. https://doi.org/10.1145/1514095.1514158

Rohrer, D., Pashler, H., & Harris, C. R. (2019). Discrepant Data and Improbable Results: An Examination of Vohs, Mead, and Goode (2006). *Basic and Applied Social Psychology*, *41*(4), 263–271. https://doi.org/10.1080/01973533.2019.1624965

Roodenrys, S., Barkus, E., Woolrych, T. J., Miller, L. M., & Favelle, S. K. (2021). The intentionality bias in schizotypy: a social matter. Cognitive Neuropsychiatry, 26(1), 55–72. https://doi.org/10.1080/13546805.2020.1865894

Rosenthal-von der Pütten, A. M., Krämer, N. C., Maderwald, S., Brand, M., & Grabenhorst, F. (2019). Neural mechanisms for accepting and rejecting artificial social partners in the uncanny valley. *The Journal of Neuroscience, 39*(33), 6555–6570. https://doi.org/10.1523/JNEUROSCI.2956-18.2019

Rosset, E., & Rottman, J. (2014). The Big 'Whoops!' in the Study of Intentional Behavior: An Appeal for a New Framework in Understanding Human Actions. *Journal of Cognition and Culture*, *14*(1–2), 27–39. https://doi.org/10.1163/15685373-12342108

Rosset, E. (2008). It's no accident: Our bias for intentional explanations. Cognition, 108(3), pp.771-780. https://doi.org/10.1016/j.cognition.2008.07.001

RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.

Salem, M., Eyssel, F., Rohlfing, K., Kopp, S., & Joublin, F. (2011). Effects of Gesture on the Perception of Psychological Anthropomorphism: A Case Study with a Humanoid Robot. *Social Robotics*, 31–41. https://doi.org/10.1007/978-3-642-25504-5_4

Salem, M., Lakatos, G., Amirabdollahian, F., & Dautenhahn, K. (2015). Would You Trust a (Faulty) Robot? *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*. https://doi.org/10.1145/2696454.2696497

Schmider, E., Ziegler, M., Danay, E., Beyer, L., & Bühner, M. (2010). Is it really robust?: Reinvestigating the robustness of ANOVA against violations of the normal distribution assumption. Methodology: European Journal of Research Methods for the Behavioral & Social Sciences, 6(4), 147–151. https://doi.org/10.1027/1614-2241/a000016

Sijtsma, K. (2008). On the Use, the Misuse, and the Very Limited Usefulness of Cronbach's Alpha. *Psychometrika*, *74*(1), 107–120. https://doi.org/10.1007/s11336-008-9101-0

Simon, M. (2018, May 17). *Everything you ever wanted to know about robots*. Wired. Retrieved February 5, 2022, from https://www.wired.com/story/wired-guide-to-robots/

Sharma, D., & McKenna, F. P. (2001). The role of time pressure on the emotional Stroop task. *British Journal of Psychology*, *92*(3), 471–481. https://doi.org/10.1348/000712601162293

Slavny, R. J. M., & Moore, J. W. (2018). Individual differences in the intentionality bias and its association with cognitive empathy. In Personality and Individual Differences (Vol. 122, pp. 104–108). Elsevier BV. https://doi.org/10.1016/j.paid.2017.10.010

Steckenfinger, S. A., & Ghazanfar, A. A. (2009). Monkey visual behavior falls into the uncanny valley. In Proceedings of the National Academy of Sciences (Vol. 106, Issue 43, pp. 18362–18366). Proceedings of the National Academy of Sciences. https://doi.org/10.1073/pnas.0910063106

Strickland, B., Fischer, M., Peyroux, E., & Keil, F. (2011). Syntactic Biases in Intentionality Judgments. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 33. Retrieved from https://escholarship.org/uc/item/89m5h9tw

Subra, B. (2021). The effect of anger on intentionality bias. *Aggressive Behavior*, *47*(4), 464–471. https://doi.org/10.1002/ab.21964

Takahashi, Hideyuki & Saito, Chinatsu & Okada, Hiroyuki & Omori, Takashi. (2013). An investigation of social factors related to online mentalizing in a human-robot competitive game. Japanese Psychological Research. 55. 144-153. https://doi.org/10.1111/jpr.12007

Takayama, L., Ju, W., & Nass, C. (2008). Beyond Dirty, Dangerous and Dull: What everyday people think robots should do. Proceedings of the Human-Robot Interaction: HRI 2008, Amsterdam, NL, 25-32. https://doi.org/10.1145/1349822.1349827

Thellman, S., Silvervarg, A., & Ziemke, T. (2017). Folk-psychological interpretation of human vs. humanoid robot behavior: exploring the intentional stance toward robots. Frontiers in Psychology, 8, 1962. https://doi.org/10.3389/fpsyg.2017.01962

Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences*, *28*(5), 675–691. https://doi.org/10.1017/s0140525x05000129

Trizano-Hermosilla, I., & Alvarado, J. M. (2016). Best Alternatives to Cronbach's Alpha Reliability in Realistic Conditions: Congeneric and Asymmetrical Measurements. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.00769

Urquiza-Haas, E. G., & Kotrschal, K. (2015). The mind behind anthropomorphic thinking: attribution of mental states to other species. *Animal Behaviour*, *109*, 167–176. https://doi.org/10.1016/j.anbehav.2015.08.011

