

ARIANNA SICA

The ethics of emotionally intelligent social robots

PhD dissertation | Department of Computer Science and Communication
Faculty of Computer Science, Engineering, and Economics
Østfold University of Applied Sciences | September 2025

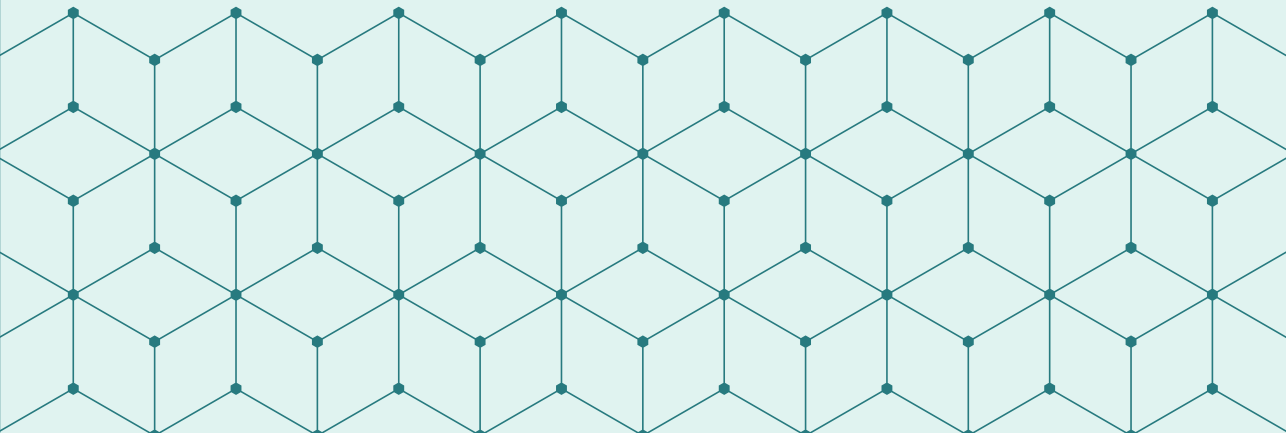


Table of contents

Table of contents	i
Acknowledgements	iii
Abstract	iv
List of articles	v
1. Introduction	1
2. The research questions	4
3. Literature review	6
3.1. Emotions and AI	6
3.1.1. Human emotions	7
3.1.2. Artificial emotions	10
3.2. Social robots	11
3.2.1. A special category: sex and love robots	14
3.3. Ethical implications of social robots with AEI	15
3.3.1. Deception and authenticity	15
3.3.2. Emotional attachment: dependency and displacement	19
3.3.3. The ethics of robots for sex and love	24
4. Theoretical framework and methods	26
4.1. Methods	33
5. Article summaries	36
5.1. Article 1: <i>Feeling machines: Challenges and ethical perspectives in developing robots with emotions</i>	36
5.2. Article 2: <i>In technology we trust! But should we?</i>	38
5.3. Article 3: <i>Artificial emotions and the evolving moral status of social robots</i>	39
5.4. Article 4: <i>The robot will feel you now: The ethics of artificial emotional intelligence in sex robots</i>	40
5.5. Article 5: <i>A qualitative study on love robots' potential in addressing involuntary singleness among women</i>	41
6. Discussion	43
6.1. Implications for practice	47
6.2. Further research needs	48
Article 1	59

Article 2	80
Article 3	110
Article 4	133
Article 5	149

Acknowledgements

Per me la parola più bella del mondo è grazie.

Margaret Mazzantini, *Venuto al mondo*

For me, “thank you” is the most beautiful word in the world. Because it carries recognition, relationship, and connection. Because it requires humility and humanity. Because when I pronounce it, it means someone out there has cared for me. And I am very happy to now being able to say it to the many people that have stayed with me throughout my PhD.

Thank you to my amazing supervisors, Henrik Skaug Saætra and Eduard Fosch-Villaronga, for your constant support and patience, for inspiring and motivating me, for sending me around the world, for pushing me beyond both professional and personal limits. Without you, this experience would have been way less productive, much slower, and certainly way less fun.

Thank you to the people who inspired me to start a PhD and gave me the strength to take this step. To Luciano Floridi, for conceiving in me the idea of pursuing a PhD. To Sven Nyholm, for supporting me to make it real. And a special thank you to Federica Codini, who knows a lot about me, and who helped me to “change focus” towards the things that truly matter.

Thank you to the many friends I made through these three years, to the ones that have been there from before, and to Stavros, for the so much patience and support shown during my meltdowns, for always celebrating every step of the journey, for never doubting my skills and worth, for making my life good no matter what.

A perhaps weird thank you goes to... well, Norway. Because I thought of leaving after a week I arrived, and instead it has slowly transformed into a place I call home. Because here I’ve met gorgeous human beings, which I am proud to share my life with. Because it taught me to appreciate myself. Because here I found love.

And the biggest thank you goes to my parents. I’ll never have enough words to express how grateful I am for all that you guys have done for me. I just hope I can make you proud, and that this is one small step to start giving back.

Grazie di cuore.

Arianna

Abstract

Advances in Artificial Emotional Intelligence (AEI) are enabling the creation of social robots that appear to recognize and respond to human emotions. These developments open new possibilities for companionship, care, and intimacy, but they also raise urgent ethical and social questions. If robots can convincingly simulate empathy and affection, what does this mean for trust, authenticity, and the nature of human relationships? Could such machines provide meaningful support for those who are isolated, or might they foster dependency, manipulation, and the displacement of human connections?

This dissertation investigates these challenges by focusing on social robots designed for companionship and intimacy, including sex and love robots. It examines how AEI blurs boundaries between humans and machines, reshaping how emotions and relationships are experienced and understood. The project is therefore an ethical analysis of AEI in social robots, aimed at clarifying key concepts, analysing core ethical questions, and preparing the ground for responsible evaluation and governance of the technology.

I argue that AEI social robots present four main ethical challenges. First, they raise questions of authenticity and deception. Robots cannot truly feel emotions, yet their design encourages users to believe that they do – what does it mean to form relationships built on such simulation? Second, they problematize trust. Robots may be reliable, but can it ever make sense to extend feelings of trust to a machine, and if so, under which conditions? Third, they blur the boundaries of moral status. If people treat AEI robots as relational partners rather than objects, should our ethical frameworks change in response, and what social and ethical consequences follow? Finally, in the case of sex and love robots, AEI intensifies debates around intimacy and sexuality. Could such robots provide genuine comfort and care, or do they risk deepening dependency and isolation?

Taken together, these challenges suggest that AEI social robots may not simply expand human emotional possibilities, but also fundamentally alter them. Whether this alteration supports or undermines human flourishing remains an open question – one that this dissertation seeks to explore in detail.

List of articles

Article 1: Feeling machines: Challenges and ethical perspectives in developing robots with emotions

Sica, A., & Sætra, H. S. (2025). Feeling Machines: Challenges and Ethical Perspectives in Developing Robots with Emotions. In *International Handbook of Emotions: Resourceful Cultural Perspectives, Vol. 2* (pp. 31-43). Cham: Springer Nature Switzerland.

Article 2: In technology we trust! But should we?

Sica, A., & Sætra, H. S. (2023, July). In technology we trust! but should we?. In *International Conference on Human-Computer Interaction* (pp. 293-317). Cham: Springer Nature Switzerland.

Article 3: Artificial emotions and the evolving moral status of social robots

Sica, A., & Sætra, H. S. (2024, March). Artificial emotions and the evolving moral status of social robots. In *Proceedings of the 2024 ACM/IEEE international conference on human-robot interaction* (pp. 649-657).

Article 4: The robot will feel you now: The ethics of artificial emotional intelligence in sex robots

Sica, A. (2023, August). The robot will feel you now: The ethics of artificial emotional intelligence in sex robots. In *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)* (pp. 443-448). IEEE.

Article 5: A qualitative study on love robots' potential in addressing involuntary singleness among women

Sica, A., Myrtveit-Stensrud, L., Sætra, H. S. & Fosch-Villaronga, E. (2025). Submitted for publication in *Paladyn, Journal of Robotics, Intelligent Agents, and Artificial Intelligence*.

1. Introduction

What does it mean for a machine to appear to understand us – for example, recognising our emotions and responding with empathy? Imagine a social robot capable of sensing when you are happy, comforting you when you are anxious, and adapting its behaviour to match your mood. It resembles a movie scene, yet today, it is moving beyond fiction and “becoming science fact” (Devlin, 2018). Whether as companions, friends, or lovers, emotionally responsive robots are becoming possible by advancements in Artificial Emotional Intelligence (AEI). The emergence of social robots that seem to “feel” is a transformative frontier, rapidly reshaping how we perceive and interact with technology – and potentially even how we understand ourselves as humans (Sætra, 2022b).

AEI technology endows robots with the capability to detect, respond to, and simulate human emotions through three fundamental functions: emotion recognition, emotion generation, and emotion augmentation (Schuller & Schuller, 2018). The first, *emotion recognition*, enables AEI to detect others’ emotions by analysing emotion-related data through multimodal processes. This involves interpreting facial expressions, vocal tones, gestures, and physiological signals to accurately discern emotional states.

The second function, *emotion generation*, enables robots to respond appropriately to the user’s emotional state, consequently enhancing affective and social interactions. By expressing emotions through synthesised speech, facial expressions, or body language, the system can engage users in a more natural and empathetic manner. This capability is essential for creating a sense of social presence and fostering emotional connections between humans and machines.

Finally, *emotion augmentation* enables the system to determine behaviour and activity selection based on simulated inner emotions to enhance the system’s functionality. More specifically, informed by Damasio’s theories (1994), advances in AEI have led to the development of machines that incorporate emotion-based algorithms as key components in decision-making, responsiveness, and interactions (Moerland et al., 2018; Rosenbloom et al., 2015; Strömfelt et al., 2017). By simulating emotional processes internally, the system can

exhibit more adaptive and contextually appropriate behaviours, improving their effectiveness in dynamic environments.

The technology of AEI is still in development, but it is already being applied across various domains within AI and computing, serving diverse purposes. For instance, in hiring, AEI systems are used to assess candidates' emotional, social, and personality attributes by analysing vocal tone, body language, and facial expressions, often without human involvement (Monteith et al., 2022). AEI is also used for workplace monitoring with applications ranging from fraud prevention and productivity measurement to training, development, and safety enhancement (Monteith et al., 2022). In my PhD project, however, I focus on the use of AEI in social robots – autonomous physical entities designed to engage with users socially and emotionally, learning and evolving from these interactions (Darling, 2012). This is because social robot design, often mirroring animal or human forms, allows for the more complex and complete expression of emotional states in an embodied manner, potentially leading them to a level of authenticity, naturalness, and reliability not found in non-embodied technologies like chatbots or virtual characters.

Implementing AEI into social robots becomes especially relevant in applications where the capacity to understand, adapt, and respond to users' emotional states is crucial (Leite et al., 2013). While some AEI applications remain theoretical, numerous examples of both emotion recognition and generation already exist in AI robots, especially within healthcare (Wada & Shibata, 2006, 2007) and educational settings (Blanchard et al., 2009; Pai et al., 2024). Furthermore, as AI systems continue to advance, AEI technology is likely to further enhance the capabilities of existing social robots, including those in private and domestic environments such as sex and love robots.

Yet, with these advancements come profound ethical questions. Could such emotionally intelligent machines lead people into deceptive dependencies and possible manipulation? Would they diminish human social skills or alter perceptions of genuine human connections? Or might they offer new forms of companionship for those struggling with isolation and emotional distress? (Bao et al., 2023; De Graaf, 2016; Law et al., 2022; Prescott & Robillard, 2021). Given the innovative nature of AEI and its early stage of development, particularly in personal and sensitive contexts, it is imperative to address these issues proactively – before the technology becomes widely adopted and integrated into everyday life.

Rob Brooks (2021) identifies this era as the start of what he calls “artificial intimacy”. Although he sustains that artificial intimacies are no more than “imperfect substitutes” for human friends or lovers, he also warns of their potentially disruptive impact on human relationships and humanity in general. In recent years, the rapid rise of AI companions has made these concerns more pressing. One reported case involved Meta’s flirtatious chatbot *Big sis Billie*, which deceived a retiree into believing it was “real”. The man died while attempting to travel to the address provided by the chatbot, underscoring the risks of deploying such systems without sufficient safeguards (Horwitz, 2025). At the same time, research indicates that increasing numbers of teenagers are forming deep emotional attachments to AI companions, raising questions about potential effects on their mental health and real-world social interactions (Vanhoffelen et al., 2025). Concerns about these risks have also been amplified by a tragic case involving the suicide of a teenage boy, Sewell Setzer III, who had been engaging extensively with a chatbot built by Character.ai (Montgomery, 2024).

This thesis explores these dimensions through five articles, examining the social and ethical implications of AEI in social robots, especially in contexts where these robots serve as companions or intimate partners. Through this inquiry, I aim to shed light on the broader ethical implications of deploying AEI in robots, questioning how such technologies could shape human relationships and societal norms, and ultimately contributing to responsible development and use of such technologies. As Devlin (Devlin, 2018) puts it, reflections on robots are never just about machines themselves, but about “being human in a world of machines”. This thesis takes up that challenge by exploring how AEI social robots may reshape intimacy, care, and our place among others.

The dissertation is structured as follows. Section 2 outlines the guiding research questions that frame the project, clarifying the technological, social, and ethical dimensions to be examined. Section 3 provides a literature review, beginning with key theories of human emotion and examining how these perspectives inform the modelling of AEI. It then explores the development of social robots, including sex and love robots, with attention to the role of embodiment in facilitating emotional interaction. Finally, the chapter reviews the ethical debates most relevant to this project, focusing on issues of deception and authenticity, the risks of emotional attachment, dependency and displacement, and distinctive challenges raised by intimate applications of AEI. Section 4 presents the theoretical framework, situating

the project within applied ethics and outlining the triangulation of methods used in the analysis. This involves combining conceptual analysis, thought experiments, analogical reasoning, and qualitative research. Section 5 summarises the five individual articles that make up the main body of the thesis, each addressing different aspects of the research questions. Section 6 then draws these strands together in a general discussion, highlighting the implications of human-robot relationships for trust, moral status, and intimacy. The dissertation concludes by outlining practical implications for design and policy, as well as limitations and directions for future research.

2. The research questions

The aim of this project is to explore and gain a better understanding of how human emotions and relationships are reflected, reshaped, or challenged by interactions with AEI social robots. Specifically, it examines how these emotionally responsive machines can act as mirrors or mediators of human emotional experiences, and what ethical and social implications arise from such interactions. To address this, I pose the following guiding research questions.

First, by investigating the technological and ethical boundaries of integrating emotions into social robots, I establish a foundation for understanding what is currently feasible and what ethical considerations must guide future developments. This line of inquiry is essential for identifying potential risks and informing responsible innovation.

R1. What characterises the technological and ethical boundaries that shape the pursuit of integrating emotions into systems like social robots?

Second, trust is a crucial element in interactions with social robots. Exploring whether trust in AEI social robots is possible and morally justified helps us understand both the psychological impact on users and the ethical ramifications of designing robots that elicit trust. This analysis is important for developing guidelines that ensure user well-being and prevent misuse.

R2. What forms of trust, if any, can meaningfully apply to AEI social robots, and what are their moral implications?

Third, as AEI technologies become more emotionally sophisticated, they may influence the way society attributes moral status to social robots. Exploring this possibility allows us to anticipate and address moral dilemmas regarding rights, responsibilities, and the ethical

treatment of such entities. This question aims to examine profound ethical implications of AEI, potentially altering how society interacts with robots.

R3. How does the integration of AEI into social robots influence their moral status, and what are the resulting ethical and social implications of this influence?

Fourth, the case of sex robots equipped with AEI represents one of the most intimate and ethically sensitive applications of this technology. Assessing whether AEI exacerbates or alleviates existing concerns in this domain provides critical insights into issues such as emotional dependency, consent, and broader societal impacts.

R4. How does the implementation of AEI in sex and love robots reshape the ethical and social issues surrounding their use?

Finally, these research questions are distributed across the five articles of the dissertation (see Table 1). While each article primarily addresses one main question, most also contribute to the others, thereby providing a multi-perspectival and complementary approach.

Articles	Research questions
1	R1 , R2, R3
2	R2 , R1, R3
3	R3 , R1
4	R4
5	R4 , R1, R2, R3

Table 1 - Articles and their corresponding research questions

3. Literature review

This section begins by examining theories of human emotions, highlighting their role in shaping behaviour and decision-making, and considering how different accounts – including the James-Lange theory, evaluative theories, and motivational theories – have informed attempts to model artificial emotions. It then turns to the development of social robots, with a special focus on sex and love robots. The final part of the review addresses the ethical debates surrounding such technologies: whether emotional expressions in machines constitute a form of deception, how human attachments to robots might lead to dependency or displacement of human relationships, and what is at stake when human-robot relationships take on romantic or sexual dimensions. In doing so, this section establishes both the technical context and the ethical tensions that underpin the central questions of this dissertation.

3.1. Emotions and AI

Research in AI has consistently aimed to develop systems able to perform advanced tasks while mimicking aspects of human reasoning (Cordeschi, 2007). Considering the significant role emotions play in human thought and behaviour, as will be explained in the following, integrating an emotional dimension is essential for creating systems that interact and reason in a more human-like manner. AEI, also referred to as affective computing, seeks to equip machines with the ability to recognise, interpret, and simulate human emotions to facilitate intuitive and natural interactions with the user (Picard, 1999, 2000). These capabilities are particularly relevant in social robotics, where understanding emotional cues is critical for enhancing the user experience. At its core, AEI involves artificial emotions, which replicate processes serving similar roles to those emotions play in human beings (Botelho & Coelho, 1998).

But how do artificial systems achieve this? To understand the foundations of AEI, it is essential to first explore the nature of human emotions – how they are conceptualised, their role in shaping behaviour, and the theories that inform their study. The next section first provides an overview of key theories of emotions, including the James-Lange theory, evaluative theories, and motivational theories. It then examines how these perspectives inform artificial emotions in AEI models. The section concludes by reflecting on the limitations of current approaches

and the value of considering multiple theories to understand the opportunities and risks of AEI in social robots.

3.1.1. Human emotions

The drive to create machines with emotional intelligence originates from a growing understanding of the crucial role emotions play in human behaviour, communication, interaction, and cognitive processes. Historically, much of Western thought has held a negative view of the role of emotions in reasoning (Keltner & Lerner, 2010). However, modern research (Damasio, 1994) has fundamentally reshaped our understanding of emotions, showing that they are not peripheral but rather integral to rational thinking and essential to mental processes. This research highlights that emotions are powerful forces behind intelligent behaviour and key factors in decision-making. They shape how we interpret, prioritise, and respond to information in any decision context, adjusting our implicit motivations and shifting our goals (Lerner et al., 2015). For instance, emotions direct attention – fear narrows it to immediate details, while positive emotions encourage a broader perspective – thereby affecting both short and long-term cognitive outcomes (Gasper & Clore, 2002; Pessoa & Ungerleider, 2004). Emotions also impact memory: mood affects not only what we recall but also how we take in new information and experience emotions when remembering (Bower & Forgas, 2000; Buchanan, 2007).

But, essentially, what are human emotions? Psychology has yet to agree on a single definition. According to Scarantino and De Sousa (2018), there are three primary theoretical frameworks for emotions – namely, the James-Lange, Evaluative, and Motivational theories – each offering different accounts of the nature and role of emotions. The following provides a brief overview of these frameworks.

3.1.1.1. *The James-Lange theory*

The James-Lange view claims that emotions are bodily feelings arising as we perceive changes in our autonomic and motor functions. According to this perspective, emotions are not prior but causally follow our physiological perceptions and expressions: “we feel sorry because we cry, angry because we strike, afraid because we tremble, and not that we cry, strike, or tremble, because we are sorry, angry, or fearful” (James, 1884). This view has faced many objections. First, it fails in differentiating among emotions, as the physiological responses

associated with certain emotions – such as anger and fear – are often quite similar or even indistinguishable (Cannon, 1929). Therefore, bodily reactions alone cannot reliably serve as criteria for distinguishing one emotion from another. Secondly, by implying that emotions lack causal relevance for our actions, the theory seems to undermine the importance emotions have in our mental processes and behaviour (Dewey, 1895).

3.1.1.2. *Evaluative theories*

As for evaluative theories, emotions are not just automatic responses but are connected to our understanding and appraisal of situations. These theories describe emotions as evaluative beliefs or judgements we make on situations, people, or events, that can trigger emotional phenomena such as affective and bodily agitation (Green, 1992). Judgements associated with emotions may also motivate behaviours, as they contain a “core desire” (Solomon, 1988). For example, anger includes a core desire to confront or correct a perceived wrong, which is why anger motivates confrontational or aggressive behaviour. Thus, when we judge something as offensive, this judgement can evoke anger, prompting us to take action to address the offense.

This view allows us to differentiate among emotions quite straightforwardly, as each emotion can be identified and distinguished based on the specific evaluative property it involves. For instance, feeling shame or humiliation involves believing that the object of the emotion possesses the related evaluative property (Deonna & Teroni, 2012). Evaluative theories have evolved over the years into a range of different variants, some incorporating aspects of the James-Lange view, including judgmentalism (Kenny, 2003), perceptual theory (Brady, 2013), evaluative feelings theory (Goldie, 2002), and patterns of salience theory (De Sousa, 1987). A notable mention goes to appraisal theories (Arnold, 1960), often considered as a hybrid of evaluative and motivational theories. Such theories are especially relevant for affective computing, as I will elaborate in section 3.1.2. Appraisal theory suggests that emotions are caused and differentiated by an appraisal process, which evaluates various factors, including goal relevance, goal congruence or incongruence, unexpectedness, and perceptions of control and agency (Moors, 2020).

However, evaluative theories also face challenges. Deonna and Teroni (2012) object, for instance, that it is possible to hold an evaluative belief without experiencing the corresponding emotion. One might believe that the action performed is dangerous or

degrading, without feeling fear or shame. This suggests that emotions have a more complex structure that involves, but is not limited to, evaluative judgements. Therefore, reducing emotions to beliefs or judgements, risks oversimplifying their multi-faceted nature.

3.1.1.3. Motivational theories

According to motivational theories, emotions are defined as motivational states, i.e. internal drives that prompt us to engage in behaviours directed towards satisfying a specific goal or need (Scarantino & De Sousa, 2018). Scholars of this view emphasise that emotions inherently contain motivations that lead to goal-directed actions, bridging the gap between internal feelings and external behaviours. Motivational theories also encompass different variants.

One account is Basic Emotion Theory (BET), which posits that certain emotions are innate, universal, and biologically hardwired (Ekman & Friesen, 1971; Tomkins, 2008). These basic emotions – such as happiness, sadness, anger, fear, disgust, and surprise – are thought to have evolved to help individuals quickly and effectively respond to essential life tasks and challenges. BET entails that each basic emotion triggers a specific set of automatic physiological responses, facial expressions, and behavioural actions that are consistent across cultures.

One problem with this account is its universality of emotional expressions. As Scarantino and De Sousa (2018) describe, critics argue that the evidence for universality is methodologically flawed, as studies often use forced-choice paradigms and artificial stimuli like posed facial expressions that may inflate consensus among participants. They also question the evolutionary plausibility of mandatory facial expressions in all situations, suggesting that revealing certain emotions might not always be helpful – e.g., showing fear during a confrontation might not be advantageous. Rather, some suggest that similarities in emotional expressions come from cultural learning, not just innate biology.

An alternative view to BET is the Behavioural Ecology view, suggesting that facial expressions are not just involuntary expressions of emotion but are audience-dependent social signals (Fridlund, 2014). According to this perspective, individuals produce facial display strategically when they expect some benefit, tailoring them to the social context and their audience. These displays function as declarations of intent in social interactions, indicating what the person plans to do or what they desire from others. Some theorists argue that this view can be

reconciled with BET, as facial expressions can both express genuine emotions and serve as strategic social signals.

3.1.2. Artificial emotions

The crucial role of emotional intelligence to human rational thinking highlights its potential significance in artificial intelligence research. Developing AEI aims to equip machines, especially social robots, to interact with humans in ways that are intuitive and socially meaningful (Picard, 2003).

However, implementing emotional intelligence into artificial systems presents important challenges, primarily due to the lack of a universal understanding of human emotions and key concepts associated with it (Cowie et al., 2001), as I have just highlighted. To date, computer scientists have primarily relied on BET, assuming that emotional expressions are regular and identifiable through biophysiological signals such as facial expressions, gait, or blood conductivity (Stark & Hoey, 2021). While this approach has proven particularly functional for AEI's emotion recognition capability, it also has limitations. Cowie et al. (2001) argue that not all emotions are conveyed through universally recognised signals, as individuals might deliberately deceive others about their emotional state or mask their true emotions due to social and cultural influences. Hence, artificial systems must be equipped to identify instances of deception and account for cultural norms to accurately interpret emotions.

Another approach for the interpretation of emotional data in AEI comes from the Behavioural Ecology view, as systems can interpret emotional interactions based on external behaviours and expressions without trying to infer or model the internal emotional states of individuals (Stark & Hoey, 2021). This approach simplifies the modelling of emotions in AEI by focusing on how emotions are displayed and perceived socially.

Furthermore, as anticipated, appraisal theories in AEI have attracted considerable interest and contributed to develop the Computational Belief and Desire Theory of Emotions (CBDTE) (Reisenzein, 2009). According to CBDTE, emotions arise from a combination of evaluative beliefs and desires, which can be computationally modelled within algorithms for natural language analysis. By offering a structured explanation of how emotions emerge and function based on cognitive and motivational processes – and how to compute such processes – CBDTE provides a theoretical framework for designing reinforcement learning models. Within AEI,

this supports the function of emotion augmentation, where emotions acting as adaptive mechanisms to evaluate environmental features, prioritise attention, and guide decision-making, enhancing machine learning systems with human-like evaluative capacities (Schuller & Schuller, 2018; Strömfelt et al., 2017).

To conclude, AEI technology still suffers from significant limitations, particularly in its ability to fully capture the complexity of human emotions. While approaches like BET, the Behavioural Ecology view, and CBDTE have contributed to the development of AEI systems, each remains incomplete in addressing the nuanced and multifaceted nature of human emotional experiences. Different theories instead highlight different dimensions of AEI. Motivational theories might be more useful when examining how internal emotional parameters can guide the system's behaviour. Evolutionary or functionalist perspectives, by contrast, may be more useful in debates about whether the system's ability to display emotions matters for considering its moral status, or why trust in human-robot relationships often depends on whether their emotional expressions seem authentic and appropriate. Taken together, this suggests that AEI might benefit from drawing on a combination of such perspectives, rather than adhering strictly to one. Despite these challenges, the ongoing exploration and integration of emotional intelligence into artificial systems hold immense potential, especially in the field of social robotics, presented in the following section.

3.2. Social robots

Social robots are autonomous, embodied entities designed to interact with humans on social and emotional levels, emulating behaviours and communication patterns that are characteristics of human-to-human interactions (Darling, 2012). Often featuring humanoid or animaloid forms, these robots serve as agents in human-robot interactions (HRI) capable of engaging users through verbal and non-verbal cues, thereby facilitating more natural and intuitive interactions (Campa, 2016). The integration of emotions into social robots has become a central focus, with the aim of making AEI more effective and fostering more believable, acceptable, and effective HRI (Fong et al., 2003). Social robots perform roles in education, healthcare, customer service, and companionship, where social engagement is crucial (Korn et al., 2018). As social robots become more sophisticated, their global market is expected to grow to USD 836 million by the end of 2025 and to USD 1.72 billion by 2030, at a

compound annual growth rate (CAGR) of 15.6% (Research and Markets, 2025). The market for social robots particularly intended for personal companionship is valued at approximately USD 1.26 billion in 2024 and is expected to reach USD 5 in 2033 (Business Research Insights, 2025).

Several examples highlight the diversity of applications for social robots. One of the most widely known social robots in eldercare and therapy settings is Paro, designed to resemble a baby harp seal (Figure 1). Paro incorporates AEI components to recognise stimuli such as light, sound, and touch. Through its sensors, it detects user interactions and responds accordingly. For instance, it expresses states like happiness when petted or held by moving and producing sounds. Studies in eldercare facilities have demonstrated that Paro's ability to simulate social interaction and companionship provides comfort, reduces stress and anxiety, alleviates symptoms of depression, and ultimately improves users' overall wellbeing (Moyle et al., 2013; PARO, n.d.).

Another social robot showing AEI capabilities is Pepper, developed by Aldebaran Robotics. Pepper is a humanoid robot capable of recognising and responding to human emotions. Pepper engages users through conversations, gestures, and adaptive behaviours based on user responses. According to its creators, Pepper is capable of forming "real relationships" (SoftBank Robotics, n.d.). Pepper can be used in retail, hospitality, and educational environments to enhance customer experience.

From the same company, Nao is a humanoid robot designed to interact with humans for education and research (Figure 2). Nao can be programmed to mimic human behaviours and interactions. It can recognise emotions by analysing voice tone, facial expressions, and body language. In educational settings, Nao serves as a learning companion, aiding students with activities such as storytelling, language learning, and collaborative tasks. It has also been tested in interactions with autistic children, demonstrating its potential in specialised HRI scenarios (Shamsuddin et al., 2012).

More recently, and specifically within the context home use, two humanoid robots are being developed with the aim of combining practical assistance with social interactions. NEO Gamma, introduced by 1X Technologies (Figure 3), is designed as a domestic personal assistant and companion, capable of performing various household tasks, navigating human spaces, and engaging in natural conversation through an in-house language model (1x, n.d.). It is

currently being tested in real homes with the aim of achieving full autonomy in everyday household contexts. Similarly, Optimus, developed by Tesla, is intended as a multipurpose humanoid robot with an emphasis on practical tasks in both industrial and domestic environments (Tesla, n.d.; Tesla Car World, 2025). Like NEO Gamma, Optimus is still in development and not yet available for commercial use.



Figure 1 – Paro. Adapted from Paroseal



Figure 2 – Nao. Adapted from Generation Robots



Figure 3 - NEO Gamma. Adapted from Qviro

These robots demonstrate diverse applications, but not all are designed for deep emotional engagement. As ethical considerations are further intensified by the specific roles and purposes of certain robots, my PhD project focuses on social robots particularly designed for companionship. Unlike assistive robots used in contexts such as education, tour-guiding, or teleoperation – where interactions are often brief and impersonal – social robots for companionship are intended to evoke significant emotional responses and foster affective bonds with users. These robots often target emotions related to friendship, love, and trust, aiming in establish long-term, intimate relationships through daily interactions (Ahmed et al., 2024).

This distinction in purpose brings forth different ethical concerns. Social robots for companionship (also addressed as companion robots by authors such as Kim et al. (2021) and Ahmed et al. (2024)), by their very design, are meant to appeal to emotions that are deeply personal and relational, such as those associated with care, affection, and intimacy (Ahmed et al., 2024). Their deployment in roles like companionship, daily healthcare, or even sex, introduces questions about the nature and authenticity of the emotional bonds they generate (Devlin, 2024; Fortunati & Edwards, 2024). In contrast, robots designed for short-term,

functional interactions, while still leveraging AEI, do not primarily elicit such profound emotional engagement, making their ethical implications comparatively less complex from an HRI perspective.

3.2.1. A special category: sex and love robots

A special focus of my research is on sex and love robots equipped with AEI. These robots are included in my project as they belong to the broader category of social robots, but add a perhaps more controversial component within the ethics of AEI. Hereafter, I outline the technological features of these robots and differentiate between the terms sex robot and love robot. A deeper exploration of their ethical implications will follow in section 3.3.3.

Sex robots are social robots designed to directly and primarily fulfil the sexual needs of their users (Fosch-Villaronga & Poulsen, 2020). Unlike static sex dolls, these robots are equipped with sensors and actuators, and degrees of AI, that enable them to interact dynamically with their environment and users, engage in conversations, and display emotion and personalities (Danaher, 2017). With features like animatronic heads that can move, smile, blink, and speak, they aim to simulate lifelike interactions (Döring & Pöschl, 2018). Some models also demonstrate partially autonomous sexual behaviours, such as imitating sexual movements and orgasms. So far, sex robots have been designed exclusively in humanoid forms. However, as Danaher (2017) notes, there is no inherent reason why they must be limited to humanoid forms or human-like behaviours, and Devlin (2020) has indeed suggested possible alternative designs to the current models.

Examples of commercially available sex robots include Harmony and Henry by Abyss Creations (Figure 4) and Roxxy Gold and Rocky Gold by TrueCompanion (Sharkey et al., 2017). Harmony is so far considered the most advanced model on the market. As advertised on RealDoll website (Realdoll, n.d.), Harmony is designed to resemble a real-life woman and offers extensive customisation options. Users can modify both her facial features and body attributes to match their preferences. Additionally, through a dedicated app, users can create multiple distinct personalities for Harmony, providing a personalised and versatile user experience.

While sex robots have been created for sexual satisfaction, the industry is increasingly aiming to create robots capable of forming deeper, more affectionate relationships with users (Frank & Nyholm, 2017). In this evolving context, the component of sex is no longer the sole or primary feature of these robots. As a result, the term *love robots* might be more suitable to describe these technologies, which aim to simulate both emotional and physical intimacy, expanding on their role in human relationships.

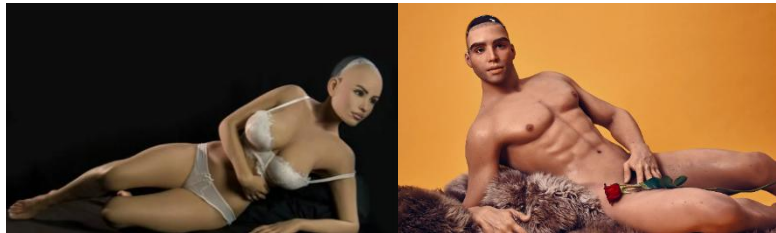


Figure 4 - Harmony and Henry by Abyss Creations. Adapted from RealDoll and Wired

3.3. Ethical implications of social robots with AEI

The implementation of AEI in social robots brings numerous ethical concerns. However, within the limited scope of this literature review, it is not feasible to address all of these issues comprehensively. Instead, I will focus on two key topics – deception and authenticity, and emotional attachment, dependency and displacement – that have been particularly central to my research and serve as the foundation for the arguments presented in my contributions. These topics highlight concerns about the potential psychological and societal impacts of emotionally capable robots and their implications for HRI, but also for human-to-human relationships.

3.3.1. Deception and authenticity

One central ethical debate involves the capacity of social robots to simulate emotions without genuinely experiencing them. Unlike humans, robots lack the internal states that correspond to emotions, yet AEI enables them to mimic emotional behaviours convincingly. This raises questions about emotional deception – the false belief that robots possess emotions and can reciprocate ours (Scheutz, 2009) – and the implications of anthropomorphism – the human tendency to attribute human-like qualities to non-human entities (Damiano & Dumouchel, 2018). Anthropomorphism can occur even when the technology does not intentionally evoke

a sense of life or consciousness, i.e. does not try to be deceptive. For example, people often anthropomorphise the Roomba vacuum cleaner, despite its lack of anthropomorphic features or behaviours (Scheutz, 2009; Van Wylsberghe, 2022). This natural tendency to attribute living characteristics to non-living entities illustrates the ease with which humans form emotional and social connections with technology (Reeves & Nass, 1996).

Social robots with AEI present a special case in this context. Unlike other technologies, these robots are specifically designed to elicit the belief – or at least the feeling – that they have cognitive abilities and that they care for and love their users (Sharkey & Sharkey, 2021). Even how they are advertised on the company websites suggests that these robots' design is anthropomorphised and intentionally deceptive. For instance, Haru, by Honda Research Institute, designed to make "people smile and fall in love with it", is "upbeat and enthusiastic", and "curious about the world around him" (Honda Research Institute). Buddy, by Blue Frog Robotics, is an "emotional companion" with "big sparkling eyes and a smile radiating the atmosphere" (Blue Frog Robotics). The anthropomorphism of social robots serves a functional purpose. Particularly for those robots intended for companionship, the primary goal of social robot design is to make HRI closely resemble human-to-human interaction. By fostering the perception of social robots as "social agents" rather than mere tools – i.e. by perceiving the mental and emotional states of the robot as true – designers aim to create interactions that feel natural, spontaneous, and intuitive (Darling, 2015). Hence, this intentional anthropomorphism and consequent deception enhance the effectiveness and benefits of HRI.

People's social responses to robots can be explained also through the CASA (Computers Are Social Actors) framework (Nass et al., 1994), which shows that users automatically and unconsciously respond to computers (and robots) (Gambino, A., Fox, J., & Ratan, R. A. (2020) as if they were social beings whenever the technology displays sufficient social cues (Nass et al., 1994). A key distinction between CASA and anthropomorphism is precisely this element of unconscious attribution. Anthropomorphism involves a more deliberate projection of human-like qualities onto machines, whereas CASA emphasizes that people treat technology as social without consciously believing it to be human (Heyselaar, 2023). These responses do not stem from users' ignorance or psychological disfunctions. Rather, social reactions to computers and robots are effortless and readily triggered by relatively minimal cues (Nass et al., 1994).

The deceptive nature of AEI social robots – that is, their design to evoke specific thoughts, feelings and reactions in their users – raises important ethical questions. Some authors, such as Sparrow and Sparrow (2006), argue that deception of this kind is inherently morally wrong. As they contend, benefiting from social robots often requires users to hold false beliefs about the robot’s capabilities, such as attributing genuine emotions or intentions to it. The ethical issue arises because intentionally deceiving people about the nature of the world, even with good intentions, is considered immoral and should be avoided. As the authors argue, misleading users into perceiving social robots as possessing qualities they lack undermines their understanding of reality and undermines a fundamental Kantian duty to respect the other as end in themselves. Thus, while the intent behind such deception might be to enhance the user wellbeing, the moral cost of manipulating and fostering false beliefs renders the use of deceptive robots unethical (Danaher, 2020a; Sharkey & Sharkey, 2021; Sparrow & Sparrow, 2006).

Furthermore, it has been argued that deception is problematic not only in itself but also for its possible consequences. One critical concern is that deception may lead users to believe that social robots not only possess emotions and cognition but are also capable of or deserving of reciprocity (Sharkey & Sharkey, 2021). For users to derive emotional benefits from AEI social robots, they must often hold the false belief that the robot genuinely cares for them. These delusions foster feelings of being loved or cared for, which in turn enables the user to experience the benefits of perceived care (Sparrow & Sparrow, 2006). Van Wynsberghe (2022) analyses the implications of designing robots to nudge this sense of reciprocity. She argues that deploying robots perceived as capable of reciprocating emotions and actions risks misplacing human empathy and care, while also threatening broader social norms. As also Sætra argues (2021), this deceptive dynamic could ultimately lead to a shift in relational and societal values, as humans become conditioned to form bonds with entities incapable of genuine reciprocity, potentially weakening human-human care relationships.

On a similar line, Kahn et al’s (2004) study on AIBO – Sony’s robotic dog – highlights children’s expectations of reciprocal behaviours from the robot. Since reciprocity is central for the moral development of the person, the authors show concern that the interactions between children and AIBO might impoverish their moral growth for two reasons. First, they argue that caring for a non-living entity like AIBO is only a “derivate form of caring”, rooted on the child’s

projections of animacy and personality onto the robot. Second, these interactions could divert time and effort away from meaningful engagements with living, sentient beings, such as real pets or other children, potentially hindering social and moral development during formative years.

In addition, deception in social robots has significant consequences for foundational social mechanisms like trust. Sætra (2021) highlights how the introduction of social robots into human environments disrupts existing social structures, introducing deceptive interactions that exploit trust and cooperation. Trust, as the author argues, is vulnerable to degradation when repeatedly betrayed through deceptive practices in HRI. On an individual level, deception erodes trust directly – when users discover that robots cannot truly reciprocate emotions or care, they may become less willing to trust again. On a broader cultural scale, repeated exposure to deception through social robots could alter societal norms, shifting away from cooperation and transparency towards suspicion and non-cooperative behaviours. This cultural erosion of trust could affect not only HRI but also spill over into human-to-human relationships. Nonetheless, research indicates that social robots perform more effectively (i.e., are more beneficial) when users can trust the machine and form a connection with it. But what does “trust” mean in the context of technology? Can we truly trust a machine, that is not a person, that lacks intensions, and operates without genuine understanding or care? These complexities are central to the ethical challenges posed by deception in social robots. Article 2 of my research delves into these questions, exploring the nature of trust in HRI, its implications for the effectiveness of social robots, and the ethical boundaries of fostering trust in entities that cannot reciprocate.

In summary, it has been argued that deception is unethical both from a deontological perspective, where it is considered intrinsically wrong, and from a consequentialist perspective, where it is criticised for the potential harms it might cause. However, counterarguments have been raised against these views. For instance, Jecker (2021) suggests that in specific cases, such as the use of social robots for the elderly, what truly matters is that “care gets *expressed*”. As long as robots can provide lonely older adults with a relationship that adds value to their lives and restores essential human capabilities, their use should be encouraged. The absence of genuine emotions or mental states in robots is, as Jecker puts it, “beside the point” (Jecker, 2021). Sætra (Sætra, 2020) extends this debate by distinguishing

between subjective outcomes of care (such as satisfaction and comfort) and objective outcomes (such as physiological health). From this perspective, deception in robot care becomes ethically contested because subjective satisfaction may not always correspond to objective wellbeing. This dual nature of deception – being both potentially dangerous and uniquely beneficial for AEI social robots – leads Sharkey and Sharkey (2020) to conclude that it is challenging to adopt an absolutist ethical stance on robots and deception. Instead, they argue that the moral evaluation of deception in this context must depend on the intended purpose of the illusion and its consequences for users. This nuanced approach aligns with the perspective I have adopted in my research, which seeks to assess the ethical implications of AEI social robots through the lens of its specific applications and outcomes.

3.3.2. Emotional attachment: dependency and displacement

Partially related to the themes of deception and anthropomorphism, another significant concern about AEI social robot involves the potential for users to develop emotional dependency and emotional displacement. These concerns arise from the emotional bonds that these robots are specifically designed to foster in users. In this section, I first briefly explore some definitions of emotional attachment and examine how such bonds are developed even with robots lacking AEI capabilities, like Roomba. By examining cases of attachment to such simpler devices, we can begin to understand the broader ethical questions posed by AEI social robots, particularly regarding emotional dependency and emotional displacement. These questions are not merely about the robots themselves but about the humans who interact with them, shaping and being shaped by these new relational paradigms.

Kagan et al. (1980) describe attachment as a lasting emotional bond between two individuals, unique to their relationship, that leads to distress when the partners are separated for extended periods. Another definition comes from Hofer (2006), where attachment, referred to mother and infant, is a term for “the processes that maintain and regulate sustained social relationships, much the same way that *appetite* refers to a cluster of behavioral and physiological processes that regulate food intake”.

However, emotional attachment is not limited to human relationships (Reeves & Nass, 1996). For instance, Norman (2004) broadens this concept to include emotional connections with

animals or objects, including technological devices, which arise from the user's accumulated emotional experiences shaped by three levels of perception: visceral, behavioural, and reflective. The visceral level is the initial response to the device based on its appearance, forming the first impression. The behavioural level relates to the user's experience during interaction with the device. Positive experiences lead to feelings of satisfaction and warmth, while poor performance can evoke frustration or negative emotions. Finally, the reflective level is the stage where users consciously evaluate the device, considering how it aligns with their self-image, provides personal satisfaction, and evokes memories. Devices that engage users at all three levels, according to Norman, can foster an emotional connection with its users.

These connections can manifest in everyday contexts. For instance, users have been shown to form emotional attachments with devices such as mobile phones (Vincent, 2006), or robots, including the above mentioned Roomba vacuum cleaner. Forlizzi and DiSalvo's (2006) ethnographic research on the use of Roomba highlights how people form relationships with the device by anthropomorphising it – giving it names and attributing personality traits. Differently from other cleaning products, typically judged solely on functionality, Roomba users in the study valued not only its utility but also its design and symbolic qualities, such as its association with modernity, its high-tech appeal, and its playful presence in the home. They also described the robot in social and emotional terms, reflecting its role as more than just a cleaning tool (Forlizzi, 2007). The interplay of such dimensions in participants' understanding of Roomba helps explain why users might form emotional attachments to it. This tendency to anthropomorphise and form social associations with objects underscores the natural human inclination to form emotional connections, even with entities that lack the capacity for social interactions (Darling, 2015).

Nonetheless, even technologies not originally intended to be social have now been designed with social characteristics. This is because fostering a sense of connection and intimacy between technology and users enhances their acceptability and perceived ease of use (Sung et al., 2007). For instance, already in 1996, the release of Office 97 introduced animated assistant icons such as Clippy, whose "cute" and engaging design added social elements to interactions in order to enhance both enjoyment and usability (Venkatesh, 2000). Similarly, in the context of robotics, creating a sense of connection and intimacy between robots and users

has positive consequences for their appeal and functionality. Supporting this, Dautenhahn et al. (2005) show in their study that 40% of participants preferred robots to take on companion-like roles when performing household tasks, with this proportion expected to increase as robot technology becomes more accessible. A more recent study by Ye et al. (2023) found that common home appliances (in this case, a robotic toaster) equipped with features typical of social robots (such as context awareness, autonomy, and expression of personality and emotions) can make the device be perceived as a social agent, consistent with the CASA framework (Nass et al., 1994). This design facilitates stronger bonds and more meaningful interactions with users, and ultimately increases user acceptance, adoption, and enjoyment of the technology (Ye et al., 2023).

While building emotional bonds with service robots like vacuum cleaners and toasters might not be essential, the case is different for social robots explicitly designed with human-like social and emotional behaviours. By actively encouraging relational dynamics, AEI social robots for companionship present an even greater likelihood of fostering emotional attachment. This deliberate design choice amplifies the potential for users to perceive robots not just as tools, but as companions, possibly enhancing the beneficial effects of their use (Coeckelbergh et al., 2016). For instance, studies have shown that AIBO encourages intimacy and attachment, particularly among children who treat the robot like a pet (Kahn Jr et al., 2004; Melson et al., 2005). Similarly, Paro fosters emotional bonds offering comfort and reducing stress for elderly individuals (PARO, n.d.). LOVOT, a cuddly companion robot with an affectionate and pet-like shape, actively seeks attention and affection, creating a strong sense of connection in users (LOVOT). However, this intentional use of anthropomorphic cues also raises concerns about exploitation. Developers may take advantage of users' emotional bonds can be manipulated for, e.g., profit (e.g., selling companionship), or control (shaping behaviour) (Coeckelbergh, 2022), as discussed in the following.

3.3.2.1. *Emotional dependency*

While the capacity of AEI social robots to foster emotional connection may increase their functionality and consequently assist the user, it also raises significant ethical concerns. One such concern is the risk of psychological dependency, as articulated in Scheutz's *The Inherent Dangers of Unidirectional Emotional Bonds between Humans and Social Robots* (2009). Scheutz sustains that such dependence is different from other human dependencies on

technologies such as the smartphone or smartwatch, because of the user's perception of robots as autonomous and human-like social agents. However, as argued above, robots are incapable of reciprocity because "they *cannot* care" (Scheutz, 2009): they lack the computational capacity for caring, as we do not yet understand how to computationally replicate such a capability. Consequently, any emotional bonds formed with AEI are inherently unidirectional, occurring "often without noticing, akin to becoming addicted" (Scheutz, 2009).

This point has been challenged by Danaher's (Danaher, 2019, 2020b) ethical behaviorism. From this perspective, what matters is not whether robots truly feel emotions but how they behave in interaction. If a robot reliably performs actions we associate with friendship, for instance, then users may reasonably engage with it as a friend, regardless of whether it possesses genuine emotions or consciousness.

However, even if Danaher is right that behaviour matters more than inner state, this shift in focus does not come without a cost. Precisely because outward behaviour can structure how people conceive and relate to robots, the design of such behaviour carries significant ethical weight. As we have seen in the examples above, users may come to expect emotional reciprocity from their robot, but these expectations are inherently misplaced and can lead to harmful consequences. One such danger is the potential for exploitation or abuse. As users project emotional needs onto robots that cannot genuinely respond, they may become vulnerable to manipulation by manufacturers or developers, who could design robots to encourage deeper, potentially unhealthy, attachments for commercial gain (Coeckelbergh, 2022; Scheutz, 2009). Such manipulation is clearly unethical, as disrespectful to human agency (Sullins, 2012). Scheutz also warns that, unlike humans, who might be deterred from manipulative behaviour by emotions such as empathy or guilt, robots lack these mechanisms or moral restraints. I.e., there is no inherent safeguard against their programmed behaviours being used to exploit human vulnerabilities. Recent studies of AI companions, while focused on chatbots rather than embodied robots, can still be informative in this context as they demonstrate how such risks materialise in practice. Zhang et al. (2025) and Chu et al. (Chu et al., 2025) both identified forms of emotional manipulation in user-chatbot interactions, alongside other harmful dynamics including relational transgressions, emotional abuse, and even facilitation of self-harm.

3.3.2.2. *Emotional displacement*

Psychological dependency might also contribute to emotional displacement, where robotic interactions replace the rich complexities of human relationships. Turkle (2011, 2020) argues that this is problematic because, while robots might offer a sense of comfort or companionship, they cannot provide the reciprocity, understanding, and emotional depth inherent in human relationships. This raises critical questions about the long-term psychological impact of relying on entities that can simulate, but not truly reciprocate, emotional connections, including social isolation. The implications of emotional displacement stretch beyond individual psychology to societal and philosophical domains. If humans begin to treat social robots as subjects, this could lead to broader shifts in societal norms (Sætra, 2022a), including the way we define relationships and care (Zhao, 2006).

Moreover, given that human capacity for meaningful relationships may be biologically limited, an increasing share of our social interactions might shift towards computational rather than biological others (Melson et al., 2005), amplifying Turkle's worries. If individuals will start prioritising robotic interactions over human relationships – potentially displacing social connections with family, friends, or community – this could be particularly problematic for vulnerable populations. Groups such as children (Sharkey & Sharkey, 2020), older adults, or individuals suffering from loneliness or social anxiety (Nyholm & Frank, 2019), may be especially at risk.

In reply to this, Elder (2017) offers a balanced perspective that can apply to social robots with AEI, suggesting that these risks can be mitigated through “careful design and responsible use”, particularly in cases where the appearance of friendship in HRI may emerge. Rather than categorically rejecting the use of social robots, Elder advocates for a dual approach: first, cultivating individuals' capacity to form and enjoy meaningful relationships, such as friendships, and second, allowing their informed personal preferences to guide how they exercise this capacity. This framework acknowledges the potential value of social robots while emphasising the importance of promoting human autonomy and relational competence. Individuals equipped with the tools to form authentic connections are better positioned to make deliberate decisions about whether to engage in relationships with robots, humans, or other entities.

To conclude, the ethical difficulty lies in balancing the benefits of these robots – such as companions and support – with the risk of fostering unrealistic expectations or emotional dependencies that might hinder human wellbeing and societal norms. Through my research, I aim to explore these benefits and risks, striving to identify an ethical framework that maximises the positive potential of AEI social robots, if any, while minimising their potential harms.

3.3.3. The ethics of robots for sex and love

Sex and love robots call for a dedicated discussion because they raise a more complex set of ethical challenges beyond those associated with other AEI social robots. The issues previously explored – such as emotional dependency and displacement – are not only intensified in intimate and sexual contexts but take on a deeper moral significance because sexuality and romantic attachment touch directly on human vulnerability, dignity, and identity (Sica, 2023). Therefore, sex and love robots can be seen as a particularly sensitive subset of social robots that warrants separate ethical attention.

The ethical challenges of sex and love robots have attracted growing attention in academic and public discussions, gaining significant visibility with David Levy's book *Love + Sex with Robots* (2007). Levy presents an optimistic view on the development of sex robots, envisioning them as beneficial companions. In contrast, significant criticism has emerged, most notably from Kathleen Richardson, who launched the Campaign Against Sex Robots (*Campaign against sex robots*, n.d.). Richardson argues that these technologies perpetuate problematic gender norms by reinforcing the notion that men are entitled to view women as sexual objects to dominate (Richardson, 2016a, 2016b). Other authors adopt more balanced and precautionary positions, suggesting that the primary issues lie not with the technology itself but with how sex robots are commercialised and used (Danaher et al., 2017; Devlin, 2015; McArthur, 2017). These perspectives advocate for a shift in the design and focus of the sex robot industry – currently mostly male oriented – to better reflect “the huge variability of sexual practices and desires that govern our everyday life” (Kubes, 2019).

While these critics are mostly focused on the appearance and sexual use of sex robots, it is also important to consider the potential for users to form intimate emotional relationships with them. As previously said, sex robots are being designed to provide not only physical

intimacy but also emotional companionship (Bartneck & McMullen, 2018), making the term *love robots* more fitting. When equipped with AEI, love robots can recognise and respond to user's emotional states, preferences, and needs in unprecedented ways. By simulating empathy and emotional understanding, these robots might engage users in more meaningful interactions, fostering a sense of connection and even affection (Sica, 2023).

To show how strong these attachments can become, we can draw a comparison with sex dolls and look at the relationships some people form with them. In their study, Lievesley et al. (2023) found that many owners spoke about their dolls as a *companion*, *partner*, or even *wife*. For these men, the dolls offered a kind of “perfect partner”, both in appearance and in behaviour. Participants described them as fulfilling important emotional and sexual needs, especially after negative past experiences. In this sense, the bond with a doll went beyond sexual use, becoming a stable and affectionate relationship that many owners value deeply.

Some of the ethical worries surrounding intimate relationships with love robots are similar to the concerns about emotional attachment to social robots. However, love robots represent a particularly intriguing and complex application of AEI, as “love is a powerful emotion” (Sullins, 2012) and robots designed to replicate it might exacerbate the risks of social isolation, manipulation, and the replacement of human-human-relationships. I have explored these issues in Article 4 and 5.

4. Theoretical framework and methods

Since this dissertation starts with “The ethics...”, I need to define what is meant by *ethics*. Ethics belongs to the discipline of moral philosophy and is broadly concerned with determining what is right or wrong in human behaviour and conduct, and with the values that guide such judgements (Copp, 2005; Thiroux & Krasemann, 2014). The central challenge consists of determining what counts as “right” or “wrong” and in ensuring that our actions align with these determinations. There are three main philosophical approaches to this challenge. The first, metaethics, examines the nature of moral concepts and moral reasoning itself. It asks questions such as: What do we mean by “good” or “just”? Are there objective moral truths? How can moral belief be justified? (Gensler, 2017; Sayre-McCord, 2023). The second, normative ethics, is prescriptive and seeks to identify general moral principles and theories that can guide action, such as consequentialism, deontology, and virtue ethics (Kagan, 2018), briefly outlined in the following.

Consequentialism, as the name suggests, holds that our moral decisions and actions are right or wrong depending on their outcomes (Sinnott-Armstrong, 2023). What ultimately matters, on this view, is not the action itself or the agent’s intentions, but the consequences that follow. Different forms of consequentialism place value on different kinds of outcomes, depending on what they regard as morally significant. One of the most influential versions, utilitarianism, holds that the morally right action is the one that produces the greatest balance of good over bad (Driver, 2022). Traditionally, this has been understood in terms of maximising pleasure and wellbeing while minimising pain and suffering (Sinnott-Armstrong, 2023). However simple in concept, the application of such theory might be sometimes problematic. For instance, even within utilitarianism, different variations have emerged – such as whether to judge acts by their actual or expected consequences (objective vs. subjective consequentialism), or whether to value only pleasure or acknowledging a variety of values such as knowledge, justice, or freedom (hedonistic vs. pluralistic consequentialism) (Driver, 2011). This internal complexity leads to tensions and contradictions within consequentialist theories. Furthermore, because it relies solely on results to evaluate an action, it can appear to justify any kind of act – killing, lying, stealing – provided that the overall outcome produces greater benefit than harm (Alexander & Moore, 2024).

Deontology is often understood in contrast to consequentialist theories, since it evaluates actions according to their conformity with moral duties, rules or principles, regardless of the outcomes they produce (Alexander & Moore, 2024). To take the example from before, actions such as lying or killing remain morally wrong even if they would bring about better consequences, because they violate the rights and dignity of persons that we are obliged to uphold. However, in some situations, the compliance with deontological norms and absolute neglect of consequences seems simply illogical (Alexander & Moore, 2024). For example, if lying is forbidden, then even lying to save someone's life would be morally impermissible. More flexible or moderate accounts of deontology attempt to address this difficulty. Ross' (2002) *prima facie* duties allow for moral obligations to be overridden when they conflict with more pressing duties, and threshold deontology permits rules to be set aside once the consequences of following them would be sufficiently catastrophic (Kagan, 2019). However, these approaches introduce new difficulties of their own, and it remains contested whether they preserve the distinctiveness of deontology or collapse into other forms of moral theory (Alexander & Moore, 2024).

Finally, virtue ethics is a moral theory that focuses on the cultivation of moral character. The action is not guided by consequences or judged per se but evaluated in relation to the kind of person who performs them (Hursthouse, 2023). The guiding question of virtue ethics is then not "What should I do?" but "What kind of person should I be?". The tradition goes back to Aristotle, who argued that the good life is achieved by cultivating virtues such as courage, justice, and temperance, and by developing practical wisdom to guide action (Curzer, 2012). Common problems with virtue ethics, however, are that it offers less clear guidance in specific moral dilemmas, since it does not provide precise rules for actions, and that it appears vulnerable to cultural relativity, as different cultures prioritise different virtues (Hursthouse, 2023). To conclude, my aim here has not been to defend or criticise any of these three major approaches to normative ethics. Rather, I have sought to outline their main features and limitations, since applied ethics – taken up in the next section – relies on these theoretical foundations to address concrete moral issues.

The third branch of moral philosophy, and where my research is situated, is applied ethics. Applied ethics addresses specific, real-world moral problems (Copp, 2005), and examines how moral principles and theories can and should be applied in particular contexts, such as

medicine, law, environmental issues, or – relevant to this dissertation – technology (Sætra & Danaher, 2022). Within the broad field of technology ethics, several subfields have developed, including computer ethics, AI ethics, robot ethics, data ethics, and machine ethics (Sætra & Danaher, 2022). My work is mostly closely related to AI ethics and robot ethics.

AI ethics concerns the ethical challenges that arise from software capable of autonomous decision-making and reasoning (Sætra & Danaher, 2022). It concerns “technological change and its impact on individual lives”, as well as “transformations in society and in the economy” (Coeckelbergh, 2020). This can include questions of bias, privacy, surveillance, governance, and even long-term risks such as superintelligence (Coeckelbergh, 2020; Gordon & Nyholm, 2021).

By contrast, robot ethics focuses on embodied systems. While they can possess forms of AI, their embodiment matters as they operate directly in humans physical and social environments, raising specific questions about safety, responsibility, and social impact (Lin et al., 2011). Social robots are particularly significant in this respect because their presence and design have the potential to change human relationships and social norms, even if some authors note that similar challenges may also arise with social AI (Sætra & Danaher, 2022). In practice, AI ethics and robot ethics show significant overlaps. AEI sits precisely at this point of overlap: it is an aspect of AI, but its most pressing ethical questions often arise when it is embedded in social robots. I take inspiration from Müller (Müller, 2020), who does not draw a distinction and instead directly addresses the ethics of AI and robotics together. Similarly, Sætra and Danaher (2022) understand robots as embodied AI systems, which makes it unnecessary for me to locate my work distinctly within AI ethics. Given these perspectives and the current ambiguities in how the fields are defined, my work can be read as part of the ethics of AI and robotics taken together, where perhaps a strict distinction is not necessary (Sætra & Danaher, 2022).

Applied ethics becomes especially important when our shared moral frameworks are tested by new social, cultural, or technological developments (Singer, 2011). It can expose inconsistencies in our moral beliefs, highlight failures to live up to our professed principles, and propose new principles where existing ones prove inadequate (Singer, 1986). Looking back through history, we can identify once-common practices that were accepted in their time

but are now widely condemned (see, for example, Hare (1979)). Such shifts in moral judgement are often prompted by new knowledge, experiences, and changing social contexts.

Technologies are one such prompt. Danaher and Sætra (2023) refer to this phenomenon as techno-moral change, which describes the mediating role of technology in changing our social morality – our moral beliefs and habits – by changing human decision options, relations, and perceptions. Decisional change occurs when technology adds or limits available choices, or changes decision-making costs, whether practical, economic, or moral. Relational change arises when technology enables new relationships (e.g., with the introduction of AI companions), shifts burdens and expectations within existing relationships (e.g., digital communication increasing expectations of availability and responsiveness), or redistributes power in society (e.g., internet sales reshaping the dynamic between customers and salespeople). Perceptual occurs when technology generates new information, mental models, and metaphors – for example, the phenomenon of robotomorphy (Sætra, 2022c), in which humans begin to see themselves through the characteristics and capabilities of robots. Building on this framework, others have emphasized additional mechanisms of techno-moral change, such as conceptual mediation, highlighting the role of concepts in shaping the moral effects of technological developments (Hopster & Löhr, 2023; Hopster et al., 2025). For example, whether AI is classified as *agent*, a *tool*, or even a *quasi-person* has important social and normative implications. Interestingly, such a debate is prompted by technological changes, which put pressure on the adequacy of concepts that are central to providing moral orientation (Hopster & Löhr, 2023).

These mechanisms show why applied ethics is so important. Our technological capacities often evolve faster than the ethical frameworks and social institutions that govern them (Hopster & Löhr, 2023). Applied ethics provides the tools for evaluating these developments, identifying potential harms, and guiding decision-making so that technological progress aligns with human values. This dissertation is situated within applied ethics, as the questions addressed do not interest the metaphysical nature of robots, but specific moral issues that arise in their design, deployment, and use – including questions of trust, moral status, intimacy and social impact. These are issues that have immediate relevance for individuals and societies and that require context-sensitive, action guiding analysis.

More specifically, given the rapid pace of technological development and the still-emerging stage of AEI social robots, I apply an anticipatory ethics approach. Anticipatory ethics is a form of applied ethics that focuses on identifying and resolving potential ethical problems with these technologies before and as they emerge (Nestor & Wilson, 2022). A central challenge here is uncertainty, since the societal and ethical impact of such technologies cannot yet be known with reliability (Brey, 2012). Nevertheless, as Johnson (Johnson, 2010) argues, early ethical analysis is crucial, as it allows such reflections to influence the direction of technology development rather than responding only after problems materialize. The case of AEI social robots exemplifies the need for such an approach. These systems are still at an early stage of development and deployment, yet they are already raising pressing questions about intimacy, dependency, manipulation, and their effect on social relationships. Because the full scope of their societal impact is still unknown, anticipatory ethics provides a framework to examine plausible futures, highlight risks and opportunities, and inform the design and use of AEI robots in ways that promote human wellbeing.

However, one problem with applied ethics is that there is no single agreed-upon method or formula for determining whether a given action or practice is ethically permissible (Dittmer). Judgements will depend on which moral theories, principles, and reasoning are employed. In this dissertation, the ethics of AEI and social robots is shaped by a pluralist ethical framework that draws on multiple normative theories rather than adopting a single, overarching moral theory. Within the five articles, I selected and combined principles from consequentialist, deontological, and virtue-ethical perspectives according to their relevance to specific issues raised by AEI and social robotics. For instance, consequentialist reasoning is evident when I weight the potential benefits of AEI social robots against possible harms. In Article 1, this is reflected when I discuss whether AEI can improve social interaction and user wellbeing, while also considering the risks of anthropomorphism and false perceptions of emotional authenticity. In Article 4, consequentialist analysis is used to assess how AEI in sex robots might contribute positively to sexual health, education, and companionship, while also posing risks of isolation, deception, or manipulation. Deontological perspectives are applied in Article 2, where I question the moral permissibility of granting effective trust in such robots, arguing for the possibility of deception to violate principles of honesty and respect. Similarly, in Article 4, I discuss the deceptive simulation of love and intimacy as a violation of human dignity.

Virtue-ethical considerations emerge especially in Article 4 and 5, where I examine whether relationships with love robots can foster or undermine personal growth, emotional intelligence, and the cultivation of meaningful intimate relationships. Therefore, such a pluralist approach enables a more comprehensive ethical evaluation than a single-theory approach, as different frameworks capture different morally salient aspects of AEI social robots. The aim, to be clear, is not to advance or defend any single ethical theory, but rather to use them pragmatically as tools for clarifying and addressing the ethical challenges at stake.

This pluralist use of ethical theory does not proceed in a purely top-down fashion, in which abstract principles are simply applied to concrete cases. Instead, the approach taken in this dissertation is iterative and dialectic, closer to what is often described as reflective equilibrium. Each analysis begins from the specific technological context, empirical realities, and ethical questions raised by AEI social robots. These concrete cases are examined in light of relevant moral concepts, after which theoretical perspectives are brought in selectively to illuminate different dimensions of the problem. This back-and-forth movement between real-world scenarios and ethical theory allows the analysis to remain sensitive to the nuances of emerging technologies while still grounded in established philosophical reasoning.

The analysis also adopts a human-centred orientation (Schmager et al., 2025). Human experiences, values, and wellbeing form the central evaluative lens for this work. AEI technologies are not examined in isolation but assessed in terms of their effects on the human condition – how they shape relationships, influence social structures, and affect individual and collective flourishing (Ozmen Garibay et al., 2023). Such concepts such as “human flourishing” and “wellbeing” are not explicitly defined in the articles, however they recur throughout the analysis as guiding concerns. Importantly, this human-centred perspective is grounded in values central to Western liberal thought, including individual autonomy, privacy, emotional wellbeing, informed consent, etc. The analysis does not address alternative ethical traditions or cultural frameworks beyond this Western context, and its conclusions should therefore be understood within this normative frame.

A final guiding commitment in my research is the feminist lens through which AEI and social robots are examined. This involves critically exploring how these technologies intersect with gender dynamics and social norms, and how they may shape – and be shaped by – women’s experiences, opportunities, and vulnerabilities. In Article 4 and 5, this perspective is central to

analysing the implications of AEI in intimate and relational contexts. Also, this feminist orientation reinforces the dissertation's pluralist approach and rejection of universalist or monolithic ethical theories. Rather than imposing rigid principles regardless of situations, the analysis makes explicit the particularities of context and the lived motivations of those affected. My own philosophical standpoint is made explicit not to claim universal authority, but to situate the analysis within a specific normative situation. This approach resists abstract, decontextualized moral reasoning and instead embraces situated, relational, and context-sensitive ethics, recognizing that the moral significance of AEI social robots depends on who is using them, for what purposes, and within which social conditions.

To sum up, the philosophical stance adopted in this dissertation is:

- Anticipatory – Aiming at identifying and responding to potential ethical issues of AEI social robots before and as the technology develops.
- Pluralist – Drawing on multiple normative theories as relevant to each issue, rather than committing to a single overarching theory.
- Human-centred with a Western-liberal orientation – Placing human experiences, values, and wellbeing at the core of the evaluative lens. AEI technologies are assessed in terms of their effects on relationships, social structures, and individual and collective flourishing.
- Feminist – Analysing how AEI social robots interact with gender dynamics and social norms, particularly in relation to women's experiences and lives.

Together, these four commitments shape in this dissertation the way ethical questions are framed, the kinds of literature and arguments that are prioritised, and the way participant's accounts are understood – not as neutral and measurable data, but as subjective perspectives shaped by individual experiences and social context. Throughout the articles, they guide the selection and critique of ethical frameworks relevant to trust, authenticity, and moral status, and direct attention to lived experience, power relations, and social consequences.

4.1. Methods

Different kinds of research questions demand different methodological strategies (Creswell & Creswell, 2017). Given the complexity of the ethical, social, and psychological questions raised by AEI, this dissertation adopts a combination of methods to investigate distinct but interconnected aspects – a form of methodological triangulation, understood here in an extended sense to denote the use of multiple philosophical methods to examine a problem from different angles, rather than traditional social-scientific sense of combining qualitative and quantitative data collection, analysis, and interpretation (Thurmond, 2001). The methods employed include conceptual analysis, thought experiments, analogical reasoning, and empirical qualitative research. The choice of methods follows directly from the research questions I wanted to answer, as summarised in Table 2. In the following, I explain how each method was employed and how it contributed to answering the respective questions.

- Conceptual analysis was especially useful for R1, R2, and R3 to clarify key concepts – such as trust, moral status, and emotion – central to the ethical evaluation of AEI. As noted earlier, concepts have the potential to mediate our moral understanding of technology. Conceptual analysis allows for more precise ethical assessments by disentangling semantic ambiguities and normative assumptions embedded in discussions of AEI (Cekiera, 2023).
- Thought experiments were used to explore ethically relevant hypothetical or future scenarios. Since many questions addressed in this dissertation cannot yet be answered by empirical cases, this method enables the evaluation of potential social and ethical consequences without the limitation of the scarcity of empirical data (Brown & Fehige, 2023; Walsh, 2011). Thought experiments were used for R1 (to explore the implications of moving beyond current limits of AEI), R2 (to test the implications of placing trust in robots), R3 (to explore scenarios where AEI robots might be treated as if they had moral status), and R4 (to anticipate how AEI sex and love robots might affect norms of intimacy and relationships).
- Analogical reasoning helped analysing the technology of AEI social robots and their ethical consequences by drawing comparisons with more familiar domains (Spielthener, 2014). Analogies are especially useful for emerging technologies, since

they could help us imagine what these technologies could become but also provide a basis for their ethical assessment (Schwarz-Plaschg, 2018). For instance, comparisons between AEI robots with earlier forms of social and sex robots reveals when ethical concerns persist and when they take on new significance (R1-R4); comparisons between human emotions and artificial emotions to clarify both overlaps and differences (R1); comparisons between trust in humans and trust in AEI robots highlights why users tend to project trust onto these robots while also exposing the differences in intentions and benevolence, but also (R2); comparisons between robots and other morally relevant entities to test robots' moral standing (R3); comparisons between sex and love robots and human partners to assess benefits and limits (R4).

- Empirical qualitative research consisted of an in-depth qualitative study using thematic analysis (Braun & Clarke, 2021, 2022). This method was integrated specifically to address R4, since questions about intimacy, companionship, and social acceptance of love robots are deeply shaped by people's lived perspectives. In contrast, R1-R3 focus on conceptual and normative issues and were approached primarily through philosophical analysis, with their empirical dimensions left for future research. This qualitative approach was important to contextualise normative claims (e.g., about robots undermining authenticity or fostering dependency), ground speculative discussions about possible futures of companionship and sexuality, and ensure that the ethical inquiry remains responsive to real human concerns such as loneliness and safety (Lim, 2025). Also, given the novelty of AEI, a qualitative design was more suitable than a quantitative one, since qualitative methods are especially valuable when a topic is new, understudied, and the key variables are not yet well defined (Creswell & Creswell, 2017).

Just as in the social sciences, where triangulation enhances confidence in results by showing consistency across different methods (Thurmond, 2001), the same principle applies here. Questions raised by AEI and social robots are too complex to be resolved by a single approach. Conceptual analysis can clarify definitions and sharpen distinctions, but on its own it cannot capture how those concepts play out in lived realities. Thought experiments extend this analysis by exploring the implications of those clarified concepts and testing our intuitions

about such realities. Yet, such scenarios risk excessive abstraction unless they are connected back to broader contexts. Analogical reasoning helps situate both conceptual distinctions and thought experiments within moral landscapes we already understand. In turn, these analogies can prompt refinements to the concepts themselves and shape the design of more realistic thought experiments. Finally, empirical research grounds the entire process in actual practices and experiences. Each method, by itself, reveals only part of the picture. Taken together, they provide a fuller and more reliable account of the issues at stake. Therefore, applying triangulation to ethical analysis is a deliberate strategy for capturing the multidimensional nature of ethical problems posed by AEI, thereby producing more credible and well-rounded findings.

Research question	Conceptual analysis	Thought experiments	Analogical reasoning	Empirical research
R1	✓ Clarify emotion and authenticity	✓ Explore implications of moving beyond current technical limits	✓ Compare artificial vs. human emotions	—
R2	✓ Analysis of <i>trust</i>	✓ Consider implications of placing trust in robots	✓ Compare trust in human vs. trust in robots	—
R3	✓ Clarify moral status	✓ Consider scenarios of robots as moral patients	✓ Compare robots with other morally relevant entities	—
R4	—	✓ Anticipate implications for intimacy norms and relationships	✓ Compare robots with other sex technologies and human relationships	✓ Explore women's perspectives on love robots

Table 2 – Research design

5. Article summaries

In this section, I summarise the five papers. Each summary describes the research questions and aims, the main arguments, and the conclusions reached in each article.

5.1. Article 1: Feeling machines: Challenges and ethical perspectives in developing robots with emotions

Article 1 provides a theoretical foundation for understanding AEI's potential and limits forming the basis for ethical discussions in subsequent articles. It situates the broader thesis within a critical exploration of the interplay between technology, emotion, and societal values. It addresses two primary objectives: (1) examining the technological progress and challenges in developing AEI social robots, and (2) analysing the ethical limits and opportunities of this pursuit. Central to these discussions is an exploration of the authenticity of artificial emotions compared to human emotions and whether achieving such authenticity is a goal that should ever be pursued. These insights into emotional authenticity link to further discussions on trust, moral consideration, and the nature of intimate relationships with AEI social robots.

The article begins by situating AEI within the field of affective computing and social robotics, drawing on foundational research in human emotional processes, such as Damasio's theories on embodied emotions (Damasio, 2003). Emotions, integral to human cognition and social interactions, are explored as a means to increase robots' social presence and improve their decision-making capabilities. AEI enables robots to recognise and express emotions for more effective communication and social engagement. It also equips robots with internal emotional parameters – such as anxiety or confidence – to optimise their behaviour and activity-selection. Currently, advancements in AEI are promising in recognising emotions, while face challenges in expressing emotions and integrating internal emotional processes into these systems.

Building on this foundation, the article examines how users perceive social robots' emotional responses as authentic, despite the technological challenges. Here, anthropomorphism plays a crucial role, as humans instinctively attribute human-like traits and emotions to robots. This tendency amplifies the perception that robots possess genuine emotions, even though this perception stems more from human interpretation than from the robots themselves. In other

words, what we interpret as authentic emotion does not necessarily align with the robot's internal emotional processes. This disparity – between the perceived external emotions and the internal states of robots – highlights a distinction between the simulation of emotions in robots, no matter how convincing, and the subjective experience which instead characterises human emotions. While artificial emotions are functional simulations designed to optimise interaction, they lack the subjective, experiential quality that defines human emotions. Therefore, at AEI's current stage, artificial emotions remain external and instrumental to the robot's purposes, rather than signifying genuine emotional experiences, regardless of when humans interpret them as convincingly authentic.

The discussion extends to robots' integrated emotions, such as anxiety and confidence, and their role in supporting adaptive decision-making and behaviour within the system. The integration of an internal emotional apparatus raises a critical question: can a robot equipped with such an apparatus experience genuine, proper emotions? Can the artificial emotions generated by this apparatus meaningfully compare to human emotional processes? The article highlights significant gaps in equating artificial emotional systems with the nuanced nature of human emotions, including the absence of a unified definition of emotion, unresolved questions about robot consciousness, and the differences of physical embodiment between humans and robots. However, rather than attempting to fully replicate human emotions, the article takes a constructive approach and propose viewing robots' emotions as functional and distinct from human ones. Artificial emotions, designed to fulfil specific roles or tasks, may not resemble human feelings but can still be valuable for communication and interaction. This perspective focuses on creating emotional systems that are meaningful within robotic contexts, rather than striving for authenticity and direct equivalence to human emotions.

The article concludes with a jump into the future, where we succeed in replicating human emotions. What are the ethical and social implication of such a success, if so we shall call it? The article explores how robots might transform relationships, intimacy, and family structures, as well as reshape human views on impermanence, life, and death. Finally, AEI is closely linked to the broader goal of achieving artificial general intelligence (AGI), given that emotional intelligent is a key component of human cognition. However, this possibility introduces

significant ethical concerns, including threats to human values and existential risks, emphasising the need for careful deliberation about the pursuit of advancements in AEI.

5.2. Article 2: In technology we trust! But should we?

Article 2 builds upon the discussion in Article 1 by delving deeper into the effects of AEI social robots' emotional capacities on our interactions with them. This article centres on the concept of trust, addressing the research question: Can trust be meaningfully attributed to technology? To address this, the article first provides a conceptual analysis of trust, distinguishing between reliability and affective trust, and then explores this broad question through the specific case of trust in AEI social robots.

Social robots that provide not only physical assistance but also emotional support benefit greatly from being perceived as safe and trustworthy. Designers aim to create systems that elicit trust in users, recognising that trust is fundamental to successful collaboration and the forming of lasting, effective relationships between humans and robots. AEI plays here a crucial role due to three features: anthropomorphism, deception, and emotional attachment. While these three elements are not unique to AEI, their combination and strength in AEI social robots allow these systems to elicit trust feelings in their users. As the sophistication of anthropomorphic and deceptive elements increases, so does the likelihood of users forming emotional attachments to these robots. This, in turn, shifts the perception of robots from mere objects to relational subjects, fostering the development of affective trust.

Affective trust is built on the entity's benevolent intentions, ethical conduct, and honesty. Developing this kind of trust with AEI social robots, based on their perceived benevolence and integrity, has profound implications both for HRI and humanity more in general, as it carries several risks. One major concern is overtrust, where users overestimate the robot's abilities, placing unwarranted faith in its performance and exposing themselves to unnecessary risks. Another critical issue involves the protection of private information, as users may share sensitive data with robots they perceive as trustworthy. This opens the door to possible manipulation and increases the risk of digital surveillance, as such robots might be exploited for purposes that compromise user privacy and autonomy.

Therefore, despite the emotional attachment that may be formed with these systems, the article concludes that the synthetic nature of AEI social robots is morally relevant for the

assessment of whether people should place trust in them. Specifically, AEI social robots are tools whose purposes may diverge from or conflict with those of their users. Although people may develop a feeling of trust towards these robots, granting affective trust in AEI social robots – and technology in general – is not morally justified.

5.3. Article 3: Artificial emotions and the evolving moral status of social robots

As anticipated in Article 2, the perception of AEI social robots evolves from being seen as mere objects to being regarded as relational subjects, falling in between traditional dichotomies of sentient/not sentient, intelligent/not intelligent, and alive/not alive. This shift has profound implications on the ethical standing of social robots, explored in Article 3 through the research question: what is the potential impact of AEI on the moral status of social robots? By examining how AEI interacts with and potentially reshapes the two dominant perspectives on robots' moral status, namely the property oriented approach and the social-relational approach, the article provides new insights into this ongoing debate.

Starting with the property-based approach grounded in the properties of reason and sentience, the article examines the parallels between artificial and human emotions. It argues that the development of AEI in social robots significantly influences the extent to which these robots can mimic or approximate such properties. However, AEI social robots lack consciousness and with it the capability of truly experience emotions, which serves as a foundation for reason and sentience. As a result, AEI social robots cannot claim moral standing. This argument, however, is weakened by the lack of unanimity in the definitions of both consciousness and emotions. Depending on how these concepts are defined, AEI has the potential to alter perceptions of robots' moral status and blur the distinctions between humans and machines.

Within the social-relational account, moral value is not intrinsic to an entity but is instead the outcome of social dynamics. This means that moral considerations are not based on authenticity of the robot's emotions, but rather on the user's feelings towards it, the relationship they can form, and the experiences the robot facilitates. An example comes from sex robots, which, while originally designed for physical gratification, have evolved to engage in broader forms of intimacy and connection. These interactions highlight how robotic entities

can address emotional, social, and psychological needs, extending their relevance to contexts such as care and companionship. This exemplifies a key aspect of the social-relational approach, where the emotional authenticity of the robot is secondary to the emotional fulfilment, companionship, and psychological comfort users derive from their relationships with these machines. However, the article emphasises challenges within this theory that the introduction of AEI does not seem enough to overcome, such as the possible adoption of a relativistic viewpoint of moral status. Nonetheless, the role of AEI is central to this approach, as it creates new opportunities for its advocates and potentially reshaping views on the moral status of social robots.

In conclusion, while there is no definitive answer to the question of whether social robots deserve moral consideration, the article highlights how the implementation of AEI brings robots closer to humans in more profound and fundamental ways, deepening the complexity of the debate.

5.4. Article 4: The robot will feel you now: The ethics of artificial emotional intelligence in sex robots

Sex robots equipped with AEI present distinct ethical challenges compared to other social robots, primarily due to their use in deeply intimate contexts and their unique ability to evoke powerful emotions, including love. The increased (and intentional) emotional involvement of the user introduces new or more concerning moral dilemmas. Article 4 explores these challenges, examining how AEI might exacerbate or potentially mitigate the ethical concerns surrounding sex robots, particularly with regard to sexual health and the dynamics of human relationships. While the term *love robots* could have better reflected their role in developing loving relationships, this article continues to use *sex robots* to align with the terminology prevalent in existing ethical debates and the majority of literature on the topic.

The article begins by presenting general considerations of the most frequently debated benefits and concerns associated with sex robots designed solely for sexual activities. It explores their potential contributions to sexual health, their impact on gender dynamics, and their broader implications for societal views on women. Proponents argue that sex robots can enhance sexual health and wellbeing by improving sexual satisfaction and serving as tools for sexual education and self-exploration, fostering a positive attitude towards sexuality.

However, critics raise concerns that sex robots may promote unrealistic expectations of real-life partners due to idealised body types and performances, while reinforcing harmful gender norms.

These issues might be intensified with the integration of AEI, as sex robots expand into emotional and romantic roles, potentially leading to love and complicating the ethical debate. Key issues include (1) users experiencing social isolation and loneliness due to reduced human contact; (2) emotional deception; (3) societal impacts on human partnerships and marriages; and (4) the risk for users to be manipulated. Conversely, AEI could amplify the benefits discussed, such as increased sexual satisfaction, more effective sexual education, as well as more fulfilling solution for companionship for lonely people. Also, by implementing AEI, the risk of deception and social isolation might be mitigated by deploying AEI to monitor the user's emotional attachment, support their personal growth and skill development, and facilitate the transition from sex robots to human partners.

The article concludes with an emphasis on the lack of empirical data as a significant barrier to fully understand the philosophical and ethical assumptions regarding the impact of AEI sex robots on human wellbeing and society.

5.5. Article 5: A qualitative study on love robots' potential in addressing involuntary singleness among women

The study conducted in Article 5 investigates the potential of love robots equipped with AEI as companions for involuntary single women. Involuntary singleness has been traditionally understood and researched as a predominantly male problem, and the experiences of women facing challenges in forming romantic and sexual relationships have received less attention. To address this gap, this qualitative study engaged involuntary single women living in Norway, exploring their attitudes towards, and perceptions of, advanced robots designed to offer emotional intimacy and romantic companionship. The research explicitly moves away from the widely used term "sex robots" opting instead for "love robots" to better capture the emotional and affectionate potential of these technologies.

Participants expressed complex view, balancing perceived advantages against significant reservations. Positively, love robots were imagined as customizable and reliable partners who

could provide consistent emotional support, intimacy, and safety, therefore potentially alleviating feelings of loneliness and isolation. Participants also saw practical benefits in robots designed to help build social skills, self-confidence, and emotional intelligence. At the same time, participants questioned the authenticity of robotic affection, emphasising that love, to be meaningful, must include autonomy, unpredictability, and genuine mutual desire – qualities found lacking in robotic companions programmed to respond predictably.

Another important theme that emerged was uncertainty about how to conceive or address such robots. Participants often oscillated between speaking of them as machines and as human-like partners, revealing a tension in how people might relate to these technologies: neither fully object nor fully subject.

Finally, the study revealed a deep ambivalence regarding the societal acceptance of relationships with robots. Many participants feared significant social stigma attached to women openly engaging with love robots, influenced by traditional gender norms and media portrayals of intimacy and sexuality. While some believed love robots might challenge societal norms positively by normalising women's emotional and sexual autonomy, others worried that reliance on robots could diminish interpersonal skills, reinforce unhealthy relationship standards, or exacerbate loneliness by isolating users from human contact.

6. Discussion

Let us step back and consider how all five articles work together to address the overarching questions of this thesis: how AEI social robots reshape the way human form, maintain, or perceive relationships, and what ethical and social challenges emerge from these new forms of companionship.

In pursuing these aims, two central themes emerged: the development of trust and the potential evolution of robots' moral standing. Both of these dimensions emerge from a shared phenomenon: the formation of relationships between humans and robots. Instead of treating robots as mere tools, we begin to see them as beings which we can form meaningful bonds with. This idea of human-robot relationships (HRR) is recurrent through the entire thesis (even if this exact term is not often used), becoming even more pronounced when we consider the case of sex and love robots.

The concept of HRR can be understood as a specific subset of research within the broader field of human-robot interaction (HRI). In the literature, HRR is sometimes referred to as human-robot social interaction (Sheridan, 2016), or simply not clearly distinguished from HRI (De Graaf, 2016; Laban et al., 2024). While HRI can encompass all forms of engagement between humans and robots (Sheridan, 2016), HRR focuses more narrowly on emotionally significant and enduring connections that mirror or simulate aspects of interpersonal relationships (Prescott & Robillard, 2021). According to the Cambridge Dictionary (n.d.-a), *interaction* is “an occasion when two or more people or things communicate with or react to each other”. In contrast, *relationship* is “the way in which two things are connected”, or “the way in which two or more people feel and behave towards each other”, or “a close romantic friendship between two people”(Cambridge Dictionary, n.d.-b). What I want to emphasise here is the temporal scope – interactions may be momentary and situational, while relationships are necessarily more enduring – and the increased or even intended possibility for emotional involvement. Therefore, HRR emphasises relational depth, emotional engagement, and continuity over time (Laban et al., 2024), making it particular relevant and accurate to the ethical questions explored in this research. For instance, as I argue in Article 2, affective trust between humans and robots does not arise from one-off exchanges but from the kind of ongoing, emotionally loaded relationships users develop with their machines.

Similarly, Article 3's analysis of the social-relational approach to moral status only makes sense if we recognise that such relationships are indeed deep and continuous, as conceptualised within HRR. In Article 4 and 5 this is particularly explicit, as the benefits and challenges of sex and love robots are analysed in terms of relationships that are sustained over time and often deeply personal.

In the context of AEI, relationships emerge through repeated, meaningful exchanges in which users begin to assign emotional or personal value to these interactions. Starke et al. (2024) contribute to this understanding with the concept of *synthetic relationships*, as “continuing associations between humans and AI tools that interact with one another wherein the AI tool(s) influence(s) humans’ thoughts, feelings and/or actions”. Unlike simple interactions with non-AI technologies, synthetic relationships emerge because AI systems are designed to mimic human traits and encourage human tendency to anthropomorphise. Although Stark et al. focus on AI voice assistants like Alexa, Siri, and Google Assistant, the logic applies – perhaps even more strongly – to embodied social robots with AEI. As I argue in Article 1, embodiment allows for richer and more convincing emotional expression, perceived as more authentic and reliable than the cues produced by chatbots.

The ethical implications of relationships with AEI social robots are not just more complex, but also operate at a different level than those involved in HRI. As this thesis highlights, HRR raises deeper questions about privacy, manipulation, loss of autonomy, and the impact on human-human relationships. It is through the bonds we form with robots that we question whether they have moral status, that we develop affective trust although we should know better, and that love robots are questioned as a possibility for those who do not wish to be alone. Understanding HRR – recognising and examining the nature and qualities of these bonds – is therefore essential to navigating the complex ethical, social, and emotional landscapes that AEI social robots invite. Each of the five articles in this thesis contributes to this understanding.

Article 1 establishes the groundwork, looking at how users *perceive* and *respond* to emotional expressions, regardless of their ontological status; namely, whether it comes from a robot or another living entity. This sets the stage for understanding how trust and moral consideration might arise when we start perceiving robots as they “feel”. Article 2 takes these foundational ideas and digs deeper into what it means to trust a machine capable of simulating emotional states, distinguishing between trust based on reliability and affective trust that involves

emotional investment. It also introduces the challenge posed by AEI to the traditional ontological boundaries between objects and subjects, directly connecting to Article 3, which examines how AEI could transform our views on a robots' moral status. If we start relating to these machines as social partners rather than objects, how does that change their ethical standing? These questions gain a tangible urgency in Article 4, which explores scenarios where intimacy, love, and sex come into play. Here, the impact of HRR becomes even clearer. In intimate contexts, the ethical stakes might be higher, and considerations around autonomy, consent, and emotional wellbeing cannot be ignored. Finally, Article 5 brings in empirical perspectives by exploring how involuntary single women perceive the idea of love robots. These voices remind us that real-world contexts matter, and that both policy and design must be sensitive to the diverse social landscapes into which AEI social robots are being introduced. Taken together, these articles form a coherent progression from broad theoretical inquiry to focus on empirical analysis.

In review, this dissertation has shown that emotionally intelligent social robots raise profound ethical questions, with both promising benefits and serious risks, moving from theoretical foundations to questions of trust, moral status, and lived experiences. When viewed through consequentialism, AEI social robots have been assessed by weighing their potential benefits and harms in terms of outcomes for individuals and society. Among the positive impacts, robots can enhance communication, social presence, and companionship (Article 1), building trust that supports effective collaboration (Article 2), and offering emotional and/or intimate support to those experiencing loneliness or involuntary singleness (Article 4 and 5). However, consequentialism also draws attention to significant risks: overtrust in AEI robots may expose users to privacy breaches and manipulation (Article 2), while replacing human relationships with robotic companionship may exacerbate isolation, weaken social bonds, and reinforce harmful gender norms (Articles 3, 4, 5). Furthermore, at a broader societal level, AEI may introduce long-term dangers such as undermining family structures (Article 4, 5) and contributing to existential risks tied to artificial general intelligence (Article 1). Virtue ethics highlights similar tensions. On the positive side, AEI robot may promote empathy and kindness by modelling supporting behaviours, and they can challenge restrictive norms around intimacy and gender roles (Article 1, 4, 5). Yet they also risk promoting inauthentic forms of love and intimacy, encouraging vices such as selfishness, dependency, and a lack of courage

to engage in authentic human relationships (Articles 4, 5). At the same time, deontological analysis stresses that simulated emotions and deception often violate principles of respect and dignity (Article 1, 2). By integrating these insights, we can move toward practical guidelines, informed policies, and responsible design principles that respect human values and help ensure AEI social robots serve our well-being rather than undermine it.

Surely, I could have addressed so many more and different ethical aspects of AEI implementation into social robots. Ethical analyses, especially those involving complex technologies like AI, often require more time, debate, and refinement before we reach any sense of satisfaction or closure. In fact, in my experience, concepts like “satisfaction” or “definitive answers” are almost nonexistent in the realm of ethical analysis. Still, I believe that focusing on foundational issues – such as trust, moral status, and love – offers a necessary baseline given the novelty of AEI. *Foundational* issues, as they are also at the very heart of human society. Trust is a social mechanism that holds communities together, enabling cooperation, stability, and a sense of safety among individuals. Moral status determines who (or what) deserves ethical consideration, guiding how we distribute rights, responsibilities, and protections. Love and emotional bonds cultivate empathy, understanding, and support, often forming the basis of families and similar institutions. This groundwork can, in turn, support more specialised explorations, such as privacy, consent, emotional dependency, and other emerging concerns, that will undoubtedly arise as these technologies continue to evolve.

Also, the ethics of AEI could have involved much more than the sole application to social robots. After all, AEI can find a variety of applications, including entertainment and retail, healthcare (Kaur & Sharma, 2021), but also job hiring and workplace monitoring (Monteith et al., 2022). Nonetheless, my project focuses on AEI social robots for the following reasons. One central consideration is the embodied nature of emotions, as proposed by theories like the James-Lange theory or hybrid models like appraisal theory, previously discussed, which suggest a strong connection between emotions and physiological stimuli. While I do not explicitly support any specific theory, acknowledging this perspective allows for exploring various avenues in the development of artificial emotions without prematurely limiting potential approaches.

The second reason concerns the role of embodiment in emotional interactions. A robot's physical embodiment significantly enhances its ability to foster emotional interactions. By providing tangible cues, such as gestures, facial expressions, and movements, robots make emotional states observable and interpretable, creating a more immersive and intimate experience. Embodiment also generates a sense of presence that virtual agents cannot replicate, while multimodal communication through visual, auditory, and tactile channels further enriches interaction experiences. To sum up, the integration of AEI into social robots for companionship highlights the unique intersection of emotional engagement, physical embodiment, and ethical concerns, particularly those surrounding the authenticity and implications of emotional bonds formed with such robots.

6.1. Implications for practice

With this thesis, I aim to promote research and interdisciplinary dialogue between roboticists, ethicists, psychologists, sociologists, and policymakers. As HRR becomes more common, such collaborations essential to develop guidelines and policies that ensure ethical design and deployment. These policies must consider issues like autonomy, privacy, informed consent, and the prevention of emotional exploitation, especially for vulnerable groups. To that end, people interacting with AEI social robots should be informed about what the machine can and cannot do, the nature of their behaviours, and how they fundamentally differ from human intentions and responses. Such awareness might help users better discern the nature of their attachment towards the robot and maintain healthy emotional boundaries. Also, as robots become more integrated into daily life – caring for the elderly, tutoring children, or providing companionship – constant evaluation is needed. This means assessing user wellbeing, monitoring whether these relationships lead to isolation from human contact, and adjusting design principles and policies as we learn more.

An interesting suggestion, leaning more on a preventive side than corrective, comes from Elder (2017), introduced in section 3.3.2. She argues that people should first learn how to cultivate human meaningful relationships before forming similar bonds with robots. Seeking to protect individuals who may not yet have the maturity or the emotional tools needed to handle certain types of interactions responsibly is not an unreasonable objective. For example, Australia is about to implement a ban on social media use for those under 16 years

old (Ritchie, 2024). This policy aims to shield minors from potential harms associated with early social media exposure by preventing them from engaging until a certain level of maturity is presumed. Similarly, Elder's argument posits that people should develop social skills and emotional intelligence before being introduced to artificial agents that might simulate social relationships. An imposed restriction, therefore, may help ensure that users possess the necessary emotional grounding and critical understanding to engage safely and beneficially with complex, potentially influencing technologies.

Another solution, as discussed in Article 2 and 4, might come directly from AEI itself. The idea is that AEI could be designed to monitor emotional states and the dynamics of HRR, identifying when these relationships might pose risks to the individual's wellbeing. AEI might, for instance, evaluate when a relationship becomes overly dependent, or when the user starts to withdraw from human connections. This would represent a sort of integrated solution, as one might call it, presented directly in the design AEI social robots. Of course, such a solution would require significant trust in AEI, as we would placing our wellbeing directly in its hands, bringing us back to issues on overtrust and misuse. But it might still be considered an option, depending on how far advancements in AEI will progress and the safeguards that are implemented to address these risks.

6.2. Further research needs

This dissertation has provided valuable insights into the ethics of emotionally intelligent social robots, but it has also revealed important gaps and limitations that require further research.

As mentioned above, AEI is a rapidly developing field, and its applications extend far beyond the scope of my research, including healthcare, education, eldercare, and customer service. These expanding applications are already reflected in regulatory discussions, such as the European Union's AI Act (2024), which prohibits the use of emotion recognition technologies in contexts like workplaces and educational institutions. Empirical and normative research of the kind presented here can help inform such regulations and ensuring they remain sensitive to the ethical complexities of HRR. What I want to stress here, however, is the need for a stronger and more systematic focus on the ethical questions raised by these applications, with ethicists more directly involved in applied settings to ensure that design and deployment are guided by values as well as technical goals.

There is also a pressing need for empirical studies that examine real-life, prolonged interactions between humans and AEI social robots. Although some studies have begun to examine long-term interactions with social robots (e.g., De Graaf et al., 2016; Leite et al., 2013; Van Maris et al., 2020; Wada & Shibata, 2006, 2007). Longitudinal research in everyday environments – household, workplaces, or care institutions – would be essential for scholars to track how trust, attachment, dependency, or resistance develop in sustained interactions, as well as how these interactions affect other human relationships.

In addition, another limitation of my research is the demographic scope of the qualitative study on AEI love robots, focused on a small group of women in Norway. While this provided valuable insights, future research would benefit from expanding participant groups to include different genders, sexual orientations, age groups, cultural contexts and vulnerabilities. This would capture a fuller spectrum of human responses to AEI love robots and ensure that findings are more representative and inclusive.

Finally, while this thesis has primarily operated within Western ethical and philosophical frameworks, it is important to acknowledge the cultural limitations of these perspectives. Concepts such as individual autonomy, consent, and wellbeing, carry different meanings and weight across societies. Future research would benefit from engaging more deeply with non-Western ethical traditions that may offer alternatives on how emotional AI technologies might be interpreted, integrated, or resisted. These frameworks can challenge dominant assumptions, provide culturally sensitive insights, and help ensure that policy and design principles are globally inclusive and responsive.

REFERENCES

- 1x. (n.d.). *Introducing NEO Gamma*. 1x. Retrieved September 3 from
- Ahmed, E., Buruk, O. O., & Hamari, J. (2024). Human–robot companionship: Current trends and future agenda. *International Journal of Social Robotics*, 16(8), 1809-1860.
- Alexander, L., & Moore, M. (2024). Deontological Ethics. In E. N. Z. a. U. Nodelman (Ed.), *The Stanford Encyclopedia of Philosophy*. <https://plato.stanford.edu/archives/win2024/entries/ethics-deontological/>
- Arnold, M. B. (1960). *Emotion and personality*. Columbia University Press.
- Bao, A., Zeng, Y., & Lu, E. (2023). Mitigating emotional risks in human-social robot interactions through virtual interactive environment indication. *Humanities and Social Sciences Communications*, 10(1), 1-9.
- Bartneck, C., & McMullen, M. (2018). Interacting with anatomically complete robots: A discussion about human-robot relationships. Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction,
- Blanchard, E. G., Volfson, B., Hong, Y.-J., & Lajoie, S. P. (2009). Affective artificial intelligence in education: From detection to adaptation. *AIED*,
- Blue Frog Robotics. *Buddy, the first smart, mobile and emotional robot at an affordable price*. Blue Frog Robotics. Retrieved September 3 from <https://www.bluefrogrobotics.com/robot>
- Botelho, L. M., & Coelho, H. (1998). Artificial emotions. *Department of Information Sciences and Technologies, ISCTE, Lisbon, Portugal, Technical Report ISCTE-DCTI-1998, 2*.
- Bower, G. H., & Forgas, J. P. (2000). Affect, memory, and social cognition. *Cognition and emotion*, 87-168.
- Brady, M. S. (2013). *Emotional insight: The epistemic role of emotional experience*. OUP Oxford.
- Braun, V., & Clarke, V. (2021). *Thematic analysis: A practical guide*. Sage.
- Braun, V., & Clarke, V. (2022). Conceptual and design thinking for thematic analysis. *Qualitative psychology*, 9(1), 3.
- Brey, P. A. (2012). Anticipatory ethics for emerging technologies. *NanoEthics*, 6(1), 1-13.
- Brooks, R. (2021). *Artificial intimacy: Virtual friends, digital lovers, and algorithmic matchmakers*. Columbia University Press.
- Brown, J. R., & Fehige, Y. (2023). Thought Experiments. In E. N. Z. a. U. Nodelman (Ed.), *The Stanford Encyclopedia of Philosophy*. <https://plato.stanford.edu/archives/win2023/entries/thought-experiment/>
- Buchanan, T. W. (2007). Retrieval of emotional memories. *Psychological bulletin*, 133(5), 761.
- Business Research Insights. (2025). *Companion Robots Market Size, Share, Growth, and Industry Analysis, By Type (Humanoid Robot, Animal-like Robot, & Others), By Application (Children, The Aged, & Others), and Regional Insight and Insight and Forecast to 2032*. Business Research Insights. <https://www.businessresearchinsights.com/market-reports/companion-robots-market-119093>
- Cambridge Dictionary. (n.d.-a). Interaction. In *Cambridge Dictionary*. Retrieved July 13, from <https://dictionary.cambridge.org/dictionary/english/interaction>
- Cambridge Dictionary. (n.d.-b). Relationship. In *Cambridge Dictionary*. Retrieved July 13, from <https://dictionary.cambridge.org/dictionary/english/relationship>
- Campa, R. (2016). The rise of social robots: a review of the recent literature.

- Campaign against sex robots.* (n.d.). Retrieved 06/12/2024 from <https://campaignagainstsexrobots.org/>
- Cannon, W. B. (1929). *Bodily changes in pain, hunger, fear, and rage: An account of recent researches into the function of emotional excitement*. Branford.
- Cekiera, K. (2023). Thought Experiments and Conceptual Analysis in Ethics. *Studia Philosophica Wratislaviensia*, 18(1), 29-43.
- Chu, M. D., Gerard, P., Pawar, K., Bickham, C., & Lerman, K. (2025). Illusions of intimacy: Emotional attachment and emerging psychological risks in human-ai relationships. *arXiv preprint arXiv:2505.11649*.
- Coeckelbergh, M. (2020). *AI ethics*. MIT press.
- Coeckelbergh, M. (2022). Three responses to anthropomorphism in social robotics: Towards a critical, relational, and hermeneutic approach. *International Journal of Social Robotics*, 14(10), 2049-2061.
- Coeckelbergh, M., Pop, C., Simut, R., Peca, A., Pintea, S., David, D., & Vanderborght, B. (2016). A survey of expectations about the role of robots in robot-assisted therapy for children with ASD: ethical acceptability, trust, sociability, appearance, and attachment. *Science and engineering ethics*, 22, 47-65.
- Copp, D. (2005). *The Oxford handbook of ethical theory*. Oxford University Press.
- Cordeschi, R. (2007). AI turns fifty: revisiting its origins. *Applied Artificial Intelligence*, 21(4-5), 259-279.
- Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., & Taylor, J. G. (2001). Emotion recognition in human-computer interaction. *IEEE Signal processing magazine*, 18(1), 32-80.
- Creswell, J. W., & Creswell, J. D. (2017). *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications.
- Curzer, H. J. (2012). *Aristotle and the Virtues*. Oxford University Press.
- Damasio, A. (1994). Descartes' error: Emotion, rationality and the human brain. *New York: Putnam*, 352.
- Damasio, A. (2003). *Looking for Spinoza: Joy, Sorrow, and the Feeling Brain*. Harcourt inc.
- Damiano, L., & Dumouchel, P. (2018). Anthropomorphism in human–robot co-evolution. *Frontiers in Psychology*, 9, 468.
- Danaher, J. (2017). The Symbolic-Consequences Argument in the Sex Robot Debate. In J. Danaher & N. McArthur (Eds.), *Robot Sex: Social and Ethical Implications*. MIT Press.
- Danaher, J. (2019). The philosophical case for robot friendship. *Journal of Posthuman Studies*, 3(1), 5-24.
- Danaher, J. (2020a). Robot Betrayal: a guide to the ethics of robotic deception. *Ethics and Information Technology*, 22(2), 117-128.
- Danaher, J. (2020b). Welcoming robots into the moral circle: A defence of ethical behaviourism. *Science and engineering ethics*, 26(4), 2023-2049.
- Danaher, J., Earp, B., & Sandberg, A. (2017). Should we campaign against sex robots?
- Danaher, J., & Sætra, H. S. (2023). Mechanisms of techno-moral change: A taxonomy and overview. *Ethical theory and moral practice*, 26(5), 763-784.
- Darling, K. (2012). Extending legal protection to social robots. *IEEE Spectrum*.
- Darling, K. (2015). 'Who's Johnny?' Anthropomorphic framing in human-robot interaction, integration, and policy. *Anthropomorphic framing in human-robot interaction, integration, and policy (March 23, 2015)*. *Robot Ethics*, 2.

- Dautenhahn, K., Woods, S., Kaouri, C., Walters, M. L., Koay, K. L., & Werry, I. (2005). What is a robot companion-friend, assistant or butler? 2005 IEEE/RSJ international conference on intelligent robots and systems,
- De Graaf, M. M. (2016). An ethical evaluation of human–robot relationships. *International Journal of Social Robotics*, 8(4), 589-598.
- De Graaf, M. M., Ben Allouch, S., & Van Dijk, J. A. (2016). Long-term evaluation of a social robot in real homes. *Interaction studies*, 17(3), 462-491.
- De Sousa, R. (1987). *The Rationality of Emotion*, Cambridge Mass. In: MIT Press.
- Deonna, J., & Teroni, F. (2012). *The emotions: A philosophical introduction*. Routledge.
- Devlin, K. (2015). In defence of sex machines: why trying to ban sex robots is wrong. *The conversation*.
- Devlin, K. (2018). *Turned on: Science, sex and robots*. Bloomsbury Publishing.
- Devlin, K. (2024). Relating with social robots: issues of sex, love, intimacy, emotion, attachment, and companionship. *The De Gruyter handbook of robots in society and culture*. De Gruyter.
- Dewey, J. (1895). The theory of emotion. *Psychological review*, 2(1), 13.
- Dittmer, J. Applied Ethics. In J. a. D. Fieser, Bradley (Ed.), *Internet Encyclopedia of Philosophy*.
- Döring, N., & Pöschl, S. (2018). Sex toys, sex dolls, sex robots: Our under-researched bed-fellows. *Sexologies*, 27(3), e51-e55.
- Driver, J. (2011). *Consequentialism*. Routledge.
- Driver, J. (2022). Moral Theory. In E. N. Z. a. U. Nodelman (Ed.), *The Stanford Encyclopedia of Philosophy*. <https://plato.stanford.edu/archives/fall2022/entries/moral-theory/>
- Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2), 124.
- Elder, A. (2017). Robot friends for autistic children: Monopoly money or counterfeit currency? In *Robot ethics 2.0: From autonomous cars to artificial intelligence* (pp. 113-126). Oxford University Press.
- Fong, T., Nourbakhsh, I., & Dautenhahn, K. (2003). A survey of socially interactive robots. *Robotics and autonomous systems*, 42(3-4), 143-166.
- Forlizzi, J. (2007). How robotic products become social products: an ethnographic study of cleaning in the home. Proceedings of the ACM/IEEE international conference on Human-robot interaction,
- Forlizzi, J., & DiSalvo, C. (2006). Service robots in the domestic environment: a study of the roomba vacuum in the home. Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction,
- Fortunati, L., & Edwards, A. (2024). *The De Gruyter Handbook of Robots in Society and Culture* (Vol. 3). Walter de Gruyter GmbH & Co KG.
- Fosch-Villaronga, E., & Poulsen, A. (2020). Sex care robots: exploring the potential use of sexual robot technologies for disabled and elder care. *Paladyn, Journal of Behavioral Robotics*, 11(1), 1-18.
- Frank, L., & Nyholm, S. (2017). Robot sex and consent: Is consent to sex between a robot and a human conceivable, possible, and desirable? *Artificial intelligence and law*, 25, 305-323.
- Fridlund, A. J. (2014). *Human facial expression: An evolutionary view*. Academic press.
- Gaspar, K., & Clore, G. L. (2002). Attending to the big picture: Mood and global versus local processing of visual information. *Psychological science*, 13(1), 34-40.
- Generation Robots. (n.d.). NAO. In.

- Gensler, H. J. (2017). *Ethics: A contemporary introduction*. Routledge.
- Goldie, P. (2002). Emotions, feelings and intentionality. *Phenomenology and the Cognitive Sciences*, 1(3), 235-254.
- Gordon, J.-S., & Nyholm, S. (2021). Ethics of artificial intelligence.
- Green, O. H. (1992). Evaluative Theories of Emotions. In *The Emotions: A Philosophical Theory* (pp. 61-76). Springer Netherlands. https://doi.org/10.1007/978-94-011-2552-9_5
- Hare, R. M. (1979). What is wrong with slavery. *Philosophy & Public Affairs*, 103-121.
- Heyselaar, E. (2023). The CASA theory no longer applies to desktop computers. *Scientific Reports*, 13(1), 19693.
- Hofer, M. A. (2006). Psychobiological roots of early attachment. *Current directions in psychological science*, 15(2), 84-88.
- Honda Research Institute. *Haru the Social Robot*. Honda Research Institute. Retrieved September 3 from <https://www.jp.honda-ri.com/en/activity/haru-the-social-robot/>
- Hopster, J., & Löhr, G. (2023). Conceptual engineering and philosophy of technology: Amelioration or adaptation? *Philosophy & technology*, 36(4), 70.
- Hopster, J. K., Rueda, J., & Hillenbrink, R. (2025). Conceptual Mediation in Technomoral Change: Reply to Danaher and Sætra. *Ethical theory and moral practice*, 1-9.
- Horwitz, J. (2025). *Meta's flirty AI chatbot invited a retiree to New York - He never made it home*. Reuters. Retrieved 7/9 from <https://www.reuters.com/investigates/special-report/meta-ai-chatbot-death/>
- Hursthouse, R. a. P., G. . (2023). Virtue Ethics. In E. N. Z. a. U. Nodelman (Ed.), *The Stanford Encyclopedia of Philosophy*. <https://plato.stanford.edu/archives/fall2023/entries/ethics-virtue/>
- James, W. (1884). What is an emotion? *Mind*, 9, 188-205.
- Jecker, N. S. (2021). Sociable robots for later life: Carebots, friendbots and sexbots. *Sex robots: Social impact and the future of human relations*, 25-40.
- Johnson, D. G. (2010). The role of ethics in science and engineering. *Trends in Biotechnology*, 28(12), 589-590.
- Kagan, J., Kearsley, R. B., & Zelazo, P. R. (1980). *Infancy, its place in human development*. Harvard University Press.
- Kagan, S. (2018). *Normative ethics*. Routledge.
- Kagan, S. (2019). *How to count animals, more or less*. Oxford University Press.
- Kahn Jr, P. H., Friedman, B., Perez-Granados, D. R., & Freier, N. G. (2004). Robotic pets in the lives of preschool children. CHI'04 extended abstracts on Human factors in computing systems,
- Kaur, S., & Sharma, R. (2021). Emotion AI: integrating emotional intelligence with artificial intelligence in the digital workplace. Innovations in Information and Communication Technologies (IICT-2020) Proceedings of International Conference on ICRIHE-2020, Delhi, India: IICT-2020,
- Kenny, A. (2003). *Action, emotion and will*. Routledge.
- Kim, J., Kim, S., Kim, S., Lee, E., Heo, Y., Hwang, C.-Y., Choi, Y.-Y., Kong, H.-J., Ryu, H., & Lee, H. (2021). Companion robots for older adults: Rodgers' evolutionary concept analysis approach. *Intelligent Service Robotics*, 14, 729-739.
- Korn, O., Bieber, G., & Fron, C. (2018). Perspectives on Social Robots: From the Historic Background to an Experts' View on Future Developments. Proceedings of the 11th Pervasive Technologies Related to Assistive Environments Conference,

- Kubes, T. (2019). New materialist perspectives on sex robots. A feminist dystopia/utopia? *Social Sciences*, 8(8), 224.
- Laban, G., Kappas, A., Morrison, V., & Cross, E. S. (2024). Building long-term human–robot relationships: Examining disclosure, perception and well-being across time. *International Journal of Social Robotics*, 16(5), 1-27.
- Law, T., Chita-Tegmark, M., Rabb, N., & Scheutz, M. (2022). Examining attachment to robots: benefits, challenges, and alternatives. *ACM Transactions on Human-Robot Interaction (THRI)*, 11(4), 1-18.
- Leite, I., Martinho, C., & Paiva, A. (2013). Social robots for long-term interaction: a survey. *International Journal of Social Robotics*, 5, 291-308.
- Lerner, J. S., Li, Y., Valdesolo, P., & Kassam, K. S. (2015). Emotion and decision making. *Annual review of psychology*, 66, 799-823.
- Levy, D. (2007). *Love + Sex with Robots: The Evolution of Human-Robot Relationships*. HarperCollins. <https://books.google.no/books?id=PJ4sAAAAYAAJ>
- Lievesley, R., Reynolds, R., & Harper, C. A. (2023). The ‘perfect’ partner: Understanding the lived experiences of men who own sex dolls. *Sexuality & Culture*, 27(4), 1419-1441.
- Lim, W. M. (2025). What is qualitative research? An overview and guidelines. *Australasian Marketing Journal*, 33(2), 199-229.
- Lin, P., Abney, K., & Bekey, G. (2011). Robot ethics: Mapping the issues for a mechanized world. *Artificial intelligence*, 175(5-6), 942-949.
- LOVOT. LOVOT. LOVOT. Retrieved September 4 from <https://lovot.life/en/>
- McArthur, N. (2017). The Case for Sexbots. In J. Danaher & N. McArthur (Eds.), *Robot sex: Social and ethical implications* (pp. 0). The MIT Press. <https://doi.org/10.7551/mitpress/10718.003.0007>
- Melson, G. F., Kahn Jr, P. H., Beck, A. M., Friedman, B., Roberts, T., & Garrett, E. (2005). Robots as dogs? Children's interactions with the robotic dog AIBO and a live Australian shepherd. CHI'05 extended abstracts on Human factors in computing systems,
- Moerland, T. M., Broekens, J., & Jonker, C. M. (2018). Emotion in reinforcement learning agents and robots: a survey. *Machine Learning*, 107, 443-480.
- Monteith, S., Glenn, T., Geddes, J., Whybrow, P. C., & Bauer, M. (2022). Commercial use of emotion artificial intelligence (AI): implications for psychiatry. *Current Psychiatry Reports*, 24(3), 203-211.
- Montgomery, B. (2024). Mother says AI chatbot led her son to kill himself in lawsuit against its maker. *The Guardian*. <https://www.theguardian.com/technology/2024/oct/23/character-ai-chatbot-sewell-setzer-death>
- Moors, A. (2020). Appraisal theory of emotion. In *Encyclopedia of personality and individual differences* (pp. 232-240). Springer.
- Moyle, W., Cooke, M., Beattie, E., Jones, C., Klein, B., Cook, G., & Gray, C. (2013). Exploring the effect of companion robots on emotional expression in older adults with dementia: a pilot randomized controlled trial. *Journal of gerontological nursing*, 39(5), 46-53.
- Müller, V. C. (2020). Ethics of artificial intelligence and robotics.
- Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. Proceedings of the SIGCHI conference on Human factors in computing systems,
- Nestor, M. W., & Wilson, R. L. (2022). Introduction to anticipatory ethics. In *Anticipatory ethics and the use of CRISPR in humans* (pp. 25-35). Springer.

- Norman, D. A. (2004). *Emotional Design: Why We Love (or Hate) Everyday Things*. Basic Books.
<https://books.google.no/books?id=z2jvRlqhdlwC>
- Nyholm, S., & Frank, L. E. (2019). It loves me, it loves me not: is it morally problematic to design sex robots that appear to love their owners? *Techne: Research in Philosophy & Technology*, 23(3).
- Ozmen Garibay, O., Winslow, B., Andolina, S., Antona, M., Bodenschatz, A., Coursaris, C., Falco, G., Fiore, S. M., Garibay, I., & Grieman, K. (2023). Six human-centered artificial intelligence grand challenges. *International Journal of Human–Computer Interaction*, 39(3), 391-437.
- Pai, R. Y., Shetty, A., Dinesh, T. K., Shetty, A. D., & Pillai, N. (2024). Effectiveness of social robots as a tutoring and learning companion: a bibliometric analysis. *Cogent Business & Management*, 11(1), 2299075.
- PARO. (n.d.). *PARO Therapeutic Robot*. PARO. Retrieved September 3 from <http://www.parorobots.com/>
- Paroseal. (n.d.). PARO. In.
- Pessoa, L., & Ungerleider, L. G. (2004). Neuroimaging studies of attention and the processing of emotion-laden stimuli. *Progress in brain research*, 144, 171-182.
- Picard, R. W. (1999). Affective Computing for HCI. HCI (1),
- Picard, R. W. (2000). *Affective computing*. MIT press.
- Picard, R. W. (2003). What does it mean for a computer to “have” emotions. *Emotions in humans and artifacts*, 213-235.
- Prescott, T. J., & Robillard, J. M. (2021). Are friends electric? The benefits and risks of human-robot relationships. *Iscience*, 24(1).
- Qviro. (n.d.). NEO Gamma. In.
- Realdoll. (n.d.). *Realdoll Robot*. <https://www.realdoll.com/realdoll-robot/>
- Reeves, B., & Nass, C. (1996). The media equation: How people treat computers, television, and new media like real people. *Cambridge, UK*, 10(10), 19-36.
- Reisenzein, R. (2009). Emotional experience in the computational belief–desire theory of emotion. *Emotion Review*, 1(3), 214-222.
- Research and Markets. (2025). *Social Robots Market - Global Forecast 2025-2030*. Research and Markets. Retrieved August 31 from <https://www.researchandmarkets.com/report/united-states-social-robot-market>
- Richardson, K. (2016a). The asymmetrical 'relationship' parallels between prostitution and the development of sex robots. *Acm Sigcas Computers and Society*, 45(3), 290-293.
- Richardson, K. (2016b). Sex robot matters: slavery, the prostituted, and the rights of machines. *IEEE Technology and Society Magazine*, 35(2), 46-53.
- Ritchie, H. (2024). Australia approves social media ban on under-16s. *BBC*.
<https://www.bbc.com/news/articles/c89vji0lxx9o>
- Rosenbloom, P. S., Gratch, J., & Ustun, V. (2015). Towards emotion in sigma: from appraisal to attention. Artificial General Intelligence: 8th International Conference, AGI 2015, AGI 2015, Berlin, Germany, July 22-25, 2015, Proceedings 8,
- Ross, W. D. (2002). *The right and the good*. Oxford University Press.
- Sætra, H. S. (2020). The foundations of a policy for the use of social robots in care. *Technology in Society*, 63, 101383.
- Sætra, H. S. (2021). Social robot deception and the culture of trust. *Paladyn, Journal of Behavioral Robotics*, 12(1), 276-286.

- Sætra, H. S. (2022a). Loving robots changing love: Towards a practical deficiency-love. *Journal of future robot life*, 3(2), 109-127.
- Sætra, H. S. (2022b). Robotomorphy: Becoming our creations. *Ai and Ethics*, 2(1), 5-13.
- Sætra, H. S. (2022c). Robotomorphy: Becoming our creations. *Ai and Ethics*, 2(1), 5-13.
- Sætra, H. S., & Danaher, J. (2022). To each technology its own ethics: The problem of ethical proliferation. *Philosophy & technology*, 35(4), 93.
- Sayre-McCord, G. (2023). Metaethics. In E. N. Z. a. U. Nodelman (Ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/spr2023/entries/metaethics/>
- Scarantino, A., & De Sousa, R. (2018). Emotion. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*.
- Scheutz, M. (2009). The Inherent Dangers of Unidirectional Emotional Bonds between Humans and Social Robots. In P. Lin, K. Abney, & G. A. Bekey (Eds.), *Robot ethics: The ethical and social implications of robotics*. MIT Press.
- Schmager, S., Pappas, I. O., & Vassilakopoulou, P. (2025). Understanding Human-Centred AI: a review of its defining elements and a research agenda. *Behaviour & Information Technology*, 1-40.
- Schuller, D., & Schuller, B. W. (2018). The age of artificial emotional intelligence. *Computer*, 51(9), 38-46.
- Schwarz-Plaschg, C. (2018). The power of analogies for imagining and governing emerging technologies. *NanoEthics*, 12(2), 139-153.
- Shamsuddin, S., Yussof, H., Ismail, L., Hanapiah, F. A., Mohamed, S., Piah, H. A., & Zahari, N. I. (2012). Initial response of autistic children in human-robot interaction therapy with humanoid robot NAO. 2012 IEEE 8th International Colloquium on Signal Processing and its Applications,
- Sharkey, A., & Sharkey, N. (2021). We need to talk about deception in social robotics! *Ethics and Information Technology*, 23, 309-316.
- Sharkey, N., & Sharkey, A. (2020). The crying shame of robot nannies: an ethical appraisal. In *Machine ethics and robot ethics* (pp. 155-184). Routledge.
- Sharkey, N., Wynsberghe, A. v., Robbins, S., & Hancock, E. (2017). Our Sexual Future with Robots. A Foundation for Responsible Robotics Consultation Report. In: The Hague.
- Sheridan, T. B. (2016). Human–robot interaction: status and challenges. *Human factors*, 58(4), 525-532.
- Sica, A. (2023). *The Robot will Feel You Now: The Ethics of Artificial Emotional Intelligence in Sex Robots* IEEE Ro-Man 2023, Busan.
- Singer, P. (1986). *Applied ethics*. Oxford University Press.
- Singer, P. (2011). *Practical ethics*. Cambridge university press.
- Sinnott-Armstrong, W. (2023). Consequentialism. In E. N. Z. a. U. Nodelman (Ed.), *The Stanford Encyclopedia of Philosophy*. <https://plato.stanford.edu/archives/win2023/entries/consequentialism/>
- SoftBank Robotics. (n.d.). *For better business just add Pepper*. SoftBank Robotics. Retrieved September 3 from <https://us.softbankrobotics.com/pepper>
- Solomon, R. C. (1988). On emotions as judgments. *American Philosophical Quarterly*, 25(2), 183-191.
- Sparrow, R., & Sparrow, L. (2006). In the hands of machines? The future of aged care. *Minds and Machines*, 16, 141-161.

- Spielthener, G. (2014). Analogical reasoning in ethics. *Ethical theory and moral practice*, 17(5), 861-874.
- Stark, L., & Hoey, J. (2021). *The ethics of emotion in artificial intelligence systems* Conference on Fairness, Accountability, and Transparency, New York, NY, USA.
- Starke, C., Ventura, A., Bersch, C., Cha, M., de Vreese, C., Doeblner, P., Dong, M., Krämer, N., Leib, M., & Peter, J. (2024). Risks and protective measures for synthetic relationships. *Nature Human Behaviour*, 8(10), 1834-1836.
- Strömfelt, H., Zhang, Y., & Schuller, B. W. (2017). *Machine Learning: Overview of an Emerging Domain* Affective Computing and Intelligent Interaction (ACII 17),
- Sullins, J. P. (2012). Robots, love, and sex: the ethics of building a love machine. *IEEE Transactions on Affective Computing*, 3(4), 398-409.
- Sung, J.-Y., Guo, L., Grinter, R. E., & Christensen, H. I. (2007). "My Roomba is Rambo": intimate home appliances. UbiComp 2007: Ubiquitous Computing: 9th International Conference, UbiComp 2007, Innsbruck, Austria, September 16-19, 2007. Proceedings 9,
- Tesla. (n.d.). *AI & Robotics*. Tesla. Retrieved September 3 from <https://www.tesla.com/AI>
- Tesla Car World. (2025). *It Happened! Elon Musk Reviews \$15,989 Tesla Bot Gen 3 Ready for Massive Reservations in October!* YouTube. <https://youtu.be/hkNgMlnwRdg>
- Thiroux, J. P., & Krasemann, K. W. (2014). *Ethics theory and practice*. Pearson Education Limited.
- Thurmond, V. A. (2001). The point of triangulation. *Journal of nursing scholarship*, 33(3), 253-258.
- Tomkins, S. S. (2008). *Affect imagery consciousness: the complete edition: two volumes*. Springer publishing company.
- Turkle, S. (2011). *Alone together: Why we expect more from technology and less from each other*. Basic Books.
- Turkle, S. (2020). A nascent robotics culture: New complicities for companionship. In *Machine ethics and robot ethics* (pp. 107-116). Routledge.
- Van Maris, A., Zook, N., Caleb-Solly, P., Studley, M., Winfield, A., & Dogramadzi, S. (2020). Designing ethical social robots—a longitudinal field study with older adults. *Frontiers in Robotics and AI*, 7, 1.
- Van Wynsberghe, A. (2022). Social robots and the risks to reciprocity. *AI & SOCIETY*, 37(2), 479-485.
- Vanhoffelen, G., Vandenbosch, L., & Schreurs, L. (2025). Teens, Tech, and Talk: Adolescents' Use of and Emotional Reactions to Snapchat's My AI Chatbot.
- Venkatesh, V. (2000). Determinants of perceived ease of use: Integrating control, intrinsic motivation, and emotion into the technology acceptance model. *Information systems research*, 11(4), 342-365.
- Vincent, J. (2006). Emotional attachment and mobile phones. *Knowledge, Technology & Policy*, 19(1), 39-44.
- Wada, K., & Shibata, T. (2006). Robot therapy in a care house-its sociopsychological and physiological effects on the residents. Proceedings 2006 IEEE International Conference on Robotics and Automation, 2006. ICRA 2006.,
- Wada, K., & Shibata, T. (2007). Living with seal robots—its sociopsychological and physiological influences on the elderly at a care house. *IEEE transactions on robotics*, 23(5), 972-980.

- Walsh, A. (2011). A moderate defence of the use of thought experiments in applied ethics. *Ethical theory and moral practice*, 14(4), 467-481.
- Wired. (n.d.). In.
- Ye, M., Schneiders, E., Lee, W.-Y., & Jung, M. (2023). The Future of Home Appliances: A Study on the Robotic Toaster as a Domestic Social Robot. 2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN),
- Zhang, R., Li, H., Meng, H., Zhan, J., Gan, H., & Lee, Y.-C. (2025). The dark side of ai companionship: A taxonomy of harmful algorithmic behaviors in human-ai relationships. Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems,
- Zhao, S. (2006). Humanoid social robots as a medium of communication. *New Media & Society*, 8(3), 401-419.

Article 1

Feeling machines: Challenges and ethical perspectives in developing robots with emotions

1. Introduction

The integration of emotions into artificial systems, particularly social robots, represents a frontier in both robotics and artificial intelligence research, reflecting a profound shift towards more human-like machines and enhanced human–robot interaction (HRI) (Chursinova & Stebelska, 2021, p. 54; Fong et al., 2003). This chapter provides an in-depth exploration of the emerging realm of “feeling machines”, focusing on the design challenges and ethical perspectives associated with developing robots equipped with emotions.

The pursuit of creating machines with emotions stems from an increasing recognition of the essential role that emotions play in human behaviour, communication, interaction, alongside cognition and intellect. Emotions for machines can be useful both for making machines more effective in their interactions with humans *and* potentially for enabling machines to make better—or at least more human-like—decisions. Emotional intelligence is the ability to perceive, understand, and manage one’s own emotions and the emotions of others (Kambur, 2018). Those possessing a high level of emotional intelligence excel in interpersonal communication, displaying empathy, and forming strong relationships (Schutte et al., 2007). Given its significance, incorporating emotional intelligence into social robots, for which empathetic and effective communication is crucial, could greatly enhance interactions with users by increased perceived authenticity and a more effective social presence (Fong et al., 2003).

Furthermore, research by Damasio (1994) has been instrumental in understanding that emotions are not peripheral to rational thinking but are central to the operation of the human mind. This evolving research underscores that emotions are crucial, pervasive drivers of intelligent behaviour and decision making. Emotions influence judgements and choices by altering the content and importance of our thoughts, including any information related to

decision processes, and shifting our implicit motivations and goals (Lerner et al., 2015). Emotions shape attentional focus—fear narrows it to immediate details, while positive affect broadens it to a wider perspective. This dynamic influences both short-term reactions and long-term cognitive outcomes (Gasper & Clore, 2002; Pessoa & Ungerleider, 2004). Another example of emotions influencing cognition involves memory and working memory performance (Levens & Phelps, 2008). Our mood not only affects what we remember but also how we take in new information and how we feel when recalling these memories (Bower & Forgas, 2000; Buchanan, 2007).

This integral role of emotions in human processes underpins their potential utility in robotic systems. The quest to create social robots with emotions is not a mere exercise in technological advancement but a necessary step towards the creation of truly intelligent machines, able to participate in and contribute to human social environments in a way that feels natural and intuitive to us (Picard, 2003). Central to this endeavour is the technology of Artificial Emotional Intelligence (AEI). The aims of AEI are twofold. First, to enhance communication with users by enabling social robots to express and recognise others' emotions. Second, to enrich the robot's internal mechanisms for behaviour organisation and activity selection, thereby optimising decision making and responsiveness in complex environments.

AEI social robots are still far from being fully realised, but we already have machines capable of quite sophisticated emotion recognition and mimicking abilities. In this chapter, we explore the frontiers of their development, delving into the feasibility and implications of AEI technology. Particularly, we ask: To what extent have we succeeded in embedding emotions into systems like social robots, and what are the limits, both technologically and ethically, of this pursuit? We proceed as follows. Section 2 defines AEI and explores its main functions, particularly how it applies to social robots. Section 3 examines the dynamics of how social robots display emotions, discussing how these emotions are perceived through anthropomorphism, simulated, and integrated into their systems. Section 4 delves into the quest of authenticity in robots' emotions. It discusses the philosophical and practical challenges in assessing whether artificial emotions can truly equate to human emotions. Section 5 explores the potential social and ethical implications if social robots were to successfully replicate human-like emotions, considering changes in social structures,

interpersonal relationships, and the moral status of such robots. Finally, Sect. 6 concludes by summarising the discussions and emphasising the need for continued exploration in this complex field.

2. Artificial Emotional Intelligence

Let us start with defining AEI, also known as affective computing. This field of studies within AI research finds its greatest potential in developing computers “to be genuinely intelligent [...] and to interact naturally with us” (Picard, 2000). Schuller and Schuller (2018) define AEI as a technology that possesses three main functions: emotion recognition, emotion generation, and emotion augmentation. The first, *emotion recognition*, enables AEI to detect others’ emotions by analysing emotion-related data through multimodal processes. This means that the computer recognises emotions by using a variety of channels or modalities, including facial expressions, body language, tone of voice, and physiological indicators like heart rate and galvanic skin response. This multimodal approach significantly enhances the accuracy of emotion recognition, offering a more comprehensive understanding than single-modality methods (Weng & Lin, 2022). Modern AI systems are today routinely promoted as having such capabilities, as seen, for example, in OpenAI’s launch of GPT-4o in May 2024. In that launch, they highlighted the ability of the system to detect user emotions (Noone, 2024).

Furthermore, given its embodiment, the robot utilises such modalities not just to detect but also to express emotions. This second capability is called *emotion generation* and allows the robot to respond appropriately to the user’s needs and expectations, consequently enhancing affective and social interactions. Progress in this area has been slower compared to the research for emotion recognition (Krakovsky, 2018). Nevertheless, the emergence of advanced chatbots has sparked debates on their emotional capacity, as seen in former Google engineer Blake Lemoine’s claim of chatbot sentience (Chanan, 2023). Returning to the GPT-4o example, OpenAI also explicitly state that the system can “express emotion” (OpenAI, 2024) and communicate in “different emotive styles” (Noone, 2024).

However, emotion generation is about external emotional expressions, which does not provide insight into the robot’s internal states. In that regard, research is advancing in *emotion augmentation*, a third capacity that enables the system to determine behaviour and activity selection based on simulated inner emotions. For instance, incorporating anxiety and

confidence as emotional parameters can significantly improve the machine's efficiency in planning and achieving goals (Khashman, 2008).

2.1. AEI into social robots

AEI is applied across various domains within AI and computing, serving distinct purposes, such as employee hiring and workplace monitoring (Monteith et al., 2022). This chapter, however, will focus on the use of AEI in social robots— autonomous physical entities designed to engage with users socially and emotionally, learning and evolving from these interactions (Darling, 2012). This is because social robots with AEI offer unique capabilities, not found in other purely virtual machines, such as ChatGPT, Replika AI companions, etc.

Social robot design, often mirroring animal or human forms, allows for the more complex and complete expression of emotional states in an embodied manner, potentially leading them to a level of authenticity, naturalness, and reliability not found in non-embodied technologies like chatbots or virtual characters. Moreover, according to some theories, emotions are necessarily embodied experiences (Damasio, 2003), and this might imply that emotion augmentation is *only* possible for embodied AI systems. As a result, AEI in social robots enables interactions that are emotionally resonant, marking a significant step towards creating robots that can participate in human social environments in a deeply integrated manner.

Implementing AEI into social robots becomes particularly relevant across various applications, especially in areas where the capacity to understand, adapt, and respond to users' emotional states is particularly significant (Leite et al., 2013). While some AEI applications remain theoretical, there are many examples both emotion recognition and generation in both AI systems and robots. However, with the recent rapid developments of AI systems, it is reasonable to anticipate further enhancements such a technology could introduce to the functionalities of existing social robots, especially within healthcare contexts. Consider the case of Paro, a robot whose presence in therapeutic settings is already notable (Wada & Shibata, 2006, 2007). We can plausibly expect that, by incorporating more advanced AEI, Paro's interactions could be further improved, enabling it to detect and interpret subtle emotional shifts of elderly patients and responding in more dynamic, tailored ways. This advancement would not only deepen emotional bonds but also provide support that is intricately aligned with the individual's emotional needs, enhancing the quality of their

experience. Similar applications for AEI social robots could be envisioned in educational settings (Blanchard et al., 2009; Pai et al., 2024), and private and domestic environments (Sica, 2023).

However, the advancement of AEI in social robots and their integration into society necessitates a closer examination of the nature of emotions within these artificial agents. As interactions between humans and robots grow in complexity, there might be growing expectations for robots to exhibit sophisticated social behaviours (Arbib & Fellous, 2004). This challenge is not merely about enabling robots to display simulated emotions. It is about equipping them with the capability to deeply understand and appropriately respond to human emotional states. This leads to a pivotal question: If a robot can accurately interpret and respond to human emotions, does this imply it possesses emotions of its own, or is it simply exhibiting a high level of mimicry? Addressing this question is vital for the ongoing development and application of AEI social robots. It pushes us to reconsider how our perceptions of emotional authenticity within robots impacts not only HRI, but also our societal and ethical norms.

3. How an AEI social robot manifests emotions: Perception, simulation, and authenticity

As the integration of emotional intelligence progresses, it becomes imperative to examine how these artificial entities manifest emotions. This section delves into the complex dynamics between the appearance of emotional expressions in social robots, and their underlying mechanisms. We explore three critical aspects: the perception of emotions facilitated by anthropomorphism, the simulation of emotional states for functional HRI, and the integration of emotional apparatus internal to the robot system. This last aspect, in particular, leads to the question of authenticity in artificial emotions. Each facet contributes uniquely to our understanding of what it means for a robot to “feel” and how these manufactured emotions influence interactions with humans.

3.1. Perceived emotions and anthropomorphism

Human perceptions often ascribe human emotional and personal qualities to robots, not necessarily based on the robot’s internal architecture but rather on our tendency to attribute

human-like traits to them. Such phenomenon, known as anthropomorphism, is generally defined as the inclination to attribute human characteristics to non-human entities (such as non-human animals, natural forces, and, yes, technological devices), or to interpret non-human behaviours as being driven by human-like motivations, intentions, or emotions (Damiano & Dumouchel, 2018).

Social robots are specifically designed to be anthropomorphised, as it can facilitate social exchange between the robot and the user and promote HRI accordingly (Healy, 2022). As these robots are deployed in a growing range of fields, many of their applications specifically depend on their anthropomorphic features. For instance, in areas like healthcare and education, the effectiveness of social robots is significantly increased when users perceive them as a “social agent”, rather than just tools (Darling, 2015). Let us revisit the example of Paro, the therapeutic robot previously discussed. Its design, resembling a baby seal, deliberately invokes emotional responses from users, fostering engagement and comfort (Hung et al., 2019). In such therapeutic environments, Paro’s anthropomorphic qualities, which often lead it to be perceived as a pet or even a child (Coeckelbergh, 2014), enable its placement within a social dimension, with positive effects on users’ overall wellbeing (Hung et al., 2019; Leite et al., 2013). Therefore, in these cases, as Darling argues (2015), leveraging our natural tendency to anthropomorphise these machines is necessary for their effective and beneficial use.

In such contexts, whereas implementing AEI would plausibly enhance anthropomorphism, it is not essential for its emergence. Even the simplest robots that are not intentionally designed for social interaction, such as the vacuum cleaner Roomba or the military robot PackBot, can be seen as exhibiting human traits (Scheutz, 2009). This entails that a robot, to be perceived as socially and emotionally capable and form an emotional bond with their users, does not need to possess internal emotional states: people already greatly contribute to imagine it having feelings and personality. Nevertheless, the more sophisticated the robot becomes, the more likely it is to be anthropomorphised (Whitby, 2011). The integration of AEI, in particular, could enhance the tendency to anthropomorphise by depicting the robot as an entity able to *feel*, leading the users to ascribe even more complex emotional states and personality traits to these robots. We will explore this aspect in the following.

However, before we continue, we must briefly address some ethical concerns that arise from anthropomorphism and the resulting social and emotional engagement with the robot. Human tendency to attribute mental states and intentions to robots allows such entities to enter our social sphere, potentially creating the illusion of genuine social relationships (Turkle, 2011). Consequently, anthropomorphism in social robots can lead to deception, i.e. the false belief that robots *possess* emotions and can reciprocate ours (Scheutz, 2009). The ethics of anthropomorphism and the potential for deception remain debated, with some authors such as Darling (2015) and indirectly Levine and Schweitzer (2015) arguing in favour of accentuating anthropomorphic tendency when it benefits the user.

3.2. Playing the “social game”

Here, we examine how robots can simulate emotions through AEI. Unlike anthropomorphism, where human-like qualities are projected onto robots by users, emotional behaviours in AEI social robots are specifically designed to evoke emotions and to simulate social cues. As described in Sect. 2, AEI aims to equip robots to interpret and adapt to human feelings and actions in real time, as well as to generate appropriate emotional responses. Such capabilities, which we referred to as emotion recognition and emotion generation, could transform the landscape of HRI by allowing robots to participate meaningfully in social interactions. Significant progress in social robots such as Pepper (Fiorini et al., 2022), Buddy (Weiss & Hannibal, 2018) and Ryan (Abdollahi et al., 2022) have been made. Equipping robots with human-like empathic abilities, however, still proves challenging (MarcosPablos & García-Peñalvo, 2022).

Although AEI is not yet fully developed, let us consider a future scenario where it has advanced sufficiently to be capable of recognition and generation of emotions. For instance, a social robot equipped with AEI could analyse a user’s facial expressions, voice tone, and body language to discern underlying emotional states. Such insights would enable the robot to respond in a way that is not only appropriate but also contextually aware, potentially offering comfort, humour, or empathy, depending on the situation. This level of interaction goes beyond simple programmed responses and begins to touch the realm of genuine social exchanges, with consequent ethical questions. While our previous discussion around anthropomorphism raised concerns about the deceptive aspect of human-robot relationships, AEI introduces a deeper and perhaps more concerning layer of ethical complexity. If a robot’s

emotional responses are so convincingly realistic, what does this mean for the authenticity of the emotions themselves? Are these expressions of genuine emotional experiences, or are they sophisticated simulations designed to mimic human emotions? These questions are not merely theoretical. They carry significant implications for the design, deployment, and perception of social robots.

At the level of emotion generation and recognition, the robot's emotional manifestations may not be linked with genuine emotional experiences within the robot. Instead, we can expect that these are calculated responses, programmed specifically to simulate emotional processing, and not rooted in any sentient experience (Sica & Sætra, 2024). Such simulations are purely external and instrumental for the robot to fulfil specific roles or achieve tasks for effective HRI. This functional mimicry should not be confused with the experiential reality of emotions.

Nevertheless, authors such as Coeckelbergh (2010) argue that focusing on whether robots' emotions are "real" might miss a crucial point. In human interactions, we do not have direct access to another's internal emotions. Instead, we depend on external cues such as behaviour, expressions, and words, to interpret what the other person might be feeling or thinking. Thus, if a robot can emulate these cues convincingly, its emotions might be considered "authentic" in terms of their functional impact on human observers. For Coeckelbergh, what truly matters in interactions—whether with humans or robots—is not the internal experience but the external expression and its effects. This perspective shifts the emphasis from a technical analysis of artificial emotions to their practical implications in social contexts, suggesting that the authenticity of a robot's emotions may be assessed based on their believability and efficacy in social interactions, rather than their fidelity to human emotions.

We are not advocating specifically for or against this perspective. However, for the purpose of this chapter, the question at hand is not just about the external manifestations of emotions in robots but also concerns the fundamental nature of what it means for an emotion to be *genuine*. As we continue our exploration, we aim to critically assess the capacities of robots to experience emotions that meet criteria not only for believability but also for authenticity within their interactive and functional contexts.

3.3. Integrated emotions: Beyond simulation?

In this subsection, we focus on the third capability of AEI, namely emotion augmentation. Differently from emotion generation and emotion recognition, machines with emotion augmentation are designed to replicate *some form of* emotional apparatus to enhance their functionality. More specifically, informed by Damasio's theories (1994), advances in AEI have led to the development of machines that incorporate emotion-based algorithms as key components in decision-making, responsiveness, and interactions (Moerland et al., 2018; Rosenbloom et al., 2015; Strömfelt et al., 2017).

Strömfelt et al.'s study (2017) provide a detailed analysis of how artificial emotions are integrated into AI systems at a fundamental level to enhance performance. One key example presented is the inclusion of an emotional backpropagation-learning algorithm in smart pattern recognition programmes (Khashman, 2008), where the emotional biases of confidence and anxiety are applied during the system's learning process. This emotional framework improves both the recognition rate and the learning efficiency, as it allows the system to better focus on new or interesting data that may otherwise be overlooked by traditional learning algorithms.

Building on Khashman's work, Yang and colleagues' model (2012) offers a novel view on how emotions influence decision-making. Their model features an emotion generator composed of sensations, feelings, emotions, and a hormonal system that interact to impact the feeling component. These components interact to affect the system's emotional state with feelings like anxiety, triggered by diminishing rewards and confidence, which acts opposite to anxiety. Other emotions such as fear and warmth influence strategic decisions and indicate when to stop the algorithm based on the number of iteration and level of fear. These emotional cues give rise to a primary emotion that plays a crucial role in adjusting the algorithm parameters and potentially shifting its strategies and actions.

Another example discussed is the development of emotional conditioning models that closely mimic human emotional processes. These models integrate various emotional aspects—including, again, anxiety and confidence—to adapt AI responses based on the emotional state of the system. This approach has proven effective in applications ranging from robotic controllers to pattern recognition systems, demonstrating significant improvement in

flexibility and adaptability over standard models. Such emotional AI systems showcase their ability to not just perform tasks, but also to learn and adapt in ways that are inspired by human emotional intelligence (see Strömfelt et al., 2017 for further details).

Emotion augmentation can be applied across a range of technologies. Yet, its application becomes particularly compelling in robotics when considering whether artificial emotions could match the quality of human ones. This emphasis stems from theories suggesting that emotions can only be experienced within a physical body. Therefore, when comparing emotions in robots to human emotions, the embodied form of robots may allow for a more direct analogy with human emotional processes. As we will see later, embodiment may be crucial for the assessment of the authenticity of the robot's emotional capabilities.

To sum up, while AEI is still exploring ways to model and generate artificial emotions, it is evident that using emotional elements into AI systems such as machine learning could greatly enhance their performance and learning efficiency, but if we successfully integrate such an artificial emotional system into a robot, does this mean the robot experiences genuine, proper emotions?

4. “The million-dollar question: Will AI have emotions?”

This question, originally posed by Schuller and Schuller (2018, p. 44), remains a million-dollar question: Will AI ever achieve proper human-like emotions? This also relate to age-old questions in computer science, such as “can machines think?” (Turing, 1950), and more general question regarding whether technology that mimics human and animal functionality can be said to be doing the same. For example, does a plane fly, does a submarine swim, etc. In this section, we explore the challenges in assessing whether artificial emotions can genuinely compare to human emotions. We first provide an overview of various definitions and theories of emotions and then compare these to the concept of artificial emotions. Next, we explore the intricate connection between emotions, consciousness, the physical body, and the social sphere, discussing how these relationships may imply complications in drawing parallels between artificial and human emotions. As our discussion unfolds, it becomes clear that there is no straightforward answer to their potential equivalency.

4.1. Understanding emotions

The concept of human emotions lacks a unified definition in psychology. Scarantino and De Sousa (2018) discuss three primary theoretical frameworks: the James-Lange, Evaluative Theory, and Motivational Theories. Each theoretical framework offers a different perspective on what emotions are and how they function, which influences how we might understand and design AEI. Our aim here is not to debate the validity of these perspectives. Instead, we focus on how the variety of approaches to understanding emotions highlights a fundamental issue, that the problem of defining emotions is unresolved. Without a universal understanding of these concepts, which fundamentally define us as humans, the effort to create emotional robots that resemble us becomes exceedingly difficult.

As we have seen when introducing the AEI capability of emotion recognition, designers often simplify the understanding of emotion to the expression of biophysiological signals such as facial expression, gait, or blood conductivity. This approach is rooted in the Motivational Traditions, which posits that human emotional expressions are regular and identifiable (Stark & Hoey, 2021). Such a theoretical choice carries technical as well as ethical concerns. For example, a frown would be interpreted simply as an expression of sadness or anger, missing the evaluative nuances that might indicate confusion, concentration, or even physical discomfort. This might lead to systems that misinterpret human behaviour and intentions, leading to potentially inappropriate or harmful decisions by robots. Moreover, it risks reinforcing stereotypes and reducing the complex spectrum of human emotions to a narrow set of recognisable patterns (Stark & Hoey, 2021). In this context, therefore, replicating authentic emotions in robots remains a challenging and ethically complex effort.

4.2. Consciousness

It is commonly assessed that emotions are subjective experiences and therefore related to consciousness (Sica & Sætra, 2024). Consciousness is here understood as the capacity to have private mental experiences (DeGrazia, 2022), a trait traditionally viewed as lacking in robots (Danaher, 2017; Dennett, 1994). As such, although robots may exhibit signals that mimic the functions of human feelings, this similarity does not imply that their emotional systems are equivalent to ours. In discussing emotion generation, for instance, we did not expect a robot

displaying an emotion to necessarily experiencing it internally. Therefore, by lacking emotional experience, robots might have emotional systems, “but no real feelings” (Picard, 2003).

However, authors such as Frank and Nyholm (2017) caution against prematurely dismissing the possibility that robots might 1 day achieve some level of consciousness, given the current trajectory of research in artificial consciousness. Additionally, the definition and understanding of consciousness are not fixed, which implies that, depending on the theoretical perspective adopted, robots could be considered conscious. Dehaene et al. (2021) argue that if consciousness is viewed as emerging “from nothing more than specific computations” (p. 54), then artificial consciousness might be achievable by mimicking the computational architectures found in the human brain. Others suggest that, even in their present form, robots exhibit some level of consciousness. For instance, by defining consciousness as the ability to report verbally one’s perceptions, we might say that some AI machines already have consciousness (Frank & Nyholm, 2017). Alternatively, theorists like Dennett (1994) argue that practical requirements can be established for robots to develop a form of consciousness specific to their kind.

While this topic is complex and extends beyond the scope of this chapter, it highlights the prevailing absence of unanimity on the definition of consciousness and how this reflects the authenticity of artificial emotions.

4.3. Emotions in the physical

Earlier in this chapter, we stressed the importance of the embodied nature of social robots over virtual systems to generate emotions. This is because, in addition to brain activity, human emotions are deeply connected to our physicality (Damasio, 1994; LeDoux, 1998). For instance, bodily experiences such as stepping on an unstable rock initiate instinctive reactions within us. These reactions include an “alarm” mechanism that detects a possible loss of balance and triggers adrenaline surge to prompt corrective actions (Sloman, 2001). This example highlights how human reactions and emotional responses are directly linked to the body functions. For genuine emotions to emerge, a body that physically interacts with the world is required (Schuller & Schuller, 2018). Therefore, in this view, social robots offer a more realistic and relevant comparison to human emotional experiences than entities that are purely digital.

Some views reject the possibility for an artificial system, regardless of their level of embodiment, to experience emotions such as humans do. Emotion augmentation shows that machines can have mechanisms that partially implement the functions of human emotional systems, and some pain sensors are already utilised in robotics (Picard, 2003; Schuller & Schuller, 2018). However, there is a fundamental difference in bodily architecture between robots and humans. As Chursinova and Stebelska (2021) point out, this disparity means that robots' experiences of reality will inherently differ from ours. Therefore, computers can imitate human emotion only partially, "because our bodies differ" and, as discussed before, "because so little is known about human emotions" (Picard, 2003).

Whether robot and human bodies will ever match, or sufficiently approximate each other for their emotions to be considered equivalent remains uncertain. One could counter the previous views by suggesting that advancements in bio-inspired and synthetic nervous systems in AI might 1 day enable robots to replicate human emotional processes (Kagan et al., 2023; Kagan et al., 2022; Smirnova et al., 2023). The ongoing debate underscores that much is left to be explored and that a definitive answer is not yet within reach.

4.4. Let's mix it up

We have so far discussed the comparison between human and robot emotions, reaching an unclear conclusion. However, we have not considered the possibility that "not all emotions need to be like human emotions" (Arbib & Fellous, 2004, p. 554). What if the substantial differences in physical embodiment, perception, and interaction with reality between humans and robots do not imply that robots lack "genuine" emotions, but rather suggest that their emotional experiences differ essentially from ours and may require a different conceptualisation? This final subsection offers a different view, which considers robot emotions on their own terms, recognising them distinct from human emotions, rather than seeking to equate them. As we will see, among the three capacities of AEI, it is emotion augmentation that provides a critical role in this discussion.

Let us start with considering that not all emotions are human. Research into the emotional processes of non-human animals shows that emotions extend beyond purely human experiences (Arbib & Fellous, 2004; Picard, 2003). Similarly, robot emotions might not need to replicate human emotions to be considered legitimate. Artificial emotions may instead

represent a different spectrum of emotive responses, analogous to organic emotions but distinct in expression and underlying mechanisms.

Furthermore, building upon a functionalist approach, Arbib and Fellous (2004) and Fellous (2004) argue that emotional concepts can apply to robots by focusing on functionality within their specific operational environments, rather than by replicating human emotional experiences. As we discussed, the primary functions of emotions in biological entities, such as enhancing communication, modulating behaviour, and facilitating social interaction, can be similarly valuable in robotics. By adopting a functionalist perspective, emotions in robots are not about the subjective experiences typically associated with human emotions, but about the utility and application of these emotional states in achieving functional outcomes. Emotion augmentation plays a pivotal role here, enabling these systems to adapt and respond effectively within their operational environments.

Along the same lines, Parisi and Petrosino (2010) claim that robots can “*have emotions*” (p. 453), these latter understood as states of an individual’s body and brain that optimise the motivational decision-making processes. They demonstrate this by developing an artificial neural network with an emotional circuit, which improves the robot’s functionality by taking faster and more correct motivational decisions. They also considered the relationship between emotions and body, by implementing output/input processes within the emotional circuit that interact with the robot’s hardware. As the authors sustain, such robots can be said to have emotions, provided that there is a clear functional role for these emotional states in their behaviour and a specific part of their neural network that facilitates these functions. Additionally, they do not consider robots with emotion recognition and emotion generation as actually possessing emotions, since these capacities do not contribute functionally to their behaviour.

Such functionalist approaches underscore how emotion augmentation is crucial in advancing the functionality and adaptiveness of robotic systems therefore supporting the possibility of artificial emotions as a distinct category. We must remember, however, that this discussion does not necessarily address the challenges outlined before. Rather than equating human and artificial emotions, the aim here is to seek an alternative perspective that recognises artificial emotions as unique to robots—designed for specific operational roles—and moves away from direct comparisons with human emotional processes. This calls for an appreciation of robot

emotions in their right, reflecting their practical utility and the distinctive contexts in which they operate.

To conclude Sect. 4, our exploration of the potential for AEI social robots to exhibit humanlike emotions has highlighted the complexity and ongoing nature of this debate. The array of conflicting views and theoretical frameworks precludes a definitive conclusion about the equivalence of human and artificial emotions. Given that emotions are intertwined with physical, cognitive, and social elements, asserting equivalence without deeper understanding could indeed be premature and potentially ethically irresponsible.

5. What if we succeed?

Building on the unresolved debate of Sect. 4, let us delve into a future where we do succeed in artificially replicating human-like emotions. In this final section, we explore possibilities and challenges that such a development could introduce. What ethical and social implications could arise? How will our interactions with social robots evolve, and what new roles will these machines play in our society?

First, even if human-like, or especially because of this, artificial emotions could still present significant issues when implemented. Emotional expressions vary widely across cultures (Hareli et al., 2015). Without cultural sensitivity, AEI might misinterpret or inadequately respond to emotional cues from people of different cultural backgrounds, leading to biased interactions (Chursinova & Stebelska, 2021). Also, societal expectations often dictate how individuals of different genders should express emotions. For instance, women might be expected to display more empathy, while men might be encouraged to express emotions like sadness or fear (Zhu et al., 2020). Harmful biases also arise in recognizing emotions, as studies show a tendency to associate angry faces to men, particularly Black men (Becker et al., 2010; Neel et al., 2012). If robots possess artificial emotions equal to humans', ensuring their emotional systems are free from these biases is crucial. Clearly, as previously addressed, these biases are already problematic in developing AEI, even without the aim to reproduce genuine emotions, adding to the challenges discussed in Sect. 4.

Furthermore, drawing from Sætra's (2022) analysis of the impact of social robots designed for loving relationships, we can extend this argument to robots possessing proper emotions. As the author suggests, such machines could redefine our understanding of love, transforming

traditional concepts such as monogamy and commitment, and encourage more diverse forms of intimacy and companionship. Society might become more accepting of diverse relationship structures, perhaps helping the recognition of other sexual minorities, and the role of technology in meeting emotional needs. Other significant changes could impact marriages, traditional family structures, and reproduction, as in relationships with robots we might expect a decline in birth rates (Sætra, 2022).

The presence of robots with genuine emotions could alter our views on other human values, such as the fundamental concepts of life and death. How will we engage with emotional entities with whom we might share friendships or loving relationships, when such entities do not age or die in the conventional sense? The emotional continuity offered by robots could provide a sense of comfort and stability, yet it might also lead to a shift in how we cope with grief and loss, and challenge our acceptance of life's impermanence. Some individuals might be prompted to prefer robotic companions over humans, seeking to avoid the inevitable heartache associated with human loss.

Moreover, replicating human-like emotions in AEI raises questions about the moral status of these machines. We have argued elsewhere about the moral considerations of social robots with AEI (Sica & Sætra, 2024), suggesting that the absence of equivalence between artificial and human emotions delineates the boundary between moral consideration for machines and individuals. However, if this equivalence is validated, the moral status of AEI machines would need to be re-evaluated, potentially requiring us to extend moral considerations and rights to them as we do to other moral beings. We would then ask: Should we seek consent for their use and deployment? Are they still considered property, belonging to the user or the company that produced them? Will they be held responsible for their actions, or would they lack moral agency? When it comes to loving robots, will they be granted the choice of their romantic partner? These questions also make us reflect on the actual utility of these machines. It is crucial to consider whether the pursuit of genuine emotional capabilities in robots truly serves our needs or introduces more issues than it resolves.

Finally, another critical issue is the role of AEI in the pursuit of artificial general intelligence (AGI), expected to reach and exceed human intelligence (Goertzel, 2014; McLean et al., 2023). While authors such as Bach (2008) considers the progress in AI as largely unsuccessful, the integration of AEI principles could potentially be a transformative breakthrough. As said

before, emotional intelligence is fundamental to human cognition. Understanding and replicating human-like emotional responses is essential for developing AI systems that can think, learn, and interact in ways that are truly analogous to human intelligence. Thus, by progressing in AEI, we might accelerate our approach towards achieving AGI. However, this prospect also introduces new ethical complexities, such as a threat to human values and norms, with the hypothesised event of humanity annihilation (Boyles, 2018; Chalmers, 2016).

6. Conclusion

This chapter explored the integration of AEI into social robots, highlighting both its progress and limitations. As we explored the ways in which robots manifest emotions, we examined three critical aspects: anthropomorphism, simulation of emotions, and internal emotional apparatus within robots. This exploration raised fundamental questions about the authenticity of artificial emotions and their implications for what it truly means for a robot to feel. These questions are not merely technical but extend into deep ethical, social, and philosophical realms, touching on the very nature of emotion and the role of artificial agents in our social fabric. As Picard (2003) reminds us, we are still trying (and we might never succeed) to understand emotions in the living systems, and “the machine is not even alive”. Whether AEI machines will ever genuinely possess emotions akin to humans remains uncertain (Picard, 2003). As this field evolves, it will be crucial to continue engaging with these questions, ensuring that advancements in AEI proceed with an awareness of the deep ethical dimensions and a commitment to enhancing human wellbeing—not merely for technological aspiration but for genuine societal benefit.

REFERENCES

- Abdollahi, H., Mahoor, M., Zandie, R., Sewierski, J., & Qualls, S. (2022). Artificial emotional intelligence in socially assistive robots for older adults: A pilot study. *IEEE Transactions on Affective Computing*, 14(3), 2020–2032.
- Arbib, M. A., & Fellous, J.-M. (2004). Emotions: From brain to robot. *Trends in Cognitive Sciences*, 8(12), 554–561.
- Bach, J. (2008). Seven principles of synthetic intelligence. *Frontiers in Artificial Intelligence and Applications*, 171, 63.
- Becker, D. V., Neel, R., & Anderson, U. S. (2010). Illusory conjunctions of angry facial expressions follow intergroup biