Hoax in the Machine:
an Ethical Analysis of Perceived Humanness in Social Robots

Alicja Halbryt

First Supervisor: Dr. Casey Lynch

Second Reader: Prof. dr. Ciano Aydin

Final Master Thesis

MSc Philosophy of Science, Technology and Society

University of Twente

February 2024

ACKNOWLEDGEMENTS

I wish to express my sincere appreciation to my first supervisor, Dr. Casey Lynch, whose shared passion for social robots has been a guiding force in my work. Our meetings created a space where I could freely discuss and explore social robotics, and I will truly miss them. I would also like to thank Prof. dr. Ciano Aydin, my second reader, for his invaluable philosophical insights that greatly enhanced the depth of my work.

Writing this thesis felt like traversing to a different dimension of robotic strangeness and peculiarities, but one where I felt comfortable, and where unriddling the concepts of human life and machine lifelessness gave me endless thrill. The voyage was illuminated by intriguing claims made in scholarly papers, and stumbling upon mind-bending, profound questions about the gist of being a human. I want to thank my boyfriend Daan for listening in and unravelling some of these questions with me, my PSTS friends for cheering and supporting my robot-fascination, and my family for not talking me out of studying philosophy.

ABSTRACT

This master's thesis delves into the concept of deception in HRI, focusing on social robotics, particularly those designed to provide companionship. The central research question explores how design features influence deception levels in HRI and, consequently, impact the formation of human-robot relationships. The thesis posits that deception is integral to social robot design and argues that intentional and unintentional design features can deceive users, seemingly creating aliveness, humanness, or intelligence in the robot to replicate human-human interaction. The thesis categorizes deception into three interconnected elements: animacy, anthropomorphism, and perceived intelligence. Analysing animacy involves key studies in human-robot interaction, discussions on anthropomorphism include research on anthropomorphic technology and relevant moderators, while perceptions of intelligence draw from historical perspectives on human intelligence and studies on machine intelligence perception in robots. Additionally, the thesis critically engages with the perspective proposing a shift from deception to performance. Counterarguments and limitations to this viewpoint are discussed, followed by an examination of reciprocity in HRI, asserting that genuine relationships with machines are not attainable. This comprehensive analysis aims to contribute to the ongoing discourse on the intricate relationship between social robot design, deception, and the formation of human-robot connections.

TABLE OF CONTENTS

1. INTRODUCTION

There is a peculiar curiosity and desire in humans to build a human copy. Interacting with a humanoid machinic form is both mesmerising and bizarre. For some reason, people often do not treat robots as an assembly of electronics, code and plastic, but rather as machines with a ghost inside. It is especially the case in social robotics, a field of study dedicated to the development of robots capable of interacting and communicating with humans in a social manner.

Social robots are a type of a machine often depicted in science fiction films and novels. The ultra-human-like machines playing main roles in, for example, "Westworld" series or "Ex Machina", show that the robot's human appearance and behaviour can have an immense influence on people's actions and emotions. Interestingly, the perfect human form is not necessary for the robot to befriend a human or to be perceived as loveable, with the iconic R2-D2 and WALL-E being the best examples. These two extreme cases – a hyper realistic humanoid robot like Ava from "Ex Machina" and a compact utility machine like R2-D2 – share some specific set of features which can make one believe in their aliveness, consciousness, and intelligence. It is about how they move, communicate, reason and express human emotions that make one perceive in them something more than a machine.

While these examples come from fiction where the technological development is often extremely advanced, the notion of attributing a sense of life to a robot is also relevant in interactions with existing social robots. The level of their advancement and type of application is very broad. They are developed to assist humans in healthcare, therapy, education and customer service. An example of such a machine is Pepper, a semi-humanoid robot often used as a concierge and information provider, but also a robot claimed to be able to read emotions, build relationships with people and sell products (SoftBank Robotics America, Inc, 2023). The role of some social robots is to provide emotional support and companionship to people in their everyday lives. LIKU, a small cute-looking machine, is an example of a robot explicitly designed to accompany people in their loneliness, with the first question on LIKU's company website being: "Have you ever felt lonely when you come back to an empty house alone?" (LIKU, 2018).

A number of social robots developed today are humanoids aspiring to resemble humans to the largest degree possible. Their range was presented at the slightly preposterous AI for Good summit in Geneva in July 2023, where humanoid robots (e.g. Ameca, Geminoid and Sophia) gathered to promote positive uses of artificial intelligence (AP News, 2023). Reporters were invited to ask them questions, making this event the first news conference of social robots. The questions regarded, unsurprisingly, the robots' opinions on them taking away jobs or organising a rebellion against humans. This show suggests that these, frankly, quite creepy and unsettling robots are already listened to and treated as

separate entities having opinions. While it might not be the case among AI and robotics experts who are familiar with how these robots actually work, it can be received differently within the general public and lay audience, who might fall for the human-imitation tricks. Assuming further advancement of humanoid social robots, both in appearance and social behaviours, and their implementation into society, there are many consequences which need to be considered in an ethical discourse. It is important to discuss the effects the human-like features of social robots have on the human-robot interaction, and how they impact the formation of human-robot relationships.

This thesis will discuss the notion of deception in human-robot interaction (HRI), which is present in social robotics in general, with the main consideration and focus given to issues connected to robots providing humans with companionship. Reading through the arguments of this thesis, noteworthy examples of robots include the already mentioned LIKU and Pepper, along with Misty II developed by Misty Robotics to support robotics research (Misty II, 2024) and iPal produced by AvatarMind Robot Technology to serve in elderly care (Avatar iPal Robot Family, 2017). Advanced humanoids such as Engineered Arts' Ameca (Ameca, 2024) and Tesla Bot (CyberGuy, 2023) should also be brought to mind, as they will likely influence the trajectory of the future companionship robot shape and development. Although the market success of some of these robots is unclear, their assumed purpose and form illustrate the direction social robotics is taking.

The thesis will propose significant conclusions regarding human-robot interaction, deception, and human-social robot relationship formation on the course of exploring the following research question:

To what extent do design features of social robots steer the level of deception in human-robot interaction, and how does this affect the formation of human-robot relationships?

The presented research will suggest that deception in social robots is at the core of their design. This means that design features of robots, implemented both intentionally and not, can deceive users and make them think the robot holds a level of aliveness, humanness or intelligence. The point is to make the interaction resemble a human-human interaction as much as possible. As it will be explained, it can be argued that without deception the *social* robot ceases to be *social* and loses its purpose of interacting with humans in every-day situations. Furthermore, I will argue that no relation with a deceitful robot can ever be genuine, that is bi-directional, reciprocal and authentic.

1.1. Methodology and research scope

The above claims were reached through a systematic, secondary research approach. The literature review included a thorough analysis of academic databases, relevant journals, and conference

proceedings in order to identify key aspects and empirical studies of human-robot interaction. The examination of video content involving social robots and commercial websites of robotic companies also served as a source of insight by offering a visual dimension to the understanding of deceptive design features.

More concretely, the aforementioned conclusions are a result of, first, exploring the key aspects of social robot design and considering user perceptions. I will propose a conversation on deception in HRI beyond just anthropomorphism and will suggest a breakdown of deception into three main elements: animacy, anthropomorphism and perceived intelligence. These elements are closely intertwined: animacy triggers anthropomorphism, and this, in turn, influences perceiving intelligence. Analysing these three fundamental features offers a more nuanced perspective on the ways deception is incorporated into robot design. Moreover, among the various features built into robot design to deceive users, animacy, anthropomorphism, and perceived intelligence stand out as the most impactful elements, capable of significantly influencing human perception and fostering deception.

The concept of animacy will be analysed based on, among others, significant work on human-robot interaction and user perceptions of animacy and intelligence by Bartneck and colleagues (2009), along with Sparrow's (2004) Turing Triage Test. Discussion on anthropomorphism will consider research on Anthropomorphic Technology (Cornelius & Leidner, 2021), kinds of anthropomorphic form (DiSalvo et al., 2004) and on moderators of anthropomorphism, for example robot gender and size (Blut et al., 2021). The thesis will also touch upon uncanny valley and Aydin's (2021) research in the area, as well as the significant contribution of Duffy (2003) to the issues of meeting user expectations and transparency about the real (machinic) nature of the robot. Finally, the analysis of perceptions of intelligence will first offer views and research on the history and understanding of human intelligence using, among others, Carson (2015) and Hally (2015). Perceived intelligence in machines will be explored through the important work of Bartneck and colleagues' (2008), as well as Duffy's (2003) perspectives on the illusion of intelligence.

Further, it will be argued that these three elements directly contribute to deception being at the heart of social robotics. To unravel the issue of deception, I will conduct an in-depth discussion of Mark Coeckelbergh's (2017) counterargument, which suggests a shift in perspective from deception to performance. Apart from deliberations about the validity of the approach, I will suggest a critical point of view and limitations to this perspective. The argument will then follow on to the case of reciprocity in HRI, based heavily on the work of Aimee van Wynsberghe (2021), to build the foundation for the claim that a bi-directional relationship with a machine is not possible. Before concluding the thesis

with a summary of the ethical issues of deception, I will a briefly analyse morally permissible and desirable instances of deception in HRI using Matthias' (2015) viewpoint.

1.2. Positioning within the discussion on deception in social robotics

The topic of deception in robots sparks polarised opinions, for example with regards to intentionality which will be discussed below, or if 'deception' is the right term to use at all. It is important, therefore, to clearly position this thesis within the discussion on deception in social robotics upfront. It has to be stressed that this thesis will present a perspective on *social robots* specifically. While most non-social robots indeed do not spark any kind of thoughts of life inside of them, social robots do fall more into that category, and quite deliberately so. The goal of the design of a social robot, especially social companionship robots, is to make the human-robot interaction resemble the one happening between humans as much as possible. In other words, and as the presented research will show, a social robot is supposed to seem as *not-a-machine* and the interaction is supposed to feel as *not* with a machine, as much as it is possible. The aim of this is to make the interaction as smooth, familiar and, in fact, as human-like as possible. What is more, social companionship robots fulfil their role the best when the user feels they can trust the machine and create a connection, which can be achieved through implementing human-like features and behaviours: speech, looks, gestures, and so on.

My opinions are heavily based on and shared by Sharkey and Sharkey (2020; 2011). The scholars hold that the efforts to develop robotic features which encourage the perception of mental life are forms of deception (Sharkey & Sharkey, 2011). There can arise a question, though, whether designing a social robot means deceiving its users by default. To illustrate, in general it can be said that designing a *social robot* means thinking to oneself "my goal is to make a robot which will be able to form social connections. If I make my robot similar to a human (in any way), then the interaction will seem more real and easy to the human, because to some extent it will resemble interacting with a person, or a living being at least". Implementing any kind of human features with such a mindset could be, one could argue, perceived as an intention to deceive, that is an intention to spark in the user the tendency to perceive in the robot something that is not there, something that is not real and not true. As will be mentioned later in the thesis, making one believe something that is not true is one way to define deception.

Through a simple observation of some of the social robots' websites (e.g., Misty II, LIKU) one can see the implemented marketing strategies which arguably reveal the underlying design intentions. These robots are developed to possess 'personality', 'emotions', and an aura of cuteness, to become your friend or companion. This strongly suggests an overarching objective: to attract and persuade users of

these inherent (artificial) qualities within the robots through how they look or behave. Nevertheless, the problem of intentionality raises significant questions. Sometimes, the robot might just look like a human, or imitate the expression of human emotions, but there is no intention to deceive the observer. In this case, it is possible that people sometimes misrecognise what is going on and assign humanity to machines which were not intended to spark this kind of reactions. This happens, though, because of humans' natural tendency to assign human features to non-living objects. As it will be deliberated later on, even a non-humanoid robot Roomba can make one feel as if it is more than a machine and that it possesses some mental capacities (van Wynsberghe, 2021). It is a result of anthropomorphism, which I will argue is one of the main components of deception.

On the other hand, the observer may not always buy into or fall for the perceived mental states of the machine, even though the intention to make the observer believe these things was there. If the robotic features were intended to deceive, but did not succeed, is it deception? And conversely, if there was no intention to deceive, but the person falls for the perceived humanness, can this still be called deception? Sharkey and Sharkey (2020) argue, and I also do share their view, that deception in social robotics does not require intention. They give an example of Paro, the robotic seal. While manufacturers did not intend to make users believe it is a real seal, some of them still perceived sentience or cognition based on Paro's behaviours. Similarly, some men believe they are in a loving relationship with their sex doll and feel loved by the doll in return, a result not originally intended by doll manufacturers, as argued by Sharkey and others (2017). Following Sharkey and Sharkey (2020), these are cases of unintentional deception. Namely, the makers of the robot did not intend to deceive, but it happened anyway.

Here, it should be emphasised that not all deception is unethical, and that claiming so would be too extreme of a statement. Robot designers might simply intend to entertain (Sharkey & Sharkey, 2020) through making the robot resemble a human, speak like one, or express and read human emotions. Some scholars even suggest stepping down from using the deception vocabulary at all and propose new approaches, for example Coeckelbergh (2017), whose point of view will be discussed in chapter 3. Nevertheless, it can be argued that even though people are not always deceived or intended to be, the potential is there, because of the human-like features of the social robot design. These perspectives resonate with the already mentioned claim of Sharkey and Sharkey (2011) that social roboticists' efforts do aspire to create an illusion of mental states or human appearance in machines, but not all these efforts are steered by inherently bad intentions to fool the users or to bring negative outcomes. It has to be noted that deception can bring positive outcomes for the deceived and can be created with good intentions in mind (Sharkey & Sharkey, 2020) which is the case for example in elderly care where placebo or white lies are a common practice (Schermer, 2007).

In light of this, one of the claims this thesis makes is that deception is at the core of social robot design. What does it mean? Following Sharkey and Sharkey (2020: 315):

"If the behaviour and appearance of a robot leads to people believing that a robot has cognitive abilities or that it cares for and loves them, then, we argue, they are being deceived whether or not anyone intended to deceive them".

The behaviour and appearance of a robot the scholars talk about here depend on its designers. The design of a *social robot* is about creating an interaction between a robot and human that feels natural, which accounts for making one perceive (at least) mental states in the machine which are in fact not there. Some may not buy into it, but it can be argued that the goal is that they do. This is because if people do not buy into it, then the *social* part of the *social robot* ceases to make sense. Through buying into it (i.e. believing the mental states, emotions, etc., are true) the human-robot interaction can resemble the levels of human-human interaction.

Lastly, following Danaher (2020), several roboticists (e.g., Shim and Arkin, 2016) have debated that social robots will have to be equipped with some level of deceptive capacity in order to integrate them seamlessly into our society. This suggests, first of all, that deception is a critical component of social robot design which can allow for smooth interactions with humans and, evidently, integration in the human society (whether that is desirable will not be discussed here). Considering this, although deception in current social robots is not at its peak yet, indications suggest that robots will increasingly strive for integration into society. Consequently, deception may inevitably become even more of a fundamental and indispensable component of their design.

1.3. Embodied Artificial Intelligence

Among the many artificial intelligence applications and use cases, social robots might be the most particular. Robots allow for a type of interaction till now possible only between humans – face to face, in a way. Additionally, the physicality of a social robot plays an extremely important role in the development of AI, as it is thought to be the best way to replicate intelligence.

Interestingly, even though it remains unclear what intelligence or consciousness are, people want to copy those in a machine. Some people claim that we already live among intelligent technologies, a few others recently started to believe in their sentience (Tiku, 2022). However, while there indeed exist very complex AI systems, such as the now popular generative AI system ChatGPT, it can be argued that they are not intelligent. As will be argued later on, this is both because there is a lack of agreed upon definition of intelligence, and because these systems do not seem to fall into the general

understanding of *human* intelligence. A very significant point of view by Rodney Brooks, the father of robotics and HRI, supports this claim. Brooks (1999) proposes that in order to successfully develop an intelligent system, it has to be embodied. The roboticist claims that a system's intelligence can emerge from its interactions with the surrounding world. The significance of embodiment of intelligent systems comes from the fact that only an embodied agent can be fully validated and accepted as one that can cope and deal with the real world around it. In addition, Brooks (1999) suggests that only through physical form can the system actually analyse the world, receive cues and give 'meaning' to what is going on inside it. What is more, it is claimed that the real world is its best model, meaning that it is not possible (and also unnecessary) to create a world model based on abstract descriptions in which the intelligent system exists. The real world is always exactly up to date and contains all the details there are to be known (Brooks, 1999). In order to be able to experience the world directly, robots need to operate in dynamic environments using real and various type of sensors. Actions of the machine are part of a dynamic exchange with the world and have instant feedback on the machine's sensations (Brooks, 1999).

This brings to mind how children first learn and experience the world – by being in it, touching different objects, exploring the boundaries, seeing changes happen around them, and so on. One can interpret Brooks' (1999) thoughts as steering towards copying the human intelligence to a very true degree. If an embodied AI system can be in the world the way a human child is, provided that it has sensors working similarly to human senses, it should develop a similar type or level of intelligence. However, while the physical form of a system might indeed help develop intelligence similar to human's, it can be argued that it could also be a limitation. One could see a body as a restraint in development or confinement. If we want to achieve a human intelligence in a machine, then it should have a body. But, if we want to create an *artificial* intelligence, does it still need a body? While this thesis will focus on embodied AI and assume that a body is indeed an important factor in AI development, it is definitely a question worth considering further.

Embodied AI concept suggests a robot. It is especially crucial to deliberate over the design of a social robot since embodied artificial intelligence is arguably the form of AI which can facilitate the development of artificial general intelligence (AGI). AGI, a 'programmers dream' of the future, is understood as AI systems striving for versatility and adaptability which can encompass various tasks and domains – contrary to the current, narrowly working AI (Everitt et al., 2018). The social aspect of the robot is perhaps the most important one to attempt to develop their human-directed behaviours and to assimilate robots in the society, which is a dream of many innovators. This assimilation is envisioned in many ways, for example through placing robots in hospitals and care homes as carers, in offices and hotels as receptionists, in museums as guides, and in our homes as assistants or, what

is already happening, sex partners. Consequently, it is foreseeable that interest in social robotics will continue to rise. In fact, predictions indicate that the social robotics market will expand from USD 4.26 billion in 2023 to approximately USD 17.32 billion by 2028 (Mordor Intelligence, 2023).

Given the profound significance of a (social) robot's embodiment, it becomes necessary to meticulously analyse its design, including not only the outward appearance but also the intricacies of its behavioural attributes. As it was stated above, some of the most challenging ethical problems in human-robot interaction seem to arise once the machine takes a human form. The following chapter will try to break down the very fundamental elements of social robot design that account for its human resemblance. Animacy, anthropomorphism, and perceived intelligence will be discussed in detail, to lay the groundwork for a later exploration of deception in HRI.

2. DESIGN OF A SOCIAL ROBOT AND USER PERCEPTIONS

Embodiment seems to be a crucial factor for successful development of machine intelligence. Robotics, therefore, is the key field in this matter. As stated above, though, it is not enough for the robot to just have a body with sensors to experience the environment. The body and its behaviours are subject to judgments and evaluations made by the eye of the observer (Brooks, 1999). This means that, in a way, the decision of whether or not the machine is intelligent, friendly, or creepy, is up to the robot's user. The way a machine looks and acts will, therefore, determine how the robot is perceived. It puts a lot of emphasis on the role of the robot's design and designers making design decisions. While it is the user who feels in a certain way about the robot, it is up to designers what kind of techniques and features to use in order to make the robot make the user feel this certain way.

This chapter will focus on the seemingly most fundamental components of the social robot design. These components are animacy, anthropomorphic features and perceived intelligence. The order of the components is not coincidental. Animacy takes precedence due to the fact that, on a very fundamental level, it distinguishes living beings from the non-living, and humans and animals from objects. Quite naturally, animacy leads the discussion towards anthropomorphism. The tendency to ascribe human traits to non-human entities is evoked and enhanced through the features in-designed in social robots. Seeing humanness in machines leads their users to perceive traces of intelligence. Despite the lack of clear definition of intelligence, a robot which is designed to look and talk like a human is expected to possess at least some level of human-like intelligence. All in all, this chapter aims to lay the groundwork for an argument demonstrating that the central aspect of social robot design revolves around deception.

2.1. Animacy

Social humanoid robots are created with an idea to be integrated into our society and take on roles ranging from companions to the lonely, intimate partners to teachers in school. If machines should one day join the human society, it is necessary to understand what people's attitudes, tendencies and expectations are in human-robot interaction. The thing that distinguishes people from machines is being alive. However, social robots can *seem* to be alive and human-like. The level of the perceived 'aliveness', which is the perception of animacy in robots, can be influenced by the robot's design.

Bartneck and colleagues (2009) conducted an experiment to see how the design of a robot influences the user perception of animacy and intelligence. They claim that if people perceive a robot to be just a machine, then it should not be a problem for them to switch it off. However, if people perceive traits of life in the robot, it is likely that they will hesitate to switch it off, since they would think of the possible consequences of doing so (Bartneck et al., 2009). Bartneck, et al. (2009) draw from the Turing Triage Test, a concept proposed by Robert Sparrow (2004). Sparrow suggests that there might come a point in the future when a machine's existence becomes equally important to that of a human. In other words, it is possible that machines will one day achieve a moral standing comparable to humans'. According to Sparrow (2004), if humans encounter a moment when they make a judgment that it is reasonable to preserve the existence of an artificial intelligence over the life of a human being, that is when machines have achieved a human-level moral standing. This dilemma is the Turing Triage Test (Sparrow, 2004). An exemplary scenario (Sparrow, 2004) in which such a dilemma might occur involves a person who is a hospital administrator during a catastrophic loss of power in the hospital, and two patients connected to a life support system. The hospital administrator has to choose which one of the two patients to continue to provide electricity to, and therefore save one of their lives. Further on, the scenario includes a sophisticated artificial intelligence medical officer aiding in diagnosing patients. It is capable of learning, reasoning independently and can make its own decisions. It can converse with doctors in the hospital and over the phone, fully passing the Turing Test. In this catastrophic scenario, its battery is failing and thus it is connected to the already limited power supply of the hospital. The hospital administrator has to decide whether the electricity should go to the human patient's life support, or to power the AI employee (who begs to be saved). Switching off the machine will fuse its circuit boards, making it fully inoperable later; cutting off power from the life support will, of course, kill the human. For Sparrow (2004), if saving the AI machine has the same character as saving the human, machines have achieved the moral status of human beings.

Inspired by the Turing Triage Test, instead of asking the study participants to choose between the life of a human and existence of a robot, Bartneck et al. (2009) want to see how hesitant users can be to switch a robot off. Particularly, they examine how the level of a machine's human-likeness (in

behaviour and appearance) and animacy influences people's perceptions and hesitation to switch it off. Before discussing the features of a robot's design which increase the level of perceived intelligence, it is worth discussing the animate-inanimate distinction first, that is the state of being (or not) alive. It is crucial to understand how assigning animacy to non-living things work, since animacy is the first fundamental step for a technological object to become more than just a non-living thing. What is more, Bartneck and colleagues (2009) found that there is a significant correlation between animacy and perceived intelligence. What they suggest is that a smarter robot might also be seen as more animate. Their study also showed that people were much more hesitant to switch off the seemingly more intelligent robot compared to the more stupid one. This leads to a possible conclusion that a robot's behaviour is more important than its embodiment (Bartneck et al., 2009), at least in perceiving intelligence and animacy.

Interestingly, an ability to make the distinction between animate and inanimate is not present in human babies. Studies in HRI suggest that small children are not able to perceive a humanoid robot as inanimate or creepy (Bartneck et al., 2009). A relevant example can be found in Kahn et al. (2006), who study preschool's children behavioural interactions with and reasoning about the back then most advanced robotic pet on the retail market, a robotic dog AIBO. Next to AIBO, the children interacted also with a stuffed toy dog. Kahn and colleagues' (2006) findings show that children perceived both AIBO and the stuffed dog in a very similar way. However, interesting differences come out when looking closely at behavioural interactions. The results of the study suggest that preschool children knew that a stuffed dog is not alive, while they tended to assign some animacy to the robotic dog and treated it almost as if it was a real dog. The children more often than with AIBO mistreated the stuffed dog, which implies that they did not in fact believe it to be the sort of entity which is able to feel. They also endowed the stuffed pet with more animation than AIBO, which suggests that children thought of AIBO as able to direct its own behaviour, while the stuffed dog needed more of their assistance (Kahn et al., 2006). One of the conclusions one can develop from this study is that it takes a child's imagination in order for the stuffed dog to become alive. AIBO dog, on the other hand, shows real-dog behaviours in reality. This seems enough for children to believe a robotic pet is animate. As children get older though, their suspicions towards robotic behaviour and appearance, and thus the ability to perceive inanimacy in robots, develop (Bartneck et al., 2009).

All in all, it is often a goal of many roboticists to not only make their robots lifelike, but also to give their users an idea of being in the presence of *someone*, not *something* (Carli et al., 2022). One of the reasons for this is that often lifelike beings, for example in computer games, can cause the users to get involved emotionally. Through this involvement, it is possible to influence them (Bartneck et al., 2008). A good example of evoking emotional reactions, and the potential for resulting manipulation,

lies in the incorporation of features associated with cuteness. Lacey and Caudwell (2019) claim that cuteness of a home robot, serving as a powerlessness aestheticization, can disguise the machine's powerful capacities for data-gathering. Moreover, studies demonstrate that cuteness can impair long-term decision-making by triggering short-term rewards-based responses. This can lead to users making choices against their long-term best interests, especially concerning information privacy (Lacey & Caudwell, 2019).

However, to achieve the perceptions of robot animacy in a user, it is already enough to embed simple movements. As little as triangles and circles moving on the screen can make one develop thoughts of the shapes being somewhat alive. Reactive behaviour and responsiveness to events, for example a robot looking up when touched on its head, can also have a great influence on how alive the machine seems to be (Bartneck et al., 2020). Similarly to anthropomorphism, analysed in-depth below, it seems unavoidable for perceptions of robot aliveness to arise in HRI. Arguably, it is then up to robot designers how this knowledge is used.

2.2. Anthropomorphic features

The problem of animacy of objects is directly linked to the notion of anthropomorphism. It can be said that anthropomorphism is animacy taken to a more detailed level – instead of perceiving aliveness by itself, people perceive human-like aliveness. What is more, research in HRI often indicates a highly positive correlation between animacy and anthropomorphism, which means that being alive is an indispensable part of being human-like. For instance, it has been found that the more human-like a mouth of a robot is perceived by the users, the more alive the robot seems to them. In more general words, the more a robot is humanised, the more lifelike the perception of it (Blut et al., 2020). Therefore, anthropomorphism can impact perceived animacy in a positive way.

It can be found in Cornelius and Leidner (2021) that anthropomorphism is an ingrained tendency in humans, which can also be called a chronic feature of human beings. To anthropomorphise means to attribute human traits and emotional states to non-living and non-human entities, including animals. It can be said that anthropomorphism is the action of treating non-human behaviours as motivated by human feelings and mental states (Damiano & Dumouchel, 2018). In other words, humans filter the behaviours of other non-human entities through their own, human lens, and perceive these behaviours as, in a way, familiar to their own. What this suggests is that anthropomorphism is people's attempt to rationalise and understand actions of non-humans (Duffy, 2003), and to make them more explainable or predictable (Fink, 2012). Attributing human traits to non-human entities is based on prior experiences with people, and it is supposed to help build social connections and enhance feelings

of belonging (Cornelius & Leidner, 2021). Interestingly, anthropomorphism has been traditionally perceived as a category mistake (Damiano & Dumouchel, 2018). It has been considered by some as a bias, or an obstacle on the way to the advancement of knowledge. Anthropomorphic tendencies have even been labelled as a psychological disposition which is typical for the immature and unenlightened, i.e. young kids and 'primitive people' (Damiano & Dumouchel, 2018). This negative approach to anthropomorphism is challenged by a re-evaluated concept of anthropomorphism which is supported by recent findings in cognitive sciences. This new approach rejects the idea of anthropomorphism being an early childhood 'indisposition' and generally a cognitive mistake, and argues for anthropomorphism constituting the permanent and fundamental dimension of the human mind (Damiano & Dumouchel, 2018). In other words, anthropomorphism is an inevitable and constant human tendency to recognise behaviours and features of non-human entities as human. What is more, any object may be perceived to be human-like, including a Coca-Cola bottle or a car (Moussawi & Koufaris, 2019).

It has to be noted that while indeed everyone does anthropomorphise non-living and non-human entities, the extent to which it happens varies from one person to the other. It has been suggested in various studies on human-robot interaction that people who have a greater need of belonging tend to anthropomorphise more (Kim et al., 2013; Blut et al., 2021; Cornelius & Leidner, 2021). Lonely people have a stronger disposition to humanise a robot, which might be caused by social exclusion, isolation or disconnection of any sort. By anthropomorphising an object, in this case a robot, they can satisfy their desire for affiliation through having a perceived human-like relation with a robot (Blut et al., 2021). Personality characteristics also have influence on how people anthropomorphise. Traits such as extraversion and agreeableness have an impact on how one interprets human-like behaviours, and therefore influence the degree of anthropomorphism (Cornelius & Leidner, 2021). Kim and colleagues (2013) state that these individual differences need further attention and research, especially when it comes to understanding user psychology and internal human nature. What they suggest is that certain populations, for instance gamers with strong immersive tendencies, or elderly users with significant need to belong, may be influenced by and be more receptive to robots than other members of society (Kim et al., 2013). This claim can make one think of a very broad spectrum of potential ethical issues faced by the end users of robots and the impact (both negative and positive) the interactions with robots can have on some people's wellbeing. This is especially relevant for the case of companionship robots, which are designed to spark emotional reactions in users and facilitate the formation of a level of attachment.

2.2.1. Anthropomorphic Technology

With this suggestion in mind, and knowing that anthropomorphism is inherent in people, clearly it should be essential to take anthropomorphism into account when designing and developing technologies. The tendency to perceive humanness in machines (such as social robots) cannot be omitted when decisions on the machines' form and design are made. This is especially important when the machine is intentionally designed to resemble a human being, in behaviour or appearance, and therefore may receive strong anthropomorphic reactions. Robotics research points out that through anthropomorphism robots seem more humanlike, and therefore more familiar. The sense of familiarity has a positive impact on the user experience of the robot, since it is familiar and very natural for people to interact with a human-like entity (Blut et al., 2020). Such kind of technology is defined by Cornelius and Leidner (2021: 1) as:

"technology that possesses design features that motivate anthropomorphism [and which] can be referred to as anthropomorphic technology (AT)".

In their paper, the researchers deliberate about the acceptance of anthropomorphic technologies and suggest that the sole fact that a machine is human-like and motivates anthropomorphism does not necessarily lead to it being accepted by the user.

AT can hold design features of anthropomorphic form or function, or usually both, which trigger anthropomorphism. In fact, following Bartneck and colleagues (2020), the form should match the function, and vice versa, in order to accurately meet users' expectations. For instance, a robot designed for companionship should match its humanoid form with its companionship function (human-like speaking, gestures, etc.). Similarly, if a machine has eyes, it will be expected to see; if it was built for cleaning purposes only, it is not expected to have human-like features.

Human-like form, according to Cornelius and Leidner (2021), is a result of integrating human-likeness in the design which is interpreted as human-like through observation. It is, therefore, the physical embodiment of the machine, which can be consistent and static, and which is by observation 'labelled' as resembling a human. The human-like form of a machine can include things like shape, movement, gestures, and general appearance. There are numerous examples of experiments and research done with devices and digital interfaces which are supposed to motivate anthropomorphism through their form. For instance, it has been shown that it takes as little as a head tilt of a robot to assign humanness to it (Mara & Appel, 2015). Robotic face design is also something that has been recently getting more attention, with the Ameca robot shocking the public with its hyper realistic facial expressions.

On the other hand, human-like functions are the behavioural traits of a machine which resemble the ways people think and behave with other humans. It is manifested in the way the machine interacts (behaves) and 'thinks' (Cornelius & Leidner, 2021). The human-like function encompasses natural language processing, conversational ability, interactivity, and also the sole purpose of the robot. A good example of such machines are emotionally expressive robots, such as BUDDY or LIKU (Lynch, 2021). The vast emotional expressiveness of these machines (or, in fact, mimicking human emotional expressions) and in turn motivating emotional reactions in their users (such as empathy or joy) is their main function. Cornelius and Leidner (2021) also mention intelligence as the human-like function type. They claim that human-like intelligence appears in a robot through, for instance, the way it converses and uses the language.

A different approach to anthropomorphic technology is presented by DiSalvo and colleagues (2004). Their research was guided by a set of questions directed at designed anthropomorphic forms. Firstly, if an aspect of any form is its material qualities and properties, then how specifically is the human form imitated? In other words, a designer makes certain decisions about the object's scale, abstraction, proportions, and so on. How are the features of the object designed to imitate human form? Secondly, DiSalvo and colleagues (2004) claim that all these design decisions to imitate human form serve some purpose, other than simply improving the object's style and design. Therefore, what is the purpose of imitating humanness? This can be answered by inspecting intentions of the designer and designed functions of the object. Lastly, it has to be clarified what is meant by human form. What constitutes it? Which parts of the human form are imitated, and why them specifically? (DiSalvo et al., 2004).

As a result of asking the how, why and what of imitating human form in design of objects, DiSalvo and colleagues (2004) distinguish four kinds of anthropomorphic form. These are structural, gestural, character and aware form. *Structural* anthropomorphic form is an imitation of both the operation and construction of the human body, with a special focus on the body's materiality. It features mechanisms, shapes, volumes, and arrangements which intend to copy the functions and looks of the human body. Structural anthropomorphic form is largely based on the knowledge of human physiology and anatomy, and is an expression of 'thing-ness' of a body of a human (DiSalvo et al., 2004). As an example of the structural anthropomorphic form, DiSalvo and colleagues (2004) refer to an artist's poseable mannequin. This product, which is around 1/6 scale of a real human, imitates a human shape and several major joints of a human's organism. These copied human body parts are universal to all human beings (DiSalvo et al., 2004). What could serve as a contemporary, strictly technological example is any humanoid robot currently in development. Robots that first come to mind might include Tesla's Optimus, Boston Dynamics' Atlas or Sanctuary AI's Phoenix. All of these

machines have a very visibly human-like form, with four limbs and a head. It is clear that the intention of the design of these robots is to resemble the human shape. Although their purpose and functions differ quite extensively, which also impacts the details in their mechanisms and appearance, all these robots aim to mimic a human's structural features as accurately as possible.

The second type of anthropomorphic form, the *gestural* form, mimics the human body in terms of how the bodies communicate with each other and how they behave. The evidence of gestural anthropomorphic form can be found through the use of motions or poses which intend to suggest human action to express intention, instruction, or meaning. The gestural form draws heavily from the understanding of the expressiveness of the human body, as well as knowledge of human non-verbal communication. DiSalvo and colleagues (2004) give a particular example of the gestural anthropomorphic form, which could have been found in the feedback feature of the Mac OS X login screen. Namely, when a user has entered their password incorrectly, the login window would quickly and briefly shake from side to side. The researchers point out that this kind of motion resembles a common human gesture of expressing "no". This gentle suggestion of a mistake made imitates a human headshake (DiSalvo et al., 2004). Another example of a gestural anthropomorphic form is LIKU, and many other emotionally expressive robots already mentioned above. LIKU is a small-sized, human-shaped robot designed to provide company to lonely people. It is supposed to mimic human behaviours and emotions, which is displayed through dancing when it 'feels' happy, or through making sad eyes when the situation requires so (LIKU, 2023). These gestures and expressions intend to reinforce feelings of anthropomorphism and empathy in humans, which in turn intensifies the relationship between the robot and the user.

The next anthropomorphic form is the form of *character*. It is about how the traits, functions and roles of people are reflected in design of objects. Manifesting the qualities or habits which intend to describe individuals serves as an evidence of the anthropomorphic form of character. This type of form is based on the knowledge of societal conventions, contexts, and biases, and aims to reflect the practices that human beings engage in. DiSalvo and colleagues' (2004) example of the anthropomorphic form of character is the Jean-Paul Gaultier's perfume bottle "Le Male". According to the scholars (DiSalvo et al., 2004), while the bottle displays elements which can categorise it also as structural and gestural anthropomorphic forms, when looking at it as a whole it is a form of character. The reason for this is that the bottle is shaped as a specific type of a human body with specific traits, rather than merely a general shape of a human body. It is erotically charged and portrays male sexuality in a specific, socially construed way (DiSalvo et al., 2004). An example of an anthropomorphic form of character found in the field of technology could be Hanson's Sophia the robot. While the technological complexity of Sophia and the way the robot is promoted are questionable, it is an

instance of a machine with an in-designed character. Sophia is evidently a female robot, which can be noticed by the facial features, as well as tone of voice and, of course, the name.

The last distinguished form is *aware* anthropomorphic form. This type of form seems to be the most abstract and complex among the remaining three. It is an imitation of the human capacity for intentionality, inquiry, or thought. Additionally, it is an expression of social qualities of being a human. A design of an object which suggests that it is aware of itself in relation to others, is able to create and process abstract ideas, and is capable of interacting with others can be recognised as an aware anthropomorphic form. At the time when DiSalvo and colleagues' (2004) research took place, they expressed difficulties with finding an aware anthropomorphic form that would not be a fictional object. In the paper, they give an example of R2D2 droid from the Star Wars movie series. R2D2 exhibits awareness in interactions with its companions, it is aware of the relationships it has with people, and can generally express its own thoughts (even though not in a human language). DiSalvo and others (2004) admit that, although they live on the border between fact and fiction, the aware anthropomorphic forms can be spotted in the field of artificial intelligence and robotics, where the level of human imitation is high. Robots are being designed to mimic humans through programmed abilities, such as learning, reasoning, adapting, and friendly interacting. It can be argued that what DiSalvo and colleagues (2004) mean is that the robots are *programmed* to be *perceived* as human. Interestingly, almost 20 years after DiSalvo and others published the paper (2004), it is still challenging to find an aware type of anthropomorphic form that does not exist only in fiction. Nevertheless, probably one of the best examples which aspires to have features of the aware anthropomorphic form is Ameca developed by Engineered Arts. The grey-skinned robot has drawn general attention with its extremely human-like face and the accuracy of emotional expressions its face is able to make. While Ameca is not (yet) able to walk, it can now answer questions, which combined with the facial expressions gives an impression of thought processes and awareness taking place inside the robot's 'mind'. Another fitting example of aware features in a machine is the emerging development of socially-aware navigation (Salek Shahrezaie et al., 2022). In essence, instead of aiming only for getting from point A to point B, a robot's movements are programmed to seem that it is aware of implicit social norms around it, for instance the robot will not get too close to someone out of respect for their personal space. Abiding to social norms is definitely a development necessary for the success of human-robot interaction and social robotics.

As is clear from the above discussion, there are multiple ways in which an object can spark anthropomorphisation. Blut and colleagues (2021) distinguish a couple of features, or moderators, which have impact on anthropomorphisation of service robots specifically. In particular, they suggest that there is a strong preference between users towards physical embodiment over lack of thereof

(digital avatars). Robots with physical bodies are more appealing, evoke empathy and, naturally, are regarded as more socially present. Further, the robot's gender plays an important role in how the robot is anthropomorphised and received. While machines are genderless, it is possible for designers to, through voice, appearance or name, add gender cues. This allows for making gender-stereotypical conclusions and assigning certain characteristics to robots, similarly as it happens between human genders. For instance, female robots are perceived by customers as more affectionate and polite than male robots (Blut et al., 2021). What also moderates anthropomorphisation of robots is their size. The range of robot sizes is extensive, and can make one feel either safe and in control, like a small robot, or inferior and at risk, in case of physically superior robots. Interestingly, Blut and colleagues (2021) claim that because large robots seem more threatening to people, the importance of human-like features embedded in them increases. Therefore, to counterbalance the fear, it is important (perhaps more important than in case of small-sized robots) to make large robots seem more familiar and friendly through implementing human-like features. The researchers also mention cuteness as a feature impacting anthropomorphism. Designing an endearing appearance of a robot brings positive responses and can strengthen user intention to use it (Blut et al., 2021). It becomes clear from this section that, given the number of known design factors and choices to be made in HRI, every feature in robot design is deliberate. What is more, in case of humanoids, most if not all of these features aim to convey the robot's level of humanity. These are all conscious design decisions that either enhance or diminish the likelihood of the robot engaging in deceptive behaviour (Matthias, 2015). This conclusion confirms this thesis' claim that robot design is rooted in deception.

2.2.2. Uncanny valley

Sometimes the level of human-likeness in the anthropomorphic features of robots can go too far. When looking at the Ameca robot, some people might feel that the human resemblance is too strong, even though Ameca is still quite far from being a perfect human copy. This kind of sensation can cause feelings of uneasiness or creepiness towards the robot. What is more, Brink and others (2017) claim that very human-like robots are considerably more creepy to people than other robots. This brings about the concept of the uncanny valley explored in 1970 by a Japanese roboticist Masahiro Mori. Although the topic is extremely broad and fascinating, the thesis will try to synthesise the main ideas of this issue, as it is one of the factors having influence on HRI and deception. Mori's (1970) uncanny valley concept states that the more closely a machine resembles a human being, the more affective reactions it is able to engender through human-like stimulus. Yet, there comes a sudden drop in the acceptability of the robot, when the level of human-likeness in the machine becomes unnerving,

causing some to feel creeped out, anxious, and generally very uncomfortable in contact with the robot.

Mori's concept of the uncanniness occurring in an interaction with a machine is further extensively investigated by numerous researchers from different fields. A relevant and interesting point of view is presented in the work of Ciano Aydin (2021), who proposes that the uncanny feelings people respond with to humanoid robots can say quite a lot about the human psychology. In addition, this eeriness points towards the ontological side of a human, namely that the uncanny valley shows the gist of our relation with the self. To understand this perspective, it is interesting to analyse the multiple hypothesis about why the uncanny valley occurs at all. One of the possible reasons is the Pathogen Avoidance hypothesis, developed by Mori himself (Aydin, 2021). He related the uncanny valley with the human instinct of self-preservation. The hypothesis states that people perceive human-like robots to be genetically similar to humans, which means that any visual anomalies spotted on the robot (especially on hyper-realistic robots) cause pathogen avoidance mechanism. It indicates to people transmissible diseases and makes one feel disgust. A hypothesis slightly related to this one is Mortality Salience, which implies that the uncanny valley induces the fear of death, and therefore is a reminder of one's unavoidable mortality (Aydin, 2021). Seeing a human robotic copy which acts and looks bizarre, humans can also experience the fear of being replaced by a soulless, odd machine – a poor copy of themselves.

A definitely relevant uncanny valley hypothesis for this thesis is the Violation of Expectations hypothesis, also proposed by Mori (Aydin, 2021). Essentially, it suggests that a human-like robot can fail to meet people's expectations by not looking the way one has thought or assumed. It is not so much about that the machine fails to be a perfect copy of a human, but more about it actually being perceived as one while at the same time not living up to the standards of acting and looking like a standard person (Aydin, 2021). This hypothesis could also be extended further to the behaviour of the robot in question. If a machine is obviously an inanimate object, meaning it is not alive, then it should not move or speak, or behave in any way human. Yet, some humanoid robots do or aspire to move or speak. This contradiction involving inanimacy connected with humanity can also be an explanation for the feelings of creepiness and unease.

It is also important to draw attention to the concept of anthropomorphism and its relevance for the uncanny valley. People do have the tendency to assign human characteristics to non-human entities, which in itself does not explain uncanny valley, nor can it be a result of attributing human features to a robot. In fact, the Dehumanisation Hypothesis suggests that the uncanny valley should be regarded as a response to the lack of humanness (Aydin, 2021). This leads to a conclusion that an

anthropomorphised human-like robot is not perceived simply as a robot, but as a robotlike human. The humanness of the robotlike human is questioned while its mechanistic nature is being revealed. In other words, the more one anthropomorphises a human-like machine, the more likely one is to notice its robotic nature, which induces a dehumanisation process and in effect decreased likeability and trust (Aydin, 2021). This all might result in the uncanny feelings.

One could argue that humanity has had the opportunity of dealing with artificial entities only for a few decades in our thousands-of-years existence. Therefore, interacting with humanoid robots (and other AI devices) is something quite new for the human brain. This naturally brings inability and inexperience with dealing with machines and can cause all sorts of emotions, from excitement to fears. Uncanny valley could be a result of a mental inability and under-preparedness to interact with a machine. Interestingly, though, this 'incompetence' to be around machines does not appear until a certain age. In a study exploring the origins of the uncanny valley, Brink and colleagues (2017) examined children's responses to human-like robots. What they found is that children aged 9 and younger do not experience uncanny valley, and that the phenomenon emerges through development. In other words, young children found both very human-like and machine-like robots equally not uncanny, while for older kids very-human like robot was much more creepy than a simple, machinic robot. This is a reaction similar to those of adults. The absence of the uncanny feelings might be a result of children's expectations towards robots to have a multitude of mental abilities. The uncanny valley sort of tracks the changing understandings of mind. Brink et al. (2017) claim that evidently only children older than 9, who have clear assumptions about human and robotic minds and their mental abilities, feel uneasy towards very human-like robots. One could raise a question whether the lack of uncanny feelings would have always been the case, or whether young children who lived centuries ago would actually experience the uncanny valley towards humanoid robots. After all, right now more than ever children are exposed to cartoons and robotic toys, which makes interaction with artificial entities quite normal and familiar. Only with time do they realise the toy is in fact far from being human-like.

It can be said that we are supposed to be deceived and believe the robot holds some level of humanness, but it somehow does not always fully work, which is manifested through uncanny valley. In a way, we reject to be fully deceived by the human-likeness of the machine. What is more, people have certain expectations towards human-like robots. Because of the aspect of anthropomorphism and perceived familiarity, humans apply human-to-human interaction patterns to the interaction with something which has been designed to imitate a human. This can often lead to disappointment of (unrealistic) expectations, and therefore to fear, irritation and uncanniness (Duffy, 2003). It is suggested by several studies that in order to avoid this disappointment, the robot's design should be visibly artificial, or robotic. Duffy (2003) talks about transparent interaction, which means that a

robot's form should include only those features that enable social interaction with humans when it is required. It should be clear that the interaction is happening with a machine. Therefore, the hyper human-likeness, smoothness of moves and realistic behaviours should not be the priority in building human-robot interaction.

2.3. Perceived intelligence

While humans see familiarity in a human-like machine and have specific expectations coming from the interaction with the robot, they do not stop at evaluating the animacy and human-like appearance. The anthropomorphisation of many human-like robots, seeing their behaviours and hearing them speak, evokes thoughts and beliefs of the presence of intelligence. It of course works similarly between people. Numerous studies show that appearance and speech has an impact on people's judgements of the other person's intelligence (Duffy, 2003). Moreover, attractive people are more likely to be rated as more intelligent than others. This tendency to evaluate intelligence based on attractiveness changes, though, when one can hear the person speak and has an opportunity to rate intelligence on verbal cues (Duffy, 2003). It might mean that, in order to achieve perceived intelligence, the appearance of the robot loses (to some degree) its significance once a human-like robot is able to express itself in words. Here, it is important to note that in the weak AI stance the issue is not whether the system in question is actually, fundamentally intelligent. It is about whether or not this system displays attributes and features that promote or facilitate people's interpretation of it being intelligent (Duffy, 2003). In other words, one's evaluation of intelligence in a human-like machine circulates only around the question whether the robot 'seems' intelligent, for it will never actually 'be' intelligent, at least not in the general understanding of what human intelligence is. This thesis strongly supports this claim and will further discuss the notion of intelligence below.

2.3.1. Intelligence and consciousness in humans

To talk about perceived intelligence in robots, we have to first identify what it is with regards to humans. Human intelligence and consciousness are subjects of a never-ending debate among philosophers and neuroscientists. It is often the case in AI research to encounter the term intelligence being used interchangeably with the word consciousness. Some researchers omit the subtle differences between these two terms. Interestingly though, there is still no one specific definition for either of these words. It is unclear what intelligence or consciousness is, where it is located, how it works, or if it exists at all. For this reason people can mean different things when they call something or someone 'intelligent' or 'conscious'.

2.3.1.1. Consciousness

It is discussed whether humans are the only entities capable of having consciousness, and if not, then what are the factors determining whether a non-human being is conscious. In light of this, it can be debated whether technological artefacts could be conscious as well. Devices and digital systems have already proven to convince their users of their consciousness or humanity. For instance, a Google's software engineer Blake Lemoine was fired after he announced LaMDA chatbot to be sentient (Tiku, 2022). There are cases of people perceiving their smart vacuum cleaners as 'enlightened' (Heffernan, 2020). Similarly, humanoid machines are able to exhibit features which resemble that of a conscious being and make their user perceive them as intelligent or conscious. This brings different consequences for the interaction between a human and a machine. Without a direct evidence or clarity on the topic of consciousness, the first step for the HRI research is not to confirm or deny whether a machine is conscious, but rather to investigate how, why and if at all users perceive it as conscious, and what implications it brings (Scott et al., 2023).

It is however worth noting some of the views on the gist of consciousness and intelligence shared among scholars in order to better understand the direction followed by this thesis. In general, consciousness is considered to be something uniquely human, although it is sometimes extended to animals too. On the other hand, it is thought by some theorists of mind that consciousness cannot exist without experience (Chalmers, 1995), or that consciousness *is* experience (Koch, 2019). In his paper "Facing Up to the Problem of Consciousness" (1995), Chalmers repeatedly asks why a performance of a specific activity (for example of seeing something) is accompanied by the experience of it (i.e. the experience of seeing that thing). Why is the performance of seeing something *interpreted* into an experience of seeing it? To him, it is worth asking why physical processing, like seeing, gives rise to rich inner life at all. Chalmers (1995) suggests a view that being conscious means experiencing things and having any sort of reflection about that experience (interpretation). It can be any cue such as a colour (e.g. experiencing the blackness of an object), a shape, or a sound, which is interpreted into an experience. One experiences themselves and their life in a way no one else experiences themselves and their lives. This unique experiencing of things makes the 'what it is like to be' of an entity (Nagel, 1974; Chalmers, 1995). With such an understanding of consciousness, it is interesting to wonder whether or not machines can experience things and what it would mean for their condition.

2.3.1.2. Intelligence

Intelligence is a similarly difficult concept to describe and analyse. Often, it is viewed as an adjacent to consciousness (Scott et al., 2023). What kind of entity is intelligence? Is it a set of processes

happening in the brain, or is it a cultural invention? Currently, the most common use of the word 'intelligence' refers to

> *"some sort of overall mental capacity, and one that particularly highlights reasoning, problem solving, and abstract thinking".*
>
> *(Carson, 2015: 1)*

In the Cambridge Handbook of Intelligence, Sternberg (2020) highlights that human intelligence is examined through various metaphors of the mind, including the biological, geographic, and anthropological metaphors, among others. These metaphors shape the questions we ask about intelligence and guide our exploration of empirical phenomena. A metaphor is akin to a language – each metaphor represents a distinct way of expressing an idea. Therefore, when investigating intelligence from a cultural perspective, then one might turn to the anthropological metaphor (Sternberg, 2020). To grasp the connection between the brain and intelligence, it is useful to explore the biological metaphor. Incidentally, as noted by Brooks (1999), AI researchers often conceptualise the brain as a machine with electrical connections to sensors and actuators, suggesting the possibility of artificially replicating the brain. However, in reality, the brain operates within a "soup of hormones" (Brooks, 1999: 164), transmitting hormonally encoded messages throughout the body. This is frequently overlooked and underestimated in our 'electrocentric' society (Brooks, 1999).

The fact that one can operate under different metaphors of mind and conceptualise the studied phenomena accordingly makes it so challenging to have one particular explanation of intelligence. What is more, research across historical periods and cultures proves that specific features linked with one's overall mental ability can vary enormously, similarly to the sole importance assigned to intelligence as a characteristic of a group or individual (Carson, 2015). If we look at intelligence from a historical perspective, it can be found that intelligence in the modern Western world has been intrinsically linked to measuring it, with the most popular measurement system being the Intelligence Quotient (IQ) test (Carson, 2015). The willingness to measure intelligence comes from the desire to find the distinguishing factor between people and animals, or as one could put it – the factor proving humanity's superiority over other beings. The first standardised IQ test was developed by Alfred Binet and Theodore Simon around 1904 (Hally, 2015). It was designed for and worked best for children. The test measured practical matters in kids that were not taught to them in school, such as memory, attention and problem solving. Results showed that some children were able to answer more advanced questions compared to their age group, which gave rise to the concept of *mental age*. Originally, a person's mental age divided by their chronological age and multiplied by 100 determined one's IQ (Hally, 2015).

The use of the IQ test was sometimes questionable, to say the least. Probably the most alarming example is the use of intelligence measurements in the times of eugenics' peaking popularity (1900s – 1930s), when the IQ test was used to, for example, screen new immigrants entering the US through Ellis Island. The test (which was English only) was used to make utterly harmful generalisations and claims of "surprisingly low intelligence" of Southern European and Jewish immigrants (Hally, 2015: 2), who simply could not understand English. This led to racially biased migration restrictions and deportations of thousands of worthy individuals labelled as undesirable or unfit. Strikingly, this took place a decade before the beginning of Nazi Germany and Hitler's eugenics (Hally, 2015).

Although this intelligence measurement technology of the early 1900s spread quickly to different parts of the world, it did not solve the very basic questions about the nature of intelligence. According to Carson (2015), the existence of the IQ test did not give a clear definition of intelligence, and did not give answers to whether or not intelligence is hereditary, or whether it is influenced by the environment, and so on. Additionally, in case of such testing there is a cultural discrepancy. First of all, a test is brought from one culture to another, often with poor or inaccurate language translations of the items in the test, which naturally influences the test results (Carson, 2015). Moreover, IQ tests in most cases require certain knowledge, for instance mathematics, which also makes them culturally biased (Hally, 2015).

2.3.2.Intelligence in robots

Interestingly, while intelligence has been historically strongly linked to methods of measuring it, literature on perceived intelligence in machines is also more focused on the methods of measuring its levels, rather than about, for example, the implications of making a humanoid machine seem intelligent. Nevertheless, this thesis will concentrate on and try to understand the process and consequences of making machines seemingly intelligent. As mentioned earlier, the discussion over intelligence of a human-like robot can circulate only around whether or not the robot seems to be intelligent, and not whether or not it actually is intelligent (in the human intelligence sense). This approach to intelligence in a robot is supported in multiple studies, including Blut et al. (2020), who understand intelligence as the extent to which the machine *appears* to be capable of learning, reasoning and problem solving. It is also explicitly claimed that anthropomorphism increases user perceptions of intelligence not only in robots, but also other various smart technologies (Blut et al., 2020). Additionally, studies have shown that the more human-like appearance and behaviour of the robot, the more people expect it to be intelligent. In other words, people expect human-level intelligence from entities resembling humans. Likewise, people expect dog-level intelligence from a

dog-like robot (Krening & Feigh, 2018). What is more, a robot which appears to be too intelligent may be regarded as selfish or prone to weaknesses like humans and therefore less reliable, making it undesirable in the society (Duffy, 2003). These human expectations and reactions might direct one's attention towards the role design plays in building robots. As discussed earlier, a lot is up to the designers' decisions when it comes to managing and setting user expectations towards robotic companions.

Robots will be perceived as intelligent as long as they act intelligently, which is a tremendous challenge for roboticists. The difficulty lays largely in formalising human behaviour (Bartneck et al., 2008). This means that human behaviour would have to be convertible into a formalised formula, an algorithm. In turn, this implies that there has to be a standard set of human behaviours which suggest intelligence, and which could be inserted into the machine. With billions of people in the world, coming from different backgrounds, cultures, and so on, with personalities developing as they grow and age, it seems impossible to find a standard human behaviour displaying intelligence. What is more, even if it was possible, the ethical question remains of who and why would be the one to decide what a normal, standard set of 'intelligent behaviours' is. A strategy which according to Bartneck and colleagues (2008) is not a real solution to a problem would be to embed randomness of behaviours in a robot, to imitate a human better. While it would perhaps make sense in short human-robot interactions, in longer contact with the robot a user would learn the patterns and become bored with the limited random behaviours and vocabulary (Bartneck et al., 2008). The perceived intelligence of the robot would in this case acquire another meaning, or vanish completely. Or otherwise, the perceived intelligence would depend on the robot's competence (Bartneck et al., 2008).

Already back in 1999, Brooks suggested that intelligence of a machine is actually in the eyes of the observer, which leads to the notion of perceived intelligence. This indicates that both the environment of the AI system, and a perception of a viewer observing it, are key factors for determining its intelligence. The thing that influences the eye of the observer is the robot design. And, as this thesis tries to suggest, the core of most social robot design features lays in deception.

3. DECEPTION

What has been said so far about animacy, anthropomorphism, and perceived intelligence connected to social robots shows that robot design is based heavily on deception. When interacting with a human-like robot, people are deceived by its anthropomorphic features of appearance and behaviour, alleged animacy and cues of intelligence, and develop beliefs regarding its humanness. Clearly, the

human-like features or traits suggesting humanity are included in the robot design with a certain goal to achieve (Natale, 2021). The robot is human-*like*, not human, and the fact that it aims to spark certain thoughts, feelings and reactions in the user requires ethical analysis. At some point the discussion goes further than asking whether or not a certain machine is deceitful, but what the implications and consequences are of the deceptive factors (Natale, 2021; Coeckelbergh, 2017).

Before forming arguments for and against deception in machines, it is important to understand its general meaning and function. Interestingly, deception is a common practice within the animal kingdom, including humans (Wagner & Arkin, 2010). From a biological and psychological perspective, deceiving is a representation of an evolutionary advantage for the deceiver. Between animals, deceiving enhances chances of survival. It can be argued that it works similarly among humans. Deception is omnipresent in personal relationships, culture, sports and war (Shim & Arkin, 2013). These are contexts in which people deceive the other to take advantage and gain what they want.

There are numerous definitions of deception, with practically all of them circulating around manipulation of the other agent involved, which puts the concept of deception in a morally wrong light. Collins Dictionary describes deception to occur

> *"when someone deliberately makes you believe something that is not true"*
> *(Collins English Dictionary, n.d.).*

Matthias (2015) brings up a convincing approach towards the moral wrong in deception, which states that deception causes people to make different choices. Although divergent choices can arise from a completely neutral deception and even benefit the deceived agent, this creates ample opportunity for unethical practices, such as deceptive marketing.

Deception present between machines and humans is a peculiar type of deception deliberated over among scholars for many years. Already in 2003, Duffy (2003) asked fascinating and very accurate questions about perceived intelligence and machine deception in humanoid robots. In his paper, Duffy calls artificial intelligence in robots an *illusion* of life and intelligence, and an act of cheating. He asks,

> "*If the robot 'cheats' to appear intelligent, can this be maintained over time? Does it matter if it cheats? Is it important what computational strategies are employed to achieve this illusion?*"
> (Duffy, 2003: 178).

Relating robotic deception to illusion is an important and curious discussion, which will be analysed further below, drawing largely on the work of Mark Coeckelbergh (2017).

3.1. Performance, not deception

One of the main ethical concerns in the field of social robotics is linked to deception. The reason for this is that the core of human-robot interaction is based on anthropomorphic tendencies of humans. In turn, one can notice that the central aspect of technological anthropomorphism is illusion (Zawieska, 2015). This illusion is created by science and technology, and appears through 'tricks' performed by the machine which, essentially, fool us into believing the robot is like a person, is our companion, is an animal, or understands our feelings. On top of that, associating some technologies with magic is strictly connected to our inability to understand the technicalities of their workings, similarly to magic tricks (Natale, 2021).

In his paper, Coeckelbergh (2017) tries to analyse magic and illusion in the context of robots, and proposes a new approach towards deception, one that would evaluate problems connected to it through a more morally neutral lens. It can be understood that, while he acknowledges and sympathises with opinions which criticise and oppose deception, Coeckelbergh (2017) suggests that treating deception in information and communication technologies (ICTs) and robots as automatically negative might not be the way forward. Regarding social robotics, he states that condemning deception does not contribute to actually understanding what it is, and asks,

"*what would be a non-deceptive design and use of this technology?*"
(Coeckelbergh, 2017: 72).

If one tried to find an answer to this question, it could be said that in the case of social robots a non-deceptive design would mean that the social robot would cease to be social. It would lose its ultimate role of serving and existing among humans, as it would have to stop imitating any sort of behaviour or appearance of a human. In other words, it is plausible to suggest that deception is at the core of social robot design; without deception the robot is not able to fulfil its social role. However, what if the language of deception is not the right vocabulary to use in the first place?

What happens between a robot and a human could be, following Coeckelbergh (2017), compared to a magician and an audience, both participating in a magic show. A robot (the magician) is a product of designers and engineers who need knowledge of psychology and different techniques to successfully engage the user (the audience) in the interaction (performance). Interestingly, the magician is created to be and believed to be a supernatural character, while in reality he/she merely plays that role. Evidently, this is exactly what happens with our perception of some technologies. A robot is not capable of emotions or human intelligence, but the character we create out of it is (Coeckelbergh, 2017). During the performance, given all the cues and tricks, we start to believe things that are not the reality. Nevertheless, there must come a point at which the audience realises (by themselves or

by explanation) that what they experience is an illusion. After all, outside of the magic show, people are perfectly aware that it is only a deceptive performance. Coeckelbergh (2017) suggests a comparison to VR technologies. While there is nothing wrong with creating an illusion of a virtual world which can be experienced through a VR headset (in fact the illusion and being 'fooled' is desired), it should be made clear to the user that it is only an illusion. Undoubtedly, maintaining or retaining this awareness needs to be a crucial part of the technology design and promotion (Coeckelbergh, 2017). Thus, to relate it back to robotics, a user has to be made aware that a robot is a programmed piece of technology, not a person.

It is interesting and relevant to analyse the term *performance* further when talking about HRI and deception. As noted by Coeckelbergh (2017), it is important to notice that this interaction between magician and audience (robot and user) can be viewed as bidirectional. In other words, it is one performance in which both sides co-perform. The illusion does not happen only on the side of the magician; the spectator is also, most definitely, involved. What happens on the side of the audience while experiencing the acts of illusion is meaning-making, constructing narratives, and so on. Significantly, without the spectator there is no performance (Coeckelbergh, 2017). This claim can be directly linked to Rodney Brooks' (1999) viewpoint, in which he stresses the importance of the role of the already mentioned eye of the observer in human-robot interaction. A spectator observes the magic performance and evaluates the reality or authenticity of the illusion. Similarly, a user observes a robot's behaviours, and starts to accordingly perceive it as intelligent, creepy, human, and so on. One could contend that the eye of the observer determines everything. But what it determines is highly dependent on how the machine that the eye sees was designed to look and act.

Transforming the approach towards deception into *performance* terms might, according to Coeckelbergh (2017), be a better and a more encompassing perspective to take. The scholar argues that this is because there are multiple performances happening during a human-robot interaction, and deception is only one of them. When a person interacts with, for instance, a companion robot, it is possible to distinguish the performance of the user (who uses the robot in a specific way), of the designer or engineer (who creates and programs the machine), and of the robot itself (acting accordingly to the written code). All these different performances are part of a whole interaction process, they involve numerous kinds of techniques, artefacts and bodies. Coeckelbergh (2017) strongly suggests that describing this set of uses and experiences, which come from all sides of the interaction, as deception, reduces the rich configuration of performances to just one. What is more, using the term *deception* naturally gives ontological priority to only one specific performance and leaves the other ones less noticed. In HRI, *performance* can be decoupled from magic and illusion and

thus can become a morally neutral concept (rather than remain a negative or derogatory term) which happens between non-humans and humans (Coeckelbergh, 2017).

The suggestion that *deception* puts emphasis only on one side of the interaction, and in general 'steals' attention from the other agents, is true, especially if the phrase someone uses is "the robot deceives". The question one could ask here, though, is who/what does actually deceive? It can be argued that because of the tendency to anthropomorphise, people assign abilities and agency to a robot, including an ability to deceive. However, taking the weak AI stance, the robot itself is nowhere near being capable of deceiving a human. Perhaps a more suitable phrase to use would be "the design of the robot deceives", which switches the meaning and includes also the role of a designer.

3.1.1. Consequences of performances

Performances, however, do not come without problems. A robot's performance can cause different reactions in users (which is part of their performance) and can bring about various consequences. In Coeckelbergh's (2017) viewpoint, a performance can end with either a success or failure. For instance, in an interaction between a human and an emotionally expressive robot, it would be a success if the user saw emotions in the robot, instead of only a robotic imitation of those (Coeckelbergh, 2017). In other words, the performance is successful if the user becomes convinced of the performed, fake reality, i.e. the presence of human emotions in a machine. However, the performance can fail, either because of a changed context or time of the interaction, or when observed from the outside. To be specific, it fails when people think and claim that these emotions in the robot are not real (Coeckelbergh, 2017). One could say, then, that the emotions of the robot are (believed to be) real until someone breaks this belief and enlightens the others about the truth. Comparing it again to magic and illusion, in a magician's performance the acts look like real magic until someone says it is a scam or explains how the magic tricks work. What is more, experiences of the performance can differ between groups of users at the same given time. One group might perceive the performance as a success, while another might see it as a failure. This is to say that some people will believe the magic, and some will not. This distinction between failure and success of a performance takes on a much more serious meaning when narratives, or contexts, are taken into account. Specifically, the consequences of humans getting involved in performances alongside robots rise many ethical questions (Coeckelbergh, 2017). For instance, how ethical is it for an older person with limited cognitive abilities to be part of a performance of receiving care from a robot? Or, is it good that a young child is involved in a companionship narrative with a machine?

It can be argued that Coeckelbergh's approach actually implies and talks about deception, rather than anything else. He suggests that it is not a matter of deception, but of how successful a performance is. Yet, is the success not up to how believable the human-likeness of a robot is, and how tricky it is to distinguish from a real human? The performance is successful when the user 'buys into' the human-like reality created by the robot and its designers. The success, then, is based on deception.

3.1.2. Ethics of honesty and the role of a designer

As already mentioned above, it seems that the right way to manage technologies which can be deceitful is to inform the users about the state of reality. This kind of ethical practice circulates specifically around the virtue of honesty. In the case of social robotics, the ethics of honesty requires designers to be honest about the capabilities of their machine. It demands from them to create their robots in a manner that clearly indicates to users that it is a machine, rather than hiding the truth about how and what the robot is designed to do (Coeckelbergh, 2017). Interestingly, Coeckelbergh (2017) points out that this approach means that a designer has to take on a double role. On one hand, it is about designing and selling the main attraction in the robot, which is the magic and illusion. On the other hand, though, the designer is required to reveal the magic tricks, or at least inform that they take place. It is very often not the case on the robot market, where roboticists sell and advertise the illusion, without exposing the truth to the viewers (Coeckelbergh, 2017). To some extent it is understandable. If a magician him/herself admits that their show is not based on real magic but only on clever tricks, the attraction and curiosity of the viewer might decrease. However, when it comes to an interaction between a human and a social robot designed to be a human's support and companion, hiding or revealing the truth about its nature has a higher level of potential risks. In some cases, it is because there are human emotions at stake. For example, if a person gets attached to their companion robot, and feels the same from the robot, an error or accidental memory wipe can be emotionally damaging. On the other hand, if a care or mental support robot reveals its true nature and does not immerse its user into the magic performance (deception), it might be that it is then defeating its purpose. Through the machine not forming an emotional relation with a human, the human might not be willing to 'cooperate' and accept the robot's help.

That being said, Coeckelbergh (2017) proposes that the discourse should be about the kinds of performances we want and do not want, instead of keeping the conversation around what is an illusion and what is real. For instance, we might actually want robots to perform friendship with us, but perhaps only with adults and never with children. This approach suggests that there are cases when being, in fact, deceived is desirable or generally accepted. The question then remains of how

exactly we want to be deceived. Following Coeckelbergh (2017), however, the conversation should be formed around the ethics of specific scenarios of interaction, without using the language of deception. In other words, what the field of HRI needs is an ethics of performance. This takes the perspective of the ethics of honesty further and implies that the robot designer should not only be transparent about the machine's capabilities, but also take responsibility for the kind of performances their design enables and the consequences coming from it. The main ethical issue for designers, then, is not whether or not the machine is fooling the user, or whether or not there is illusion involved. It is generally known and accepted that trickery takes place, and that we in fact witness a show. The main ethical question asked about the robot design should be, given the techniques and tricks used by designers, and given that designers co-shape and co-perform in the interaction, how might we ensure the performances and their consequences are good (according to a specific definition of 'good') (Coeckelbergh, 2017)?

One can notice that since the theory of performance assumes that different sides of the HRI co-write narratives and co-perform them, the responsibility lays also on the part of the users. It means that the designer does not have a full control over what happens within the performance and over its ethical qualities (Coeckelbergh, 2017). In other words, if we reject the deception approach in which the user is passive and 'falls victim' of a deceptive robot, then the user should also be held responsible for the performance. This in fact implies that both the designer and the user have limited responsibility. Both sides need to accept that what happens during the performance is not entirely under their control. Accepting this means accepting unintended, unforeseen consequences. The role of the designer is to enable various kinds of uses through the robot design, as well as to ensure that a large scope of potential unintended uses are taken into account (for example by analysing all kinds of worst case scenarios) (Coeckelbergh, 2017). While Coeckelbergh (2017) does not clearly state what exactly would be the users' responsibility, it can be concluded that since every user is different, their varying intentions and desires impact the way they use the machine. By this, each user is responsible for using the robot in their own way. For them, the interaction is what they make of it (after all, the way the robot is is in the eye of the observer (Brooks, 1999)). The designer should make space for these diverse interactions, as long as consequences coming from them are not negative, in order to ensure a successful co-performance.

3.1.3. Limitations of the performance metaphor

While it is true that using a metaphor of performance, instead of the language of deception, to describe what happens in human-robot interaction gives the whole process a much more morally

neutral tone, it might seem to some that it is only a 'cover-up' of deception. It is also true that, by stating that the users are being deceived by a robot, they are made much more passive, while the performance metaphor gives the user agency and responsibility. It can be argued that Coeckelbergh (2017) proposes to use the term *performance* because in his view it is time to accept that deception is and will be present in the social robot design. Hence, it makes the discussion much more fruitful if we do not deliberate whether or not people are deceived, but rather what kind of deception we are willing to accept.

On the other hand, though, the performance metaphor has significant limitations and contradictions. For instance, when a human performs a certain role, it implies that there is also a hidden self of that person; when performance ends, the person comes back to his or her 'not-performed', real self. Is this also the case for a robot? Does the robot have a hidden true self, which it comes back to after performing a specific role with a human? Unhesitatingly, no. This disparity can make one wonder if what the robot does is also then still performance, or something else. Moreover, it can be argued that executing a performance suggests a conscious, intentioned action of pretending or acting, which a machine cannot achieve due to lack of consciousness, awareness, and so on. Another problematic aspect of the performance metaphor is that performance is normally bounded by time and space. Usually, a performance ends after a specific amount of time. Coeckelbergh (2017) himself suggests that a performance *ends* with either failure or success, indicating that there is a clear beginning and end to performances. Because of this time boundedness of performance, the metaphor is accurate only in some contexts. It means that some social robots do not fit with the performance theory, depending on their intended function. A performance that is ongoing for a long period of time, for instance a few months or years as could be the case with care and companion robots, is not a performance anymore but a relationship. Normally, care relationships and companionships are not meant to end quickly, in contrary to performances. It is clearly apparent in studies on the Roomba robot. It has been shown that after a long time of having floors cleaned by the Roomba, its owners would start feeling a need to return the favours (van Wynsberghe, 2021). This example suggests that over-time the machine's performance turns into something much more. It results in moral reactions in users and a type of emotional attachment characteristic for human-human relationships. On the other hand, an interaction lasting minutes or hours with specific goals to complete, for example delivering useful information to hotel customers by Connie the robot, could indeed be called a performance of a role. Any bond with a user is not exactly established in that case and what matters is a satisfied customer. It can be claimed, therefore, that while it is valid to call short human-robot interactions performances, it cannot be applied to longer interactions.

The above discussions show that social robot design, and therefore the shape of human-social robot interaction, rely on the robotic features which are supposed to make users believe, or be immersed, in real intentions, real emotions, or real face expressions, of the machine. While it is not a problem in itself to become immersed in a robotic illusion, it can have serious consequences on human mentality.

3.2. The case of reciprocity as a result of deception

Upon closer inspection of Coeckelbergh's (2017) stance on deception in HRI, a question arises as to whether everything can be performed. It is an extremely difficult matter to unravel. If we take values and morality into account, then nothing in social relationships can be and should be performed, especially not emotions towards another. On the other hand, it can be argued that indeed everything can be performed, as it is the case for example in a theatre. In HRI, though, the human is not supposed to believe that the robot merely performs, but on the contrary. One of the examples of a social concept which, as it will be justified below, a machine cannot co-perform is reciprocity, which roboticists are increasingly designing for (van Wynsberghe, 2021). Moreover, it is possible to contend that reciprocity serves as an umbrella term for other social concepts of mutual exchange of actions and emotions between people, for instance friendship and love. This might mean that, because of the deceitful nature, robots cannot co-perform any social role on the emotionally true level.

In the case of social robots, it can be therefore argued that using them as carers or companions is akin to deception, not performance. Sparrow and Sparrow (2006) claim that this is because robots rely on people's belief that robots are something that they are in fact not. Social robots which are designed to provide care to people deceive their users into thinking and feeling that this care and love are true, while the truth is that these beliefs are false (Sparrow & Sparrow, 2006). A robot cannot truly care and love, it only acts as if it does. Following Coeckelbergh (2017), this sort of deception is morally wrong, since it makes people fail to view the world accurately. In addition, some may argue that this kind of human-robot interaction is a one-way transaction. In her research, Aimee van Wynsberghe (2021) analyses the issues of reciprocity in human-robot interaction from the perspective of care ethics. The tendency to feel the need to return favours, i.e. reciprocity, has been noticed by robot engineers as a pro-social human behaviour and a very important mechanism in interactions among humans. It is thought of as the key characteristic of moral life across many disciplines. For this reason, reciprocity is increasingly taken into account in studies on HRI and social robot design for care and therapy (van Wynsberghe, 2021). In particular, it has been examined to what extent the social rule of reciprocation is actually present in HRI (Moberg, 2023), and stressed that reciprocity might be used as an instrumental value to manipulate and enhance the robot's acceptability (Gill, 2022).

As mentioned above, it can be argued that the interaction between a social robot and a human, in this case especially the one based on care, works only one way. Between humans, care is a bidirectional concept which also aligns with reciprocity (van Wynsberghe, 2021). While there are care-givers and care-receivers, both performing their own roles and actions, there is also a relation between two agents when it comes to reciprocity. Namely, if someone gives one a favour, the other person wants to do something similar in return. When it comes to reciprocity between a human and a robot, it takes quite an 'irregular' form. First of all, the 'robot needs' which a human is supposed to reciprocate are very different from those of a human. These could be software update, changing parts or recharging the battery. Because humans and robots have discordant needs, reciprocity towards a human is not the same as reciprocity towards a robot (van Wynsberghe, 2021). What is more, a general approach in HRI is that the robot is subservient and obedient to the human. This already stands in opposition to the paradigm of reciprocity present in human-human interaction (HHI). In HHI, desirably, both agents are equal to each other (van Wynsberghe, 2021). Given all this, one could ask whether interacting with social robots could eventually lead to humans seeking relationships with other humans in which one of the sides plays a subservient role, like a robot.

Similarly to the perspective taken in this thesis on *perceived* intelligence in machines, van Wynsberghe (2021) also approaches robots as *perceived* as social, rather than actually being social agents. One might posit that any robot can be perceived as social, as long as it has embedded features which can make it seem social in an interaction with a human. To create this perception, the foundations of human-robot interaction are most of the time built on and heavily inspired by the human-human interaction (van Wynsberghe, 2021). What is more, because of the social robot's design and its programmed behaviours, people can project their own beliefs regarding the robot's social skills onto it. In the already mentioned studies concerning the robotic vacuum cleaner Roomba, it has been found that after an extended period of time during which the robot meets user needs (i.e. keeping the floors clean), the users start to treat Roomba as deserving of reciprocity. In other words, because a machine keeps working for its users every day for a long time, it makes the users feel an urge to do something nice for the Roomba in return. This is all while the Roomba does not even know about the existence of its owners, or that it has owners at all in the first place (van Wynsberghe, 2021). Here, one can again find a significant influence of anthropomorphisation and perceived animacy steering the users' feelings towards a robot. It is apparent in the Roomba case that a machine, even a very simple one presenting traces of humanness, is able to put a human in circumstances in which the human starts to think that they should return all the good and favours provided to them by the robot. Instead of recognizing that the machine simply fulfils the given commands or performs the role it was

programmed for, users are subject to deception and develop a need to reciprocate (van Wynsberghe, 2021).

Relating it back to Coeckelbergh's (2017) performance theory, in the case of reciprocity the co-performance is definitely successful, at least at the first glance. To put it simply, the robot performs a role of a caregiver, and the human acts as the care-receiver. If the user wants to give back favours to the machine, it means that the level of deception was sufficient (successful performance), because the user came to believe that the robot helps or cares for them with genuine intentions. This can be achieved through various design methods and tricks manifested through anthropomorphic features of the robot. However, giving back favours to a machine does not seem to be logical and rational. First of all, exchanging favours is a human behaviour that comes from deep within, and it is a great human value (van Wynsberghe, 2021). The machine acts a certain way because it has been programmed to do so, hence co-performing reciprocity with it is not possible for it has no genuine fundaments. One could even argue that in case of the human-robot interaction and reciprocity, the returned favour has no receiver. In other words, a human who feels the need to return a favour to a robot in fact wants to return it to an unaware, non-living object. A robot is not a conscious, moral person who can perform favours and accept something in return.

It can be claimed based on this, and following van Wynsberghe's (2021) point of view, that the value of reciprocity cannot be achieved in HRI. The main reason for this is the coercion to deceive the user and the resulting lack of genuinity. Reciprocity requires emotional understanding and empathy, which robots (as of now) do not have.

3.3. The impossibility of a genuine relationship

"To say "I love you" as a human may or may not be truthful, but it becomes necessarily deceptive when uttered by a machine, since the machine is lacking the corresponding mental state."

(Matthias, 2015: 175)

The above quote is a great illustration of the main conclusion this thesis arrives at, namely that it is impossible to have a genuine relationship with a machine. What is meant by a genuine relationship is a reciprocal relationship, or a bidirectional, authentic and intentional exchange of impressions and feelings going on between two agents. To explain this concept better, let us compare once again human-human interaction to human-robot interaction. Specifically, let us take companionship as an example – a type of relationship between people which is attempted to be mimicked by robots.

Presumably within the prevailing consensus, companionship could be defined as a supportive relationship between individuals who provide each other with a sense of connection, emotional support, mutual care. Usually, it involves offering mutual understanding and support, spending time together, sharing passions and experiences. It always goes, or should go, in both directions. Additionally, one of the pillars, or values, of companionship is authenticity. As Matthias states

"*Authenticity in relationships is a human purpose*".

(Matthias, 2015: 170)

This view can be interpreted that reaching authenticity in relationships is what people strive for in their lives. Authenticity in this case can be understood as being genuine, real, true, and not pretending towards the other. Having someone authentic as a companion is regarded to be a great value. It can be argued that a relationship cannot be built, or is destined to failure, if both of the individuals lack authenticity.

Recreating companionship in human-robot interaction based on the dynamics observed in human-human interaction is challenging arguably for only one reason: the robot cannot feel. If we take authenticity as an example of a significant trait in companionship, and the fact that robots cannot feel, forming companionship with a robot is not possible. Companionship involves exchanging emotions, which the robot does not have. The robot pretends (deceives) to have these emotions, which automatically makes the relationship not authentic, not genuine, not reciprocal, and so on. Companionship based on deception is not companionship. Friendship based on deception is also not friendship. The above argument shows that forming any kind of bidirectional relationship with a machine is not possible.

One could wonder why it matters that a genuine relationship with a robot is not possible. The simple answer could be that it does not matter as long as the user knows the truth. If the user is aware and accepts the fact that their relationship with the machine is unidirectional, then the potential harm is perhaps less. This will be deliberated over in more depth in the following chapter.

3.4. When is deception morally permissible?

Some voices raised in the discussion about deceitful robots question the immorality of deception. Does it matter that the robot deceives its user? Does it matter that someone is made to believe something untrue while having a great interaction with the machine? Is all deception bad? Constructing an argument supporting the role of deception in HRI requires analysing how deception operates within human-to-human contexts, especially because operations and roles of social robots

are inspired by interactions between humans. To illustrate and argue for the moral permissibility of deception, it is useful to take a look at the case of healthcare and elderly care. In the context of human-human interactions in these environments, it is evident that deception is a prevalent and accepted practice, frequently motivated by the best intentions (Schermer, 2007). For example, in the context of dementia care, staff members may foster a belief among clients (patients) that they in fact come to work every day, giving them a sense of purpose and enjoyment of social interactions. Additionally, the staff can engage the clients in what they perceive as useful tasks, for instance to fold towels which were previously intentionally disarranged by a staff member (Matthias, 2015). In light of this, some caregivers may agree that deception could be deemed acceptable if it is in the best interest of the care receiver. What is more, perhaps elderly care is an example of a setting in which deception is actually necessary and required to perform the best care (Segers, 2022). Following this, a social robot could be classified similarly when it engages in deception by imitating human features and creating an illusion of emotional interaction, provided it is for the well-being of the user. This also means that, for example in such instances, there is a potential for an improvement of the overall quality of life through engagements with a deceptive robot (Coeckelbergh, 2015).

Through exploring issues of trust, autonomy and erosion of trust resulting from deceptive robot behaviours, Matthias (2015) proposes a set of requirements needed to be met in order for deception in robots to be morally permissible, or even desirable, in the healthcare setting. First of all, deception must serve the best interests of the patient in order to maintain their trust. It means that if a care robot performs any form of deception, it should be for the well-being of the patient and not for any harmful or wrong purpose. The patient's trust should not be violated through deceptive actions, and any deception should be carried out with the patient's well-being in mind. Matthias (2015) claims that if it is evident that a deceptive behaviour serves the patient's interests, it should not be considered a breach of trust, which in turn makes the particular deceptive behaviour morally permissible. Secondly, deception has to be used to increase the person’s autonomy by allowing them to make their own choices with regards to their own values and by increasing the level of control the patient has over the robot. Thirdly, to some extent, deception must be made transparent. The machine has to make it clear that the behaviours which are taking place are deceptive. Furthermore, the user should have the capability to stop any deceptive actions by signalling it to the robot. Conversely, in situations where the user is willing to suspend disbelief in the machine's abilities, the machine should refrain from revealing truths that the user may prefer not to confront. Finally, deception cannot lead to an actual harm. If the robot detects that the patient relies on it to do things it has no capacity or ability to perform, then the machine has to clearly signal it to the patient. For example, an actual harm resulting from deception could happen when, through the robot’s deceptive behaviours, the patient is made to

believe that the robot can remind them to take their medication, while it in fact cannot (Matthias, 2015).

It can be noticed, quite positively, that these requirements stand on a fundamental premise that the well-being of the patient is the ultimate good. The users' 'wishes' regarding how they want their interaction with the robot to look are of great importance. It is therefore crucial to understand what user expectations towards a robot are, which is a challenge for HRI designers. In the context of a care robot, does a patient expect it to unconditionally speak the truth, or is the robot expected to offer comfort and contribute to the patient's recovery? Should the robot therefore constantly destroy the illusion of genuinity and care, or provide comfort to patients through being deceitful about its own emotions? It is definitely dependent on the context of interaction, and also varies from case to case. Matthias (2015) refers to an example of the case of not wanting to know one's full diagnosis. People from different backgrounds and cultures approach learning about their health condition differently, with some people actually clearly asking the doctor to, to some extent, deceive them about their diagnosis. Hence, one could conclude that whether or not deception is morally permissible depends quite largely on whether or not the user wishes for it to happen.

4. SUMMARY ON THE ETHICAL ISSUES OF HRI BASED ON DECEPTION

The aim of this chapter is to summarise findings from the above sections and to deliberate over ethical issues connected to deception in social robots. Overall, the thesis tries to argue for that deception lays at the core of social robot design. If this is the case, then it is important to investigate and understand consequences coming from it. Clearly though, the role of the deceptive design differs from case to case and can be implemented with different intentions in mind. This means for example that deception, while it is a notion of negative connotations, might be used with an aim to improve or work towards the well-being of the robot's user. As was discussed above about permissible deception cases, there are instances when deception is wanted and desired, and can be morally allowed with certain design requirements in mind.

The question of when deception is wrong, on the other hand, has been a long-lasting debate in the field of social robotics. It might seem that deception in itself is not a problem, or that it is not a harmful context to be in for a robot user, at least not in the current state of robotics. It is true in some contexts, since some deceptions are simply harmless fun (Sharkey & Sharkey, 2020). A robot which welcomes and answers questions of hotel clients, such as the Hilton robot concierge Connie, can be regarded as just an attraction and an interesting addition to the whole customer experience. It does not really

matter if people interacting with it believe it is intelligent, alive, whether it understands emotions or expresses human emotions itself. On the other hand, the human-likeness and anthropomorphic traits of a robot like Connie could be used for morally wrong purposes, such as talking customers into buying more products and services which they do not need. Nevertheless, as has been found on the course of developing this thesis, deception in social robots becomes an ethically challenging issue once it is practiced in interactions with vulnerable groups over extended periods of time.

This thesis focused quite largely on the notions of anthropomorphisation and perceptions of animacy and intelligence caused by design features of social robots. Elements of these concepts are ethically problematic and questions they spark are interesting to unravel. Arguably, the main issue which causes ethical dilemmas is the robot's imitation of a human form. In general, it can be stated that once the robot takes a human-like form, the ethical problems start to arise. The fact that social robots are supposed to and do resemble a human, of course to a different extent, is questionable on many levels. What is more, the robot's design features do not have to be complicated or advanced to spark reactions in users which can be seen as ethically undesired. For instance, even as little as displaying robot's animacy through its head movement can be a starting point for a user's emotional approach towards the machine. Although it is now quite unlikely on a large scale, in the future it can bring us to a point where human life will be seen as equal to the life perceived in a robot. While this thesis has not discussed this future scenario, and whether this would be morally right or wrong, although perhaps majority of society would argue for it being definitely morally wrong, it is certainly a problem which should already be taken into account in the development of social robotics.

It has been argued in this thesis that the robot's perceived animacy and a human-like appearance trigger anthropomorphisation. On one hand, it is claimed that this tendency to attribute emotional states and human traits to objects is an attempt to make sense out of and rationalise non-human behaviours. On the other hand, a case could be made that the robot's animacy and human-likeness incite one's imagination, which make one anthropomorphise and believe in something the robot is not. Nevertheless, anthropomorphic technologies, especially social robots, pose significant ethical concerns. It is questionable whether utilising the inevitable human tendency to anthropomorphise in human-robot interaction design is to a real benefit of the user. It is not difficult to envision a human-like machine capable of eliciting empathy or care in its user solely for the benefit of its creator company. For instance, this could occur through private data collection or nagging for the purchase of unnecessary products. This brings a very significant ethical problem connected to anthropomorphism in social robotics. As was already mentioned above, it has been proven that certain groups of people have a stronger tendency than others to anthropomorphise. This is to say that there are differences in how people of varied psychological profiles react to robots. A lonely person, for instance, will ascribe

more human features and emotions to the robot than a person who is in stable social relationships. While still a lot of research needs to be done in this area, it is important to already use this knowledge in designing interactions between robots and humans.

This thesis claims that in an interaction with a social robot humans tend to go further than simply perceiving aliveness and human-likeness. People see intelligence similar to that of a human when in contact with a social robot. Perceiving intelligence in a robot is often induced by the robot's speaking abilities or expression and recognition of emotions. However, assigning intelligence to machines, AI-based technologies, or AI itself, is problematic since there is still no agreed definition of what intelligence actually is. This is to say that people often mean different things when they call something 'intelligent'. Nevertheless, an entity which looks and behaves like a human is expected to hold some level of human-like intelligence. These expectations, if unmet, disrupt the human-robot interaction, affect user experience, and might be one of the causes of the presence of the uncanny valley. It is similar to expectations coming from an interaction with a very human-looking robot which, for example, presents inaccurate human body movements, causing a sort of disappointment and eeriness in the viewer. Another issue with perceived intelligence, and more specifically with designing a robot in a way that it seems intelligent, is that it encompasses standardisation of behaviours. This means that in order to make a robot seem intelligent, it has to have a specific set of embedded behaviours which display intelligence. The question of which behaviours are perceived as intelligent, given that it can vary from culture to culture, and of who decides which behaviours display intelligence is yet another ethical problem to further explore by scholars in the field. Undoubtedly, this also applies to the design of a robot's appearance and deciding on which appearance standards to follow.

As it is clear from the above, every element of 'life' in a social robot has to be programmed or designed. Every element of a robot's existence is artificially made, including its animacy, intelligence, appearance, behaviours, facial expression, tone of voice, the things it says, and so on. Evidently, all these features are included in the robot design in order to make it resemble a human. The social robot is supposed to pretend to be human, or almost human, and play a specific social role. The pretending seems to be at the centre of the robot's design, and the user failing to detect this 'hoax' is one of the indicators of success for the robot's creators. Therefore, it is claimed that deceiving the user into thinking the robot is alive, thinks, or emotionally reacts, is at the core of robot design. This thesis also brought up the concept of the term *performance* to be used to replace *deception*. One of the main takeaways from this discussion is that indeed everything can be co-performed, but the user should be aware of it happening and willing to participate. In other words, it is crucial to make the user aware of the deception taking place. What is more, this thesis arrives at a claim that it is impossible to form a reciprocal, genuine relationship with a social robot, whose core of functioning is based on deception.

4.1. Discussion

Evidently, every facet of a social robot's 'life' is deliberately designed with and aim to emulate a human. The question of why this is, i.e. why people desire to develop a robot which would resemble a human to the tiniest detail, always receives a different answer. In the engineering world, one can notice strong tendencies to build an as-human-as-possible robot with the highest desire being to make the user forget it is only a machine. Developing a human-like robot resembles a fun challenge to build an advanced toy. It also appears to be a dream of many roboticists to build a robot with personality, a soul, essentially to build a robotic friend and make it alive (Thomas Burns, 2023). It is often argued that this desire is motivated by the fact that a human-like machine brings better user experience, since interacting with a human form is the most familiar to all humans. Therefore, creating and interacting with a human-like robot brings the most incredible and fun experiences. Taking a different perspective, though, this approach might be perceived as lacking creativity and being limiting to what robotics could achieve. According to HRI scholar Kate Darling interviewed in the Lex Fridman Podcast (Lex Fridman, 2022), it is unclear why a robotic companion has to look like us. On the course of the interview, Darling claims that the human form is actually not necessary to create the robot's social component. She claims that making a robot look like a person is a 'boring' approach and suggests that robot design could get much more creative without losing the satisfying level of user experience. For instance, she challenges the fact that social robots have two arms instead of three, or that they do not move on roller skates (Lex Fridman, 2022). This unique comment suggests a vast potential of social robotics and a view which is interesting to explore further. It goes hand in hand with scholars such as Duffy (2003), who claim that a human-like form of a robot might actually be limiting or constraining the technology's potential.

Nonetheless, while the use of human form can be justified in robots involved in interactions with humans, it does not have a justification in other cases where the robot is not meant to interact with humans. A perfectly odd example are robots designed to work in factories, such as Tesla's Optimus. According to Kate Darling (Lex Fridman, 2022), it is highly unsustainable, short-sighted and redundant to apply human form to robots which are supposed to work in factories and warehouses, as these environments are and will be designed increasingly less with humans in mind. While the human shape makes sense in places currently specifically designed to accommodate humans, such as an aircraft or a submarine, some sites such as a factory do not require a machine working there to be shaped like a human (Lex Fridman, 2022). It is understandable that people want to delegate repeatable or dangerous tasks to machines, but why do we want them to be humanoids?

5. CONCLUSION

What has been said in this thesis could be shortly concluded as follows: a *social robot* without deceptive features ceases to be social. Without deception, it cannot achieve its main purposes and roles. It is for the deceitful design features that the robot can seem alive, intelligent and human, and can through this influence the user interaction. Without deception, the user would not be able to become immersed into the interaction with the robot, which would in turn make it impossible to build a bond and allow the robot to fulfil its role as a companion.

To put it differently, roboticists design *social robots* with the intention of providing people with a new friend or companion. For this undertaking to be successful, individuals need to believe, even if only to a small extent, that the robot is something more than just a machine. There is always a potential for a user of a social, companionship robot, to become deceived, especially in case of people of higher vulnerability. In my judgement, achieving this requires incorporating deceptive features into the robot—whether through human-like speech, a pair of expressive eyes, emotion imitation, and so forth. This is the reason to claim that deception is at the core of social robot design. Additionally, if social robots are built on deception, it is then evident that there is no possibility of establishing a bi-directional, authentic relationship between a robot and a human.

The above claims were made on the basis of an extensive research on ethics of social robotics. The research was aimed at determining the extent to which design features of social robots cause and steer the level of deception in human-robot interaction, and the impact it has on the formation of human-robot relationships. Firstly, the position on deception was extensively clarified, specifically exploring matters of intentionality and aligning with the perspectives of Sharkey and Sharkey, who argue that individuals can be deceived regardless of intent. Additionally, reference was made to marketing materials from certain social robot companies indicating that these machines are developed to possess 'personality' or to become a human's friend. This was followed by assertions found in Danaher's that social robots will need deceptive capabilities for seamless integration into society. The concept of embodied AI was then introduced in order to emphasise the societal relevance and importance of research on human-AI interaction, specifically human-robot interaction. The discussion then moved on to robot design elements, in particular the notions of animacy, anthropomorphism and perceived intelligence, as I regard them to be the most fundamental in achieving deception in HRI. From this, the thesis moved on to a broad argument about deception, analysing and criticising particularly Mark Coeckelbergh's perspective on approaching deception in a context of performance. A discussion over the problem of reciprocity in social robotics was also considered, which was followed by a claim of the impossibility of reciprocal human-robot

relationships. Towards the end, the thesis presented the existing positive approach to deception in HRI and talked about requirements necessary for deception to be morally permissible. Finally, the research was concluded with a summary of ethical concerns regarding deceptive features of social robots.

What I have learnt on the course of exploring the topic of this thesis is that some of the main ethical concerns regarding deceitful robots are rooted in the capitalistic tendencies. Applying deceitful elements is often aimed at acquiring profit from private data gathering. Through deception, it is possible to evoke trust in the user, who will be more willing to confide in the machine and allow it in their private parts of life. While I do acknowledge the fact that there are cases when deception is permissible, I believe more research should be done to find solutions to merge the desirable deceptive robotic behaviours with private data protection. Additionally, more attention should be given to the possible malicious manipulations of users resulting from robotic deception. On the other hand, though, as anthropomorphism and the perception of intelligence are prevalent not only in social robots but also in various other technologies (e.g., smart home assistants, digital avatars), it would be intriguing to explore the possibility of recontextualising deception. This could involve considering alternative terms and examining it in a more nuanced way than, for example, Coeckelbergh's suggestion of performance. Moreover, a view worth ongoing promotion emphasises that social robots should be in the role of support, not replacement, of the companionship and care between humans, a view Aimee van Wynsberghe strongly advocates for. Another ethical matter directly linked to this thesis' topic which should be explored in further research is the problem of forming attachment with non-living entities. Since human-robot interaction with robotic companions has, and intends to have, similar traits to human-human social encounters, as a result human-robot relationships might contain features of attachment. Although there is a lack of rigorous definition of attachment in HRI, warnings about forming attachment with robots and unethical situations resulting from it are pervasive in the field. Clearly, this kind of matters and questions are some of the burning issues in our AI-driven world, a world increasingly concerned with the future of living among artificially intelligent beings.

References

Ameca (2024). *Engineered Arts*. Available at: https://www.engineeredarts.co.uk/robot/ameca/ (Accessed: 27 January 2024).

AP News (2023). *UN Tech Agency Rolls Out Human-looking robots for questions at a Geneva News Conference*. Available at: https://apnews.com/article/humanoid-robots-better-leaders-ai-geneva-486bb2bad260454a28aaa51ea31580a6 (Accessed: 12 October 2023).

Avatar iPal Robot Family (2017). *iPal Robot*. Available at: https://www.ipalrobot.com/ (Accessed: 27 January 2024).

Aydin, C., (2021). *Extimate Technology*. Taylor & Francis.

Bartneck, C. et al. (2008). 'Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of Robots', *International Journal of Social Robotics*, 1(1), pp. 71–81. doi:10.1007/s12369-008-0001-3.

Bartneck, C. et al. (2009). 'Does the design of a robot influence its animacy and perceived intelligence?', *International Journal of Social Robotics*, 1(2), pp. 195–204. doi:10.1007/s12369-009-0013-7.

Bartneck, C. et al., (2020). *Human-Robot Interaction. An introduction.* Cambridge University Press.

Blut, M., Wang, C., Wünderlich, N. V. & Brock, C., (2021). Understanding anthropomorphism in service provision: a meta-analysis of physical robots, chatbots, and other AI. *Journal of the Academy of Marketing Science.*

Brink, K.A., Gray, K. and Wellman, H.M. (2017). 'Creepiness creeps in: Uncanny Valley feelings are acquired in childhood', *Child Development*, 90(4), pp. 1202–1214. doi:10.1111/cdev.12999.

Brooks, R. A., (1999). *Cambrian Intelligence. The Early History of the New AI*. Cambridge, Massachusetts: The MIT Press.

Carli, R., Najjar, A. & Calvaresi, D., (2022). *Human-Social Robots Interaction: The Blurred Line between Necessary Anthropomorphization and Manipulation.* Christchurch, New Zealand, Association for Computing Machinery.

Carson, J. (2015). 'Intelligence: History of the concept', *International Encyclopedia of the Social & Behavioral Sciences*, pp. 309–312. doi:10.1016/b978-0-08-097086-8.03094-4.

Chalmers, D. (1995). Facing Up to the Problem of Consciousness. *Journal of Consciousness Studies*, 2(3), 200-219.

Coeckelbergh, M., (2015). Care robots and the future of ICT-mediated elderly care: a response to doom scenarios. *AI & Soc*, pp. 455-462.

Coeckelbergh, M., (2017). How to describe and evaluate "deception" phenomena: recasting the metaphysics, ethics, and politics of ICTs in terms of magic and performance and taking a relational and narrative turn. *Ethics and Information Technology*, p. 71–85.

Collins English Dictionary (no date). *Deception*. Available at: https://www.collinsdictionary.com/dictionary/english/deception (Accessed: 22 June 2023).

Cornelius, S. & Leidner, D., (2021). Acceptance of Anthropomorphic Technology: A Literature Review. *Proceedings of the Annual Hawaii International Conference on System Sciences.*

CyberGuy (2023). *The next generation of Tesla's humanoid robot makes its debut*. Available at: https://cyberguy.com/future-tech/next-generation-teslas-humanoid-robot-makes-debut/#:~:text=Optimus%20Gen%202%20is%20the,construction%2C%20healthcare%2C%20and%20entertainment. (Accessed: 27 January 2024).

Damiano, L. & Dumouchel, P., (2018). Anthropomorphism in Human–Robot Co-evolution. *Frontiers in Psychology*, 9.

Danaher, J. (2020). 'Robot betrayal: A guide to the ethics of robotic deception', *Ethics and Information Technology, 22(2), pp. 117–128. doi:10.1007/s10676-019-09520-3.*

DiSalvo, C., Forlizzi, J., and Gemperle, F. (2004). Kinds of Anthropomorphic Form., in Redmond, J., Durling, D. and de Bono, A (eds.), *Futureground - DRS International Conference 2004*, 17-21 November, Melbourne, Australia. https://dl.designresearchsociety.org/drs-conference-papers/drs2004/researchpapers/45

Duffy, B. R., (2003). Anthropomorphism and the social robot. *Robotics and Autonomous Systems,* Volume 42, pp. 177-190.

Everitt, T., Lea, G. and Hutter, M. (2018). 'Agi Safety Literature Review', *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence* [Preprint]. doi:10.24963/ijcai.2018/768.

Fink, J. (2012). 'Anthropomorphism and human likeness in the design of robots and human-robot interaction', *Social Robotics*, pp. 199–208. doi:10.1007/978-3-642-34103-8_20.

Gill, K.S. (2022). 'Autonomous Reciprocity: Context Matters', *AI & SOCIETY*, 37(2), pp. 415–416. doi:10.1007/s00146-022-01419-w.

Hally, T.J. (2015). *A Brief History of IQ Tests*.

Heffernan, V. (2020). *My Roomba Has Achieved Enlightenment.* WIRED. https://www.wired.com/story/roomba-robot-consciousnessenlightenment/

Kahn, P.H. et al. (2006). 'Robotic Pets in the lives of preschool children', *Interaction Studies*, 7(3), pp. 405–436. doi:10.1075/is.7.3.13kah.

Kim, K.J., Park, E. and Shyam Sundar, S. (2013). 'Caregiving role in Human–Robot Interaction: A Study of the mediating effects of perceived benefit and social presence', *Computers in Human Behavior*, 29(4), pp. 1799–1806. doi:10.1016/j.chb.2013.02.009.

Koch, C. (2019). *The feeling of life itself why consciousness is widespread but can't be computed*. Cambridge, MA: The MIT Press.

Krening, S. and Feigh, K.M. (2018). 'Characteristics that influence perceived intelligence in AI design', *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 62(1), pp. 1637–1641. doi:10.1177/1541931218621371.

Lacey, C. and Caudwell, C. (2019). 'Cuteness as a "dark pattern" in home robots', *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* [Preprint]. doi:10.1109/hri.2019.8673274.

Lex Fridman (2022). *Kate Darling: Social Robots, Ethics, Privacy and the Future of MIT | Lex Fridman Podcast #329*. Available at: https://www.youtube.com/watch?v=ZFntEFXKDHM&t=6855s (Accessed: 20 August 2023).

LIKU (2018). *LIKU Story*. Available at: http://www.likuwith.me/ (Accessed: 12 October 2023).

Lynch, C. R. (2021). Artificial Emotional Intelligence and the Intimate Politics of Robotic Sociality, *Space and Polity*, 25:2, 184-201, DOI: 10.1080/13562576.2021.1985853

Mara, M. and Appel, M., (2015). Effects of lateral head tilt on user perceptions of humanoid and android robots. *Computers in Human Behavior*, 44, pp.326-334.

Matthias, A., (2015). Robot Lies in Health Care: When Is Deception Morally Permissible?. *Kennedy Institute of Ethics Journal*, 25(2), pp. 169-162.

Misty II (2024). *Misty Robotics*. Available at: https://www.mistyrobotics.com/misty-ii (Accessed: 27 January 2024).

Moberg, R. (2023). *Humanoid Robot - Human Interaction: Towards Compliance and Reciprocity with a Social Robot Through Completion of a Pregiving Favor*. thesis.

Mordor Intelligence (2023). *Social Robots Market - Size, Share & Growth*. Available at: https://www.mordorintelligence.com/industry-reports/social-robots-market (Accessed: 08 August 2023).

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

Moussawi, S. and Koufaris, M. (2019) 'Perceived intelligence and perceived anthropomorphism of personal intelligent agents: Scale Development and validation', *Proceedings of the Annual Hawaii International Conference on System Sciences* [Preprint]. doi:10.24251/hicss.2019.015.

Nagel, T. (1974). What is it like to be a bat? *Philosophical Review*, 4:435-50.

Natale, S. (2021). *Deceitful Media. Artificial Intelligence and Social Life After the Turing Test*. Oxford University Press Inc.

Salek Shahrezaie, R. et al. (2022). 'Advancing socially-aware navigation for public spaces', *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)* [Preprint]. doi:10.1109/ro-man53752.2022.9900653.

Schermer, M. (2007). 'Nothing but the truth? On truth and deception in dementia care', *Bioethics*, 21(1), pp. 13–22. doi:10.1111/j.1467-8519.2007.00519.x.

Scott, A.E. et al. (2023). 'Do you mind? user perceptions of machine consciousness', *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* [Preprint]. doi:10.1145/3544548.3581296.

Segers, S., (2022). Robot Technology for the Elderly and the Value of Veracity: Disruptive Technology or Reinvigorating Entrenched Principles?. *Science and Engineering Ethics*

Sharkey, A. and Sharkey, N. (2011). 'Children, the elderly, and interactive robots', *IEEE Robotics & Automation Magazine*, 18(1), pp. 32–38. doi:10.1109/mra.2010.940151.

Sharkey, N., van Wynsberghe, A., Robbins, S., and Hancock, E. (2017). Our sexual future with robots. A foundation for responsible robotics consultation report.

Sharkey, A. and Sharkey, N., (2020). We need to talk about deception in social robotics!. *Ethics and Information Technology*, 23(3), pp.309-316.

Shim, J. and Arkin, R.C. (2013). 'A taxonomy of robot deception and its benefits in HRI', *2013 IEEE International Conference on Systems, Man, and Cybernetics* [Preprint]. doi:10.1109/smc.2013.398.

SoftBank Robotics America, Inc (2023). *Meet Pepper: The Robot built for people* | SoftBank Robotics America. Available at: https://us.softbankrobotics.com/pepper (Accessed: 12 October 2023).

Sparrow, R. (2004). 'The Turing Triage Test', *Ethics and Information Technology*, 6(4), pp. 203–213. doi:10.1007/s10676-004-6491-2.

Sparrow, R. and Sparrow, L. (2006). 'In the hands of machines? the future of aged care', *Minds and Machines*, 16(2), pp. 141–161. doi:10.1007/s11023-006-9030-6.

Sternberg, R.J. (2020). *The Cambridge Handbook of Intelligence*. Cambridge, United Kingdom: Cambridge University Press.

Thomas Burns (2023). *The coolest robot I've ever built!* Available at: https://www.youtube.com/watch?v=bO-DWWFolPw (Accessed: 19 September 2023).

Tiku, N. (2022). *The Google engineer who thinks the company's AI has come to life.* Washington Post. https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/

van Wynsberghe, A., (2021). Social robots and the risks to reciprocity. *AI & SOCIETY*, 37(2), pp.479-485.

Wagner, A.R. and Arkin, R.C. (2010). 'Acting deceptively: Providing robots with the capacity for deception', *International Journal of Social Robotics*, 3(1), pp. 5–26. doi:10.1007/s12369-010-0073-8.

Whaley, B. (1982). *Towards a general theory of deception*. J Strateg Stud 5(1):178–192

Zawieska, K. (2015). Deception and Manipulation in Social Robotics.