Vohs, K. D., Mead, N. L., & Goode, M. R. (2006). The Psychological Consequences of Money. *Science*, *314*(5802), 1154–1156. https://doi.org/10.1126/science.1132491

Wang, M., Deng, W., Hu, J., Tao, X., & Huang, Y. (2019). Racial Faces in the Wild: Reducing Racial Bias by Information Maximization Adaptation Network. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. https://doi.org/10.1109/iccv.2019.00078

Waytz, A., Cacioppo, J. T., & Epley, N. (2010). Who Sees Human? *Perspectives on Psychological Science*, *5*(3), 219–232. https://doi.org/10.1177/1745691610369336

Waytz, A., Morewedge, C. K., Epley, N., Monteleone, G., Gao, J.-H., & Cacioppo, J. T. (2010). Making sense by making sentient: Effectance motivation increases anthropomorphism. In Journal of Personality and Social Psychology (Vol. 99, Issue 3, pp. 410–435). American Psychological Association (APA). https://doi.org/10.1037/a0020240

What is The Secret of Don Norman's Success? (n.d.). The Interaction Design Foundation. https://www.interaction-design.org/literature/topics/the-secret-of-don-norman-s-success?gclid=CjwKCAiAu5agBhBzEiwAdiR5tCcI8RRL2qFOOAZHRAMmKAQMBvdmFrq0AwZFYVtiTa8NyTi0g2-GqhoC9XwQAvD_BwE

Wheatley, T., Weinberg, A., Looser, C., Moran, T., & Hajcak, G. (2011). Mind Perception: Real but Not Artificial Faces Sustain Neural Activity beyond the N170/VPP. *PLoS ONE*, *6*(3), e17960. https://doi.org/10.1371/journal.pone.0017960

Wiese, E., Metta, G., Wykowska, A. (2017). Robots As Intentional Agents: Using Neuroscientific Methods to Make Robots Appear More Social. Frontiers in Psychology. 8. 1663. https://doi.org/10.3389/fpsyg.2017.01663

Willemse, C., Marchesi, S., Wykowska, A., (2018). Robot Faces that Follow Gaze Facilitate Attentional Engagement and Increase Their Likeability. Frontiers in Psychology. 9. https://doi.org/10.3389/fpsyg.2018.00070

Wilson, B., Hoffman, J., & Morgenstern, J.H. (2019). Predictive Inequity in Object Detection. *ArXiv, abs/1902.11097*.

Wilson, J. H. (2015, April 15). *What is a robot, anyway?* Harvard Business Review. Retrieved January 17, 2022, from https://hbr.org/2015/04/what-is-a-robot-anyway

Wojciszke, B. (2004). *Człowiek wśród ludzi: zarys psychologii społecznej.* Wydawnictwo Naukowe Scholar.

Wykowska, A., Chaminade, T., & Cheng, G. (2016). Embodied artificial agents for understanding human social cognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1693), 20150375. https://doi.org/10.1098/rstb.2015.0375

Wykowska, A., Wiese, E., Prosser, A., & Müller, H. J. (2014). Beliefs about the Minds of Others Influence How We Process Sensory Information. *PLoS ONE*, *9*(4), e94339. https://doi.org/10.1371/journal.pone.0094339

Yong, E. (2012). Replication studies: Bad copy. *Nature*, *485*(7398), 298–300. https://doi.org/10.1038/485298a

Zanatto, D., Patacchiola, M., Goslin, J., & Cangelosi, A. (2016). Priming Anthropomorphism: Can the credibility of humanlike robots be transferred to non-humanlike robots? *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. https://doi.org/10.1109/hri.2016.7451847

Zerka, K. (2022, February 14). Robots intentionality perception. Retrieved from osf.io/2m45u

Zhang, Z., & Yuan, K.-H. (2016). Robust coefficients alpha and omega and confidence intervals with outlying observations and missing data: Methods and software. Educational and Psychological Measurement, 76(3), 387–411. https://doi.org/10.1177/0013164415594658

Appendix A

Figure A1.

The relationship between NATIR scale scores and scores for prototypically accidental (PA) behaviours.

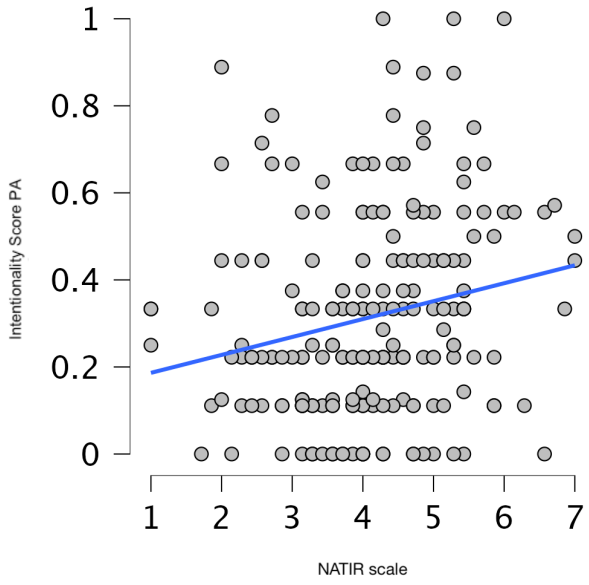


Figure A2.

The relationship between NATIR scale scores and scores for prototypically accidental (CA) behaviours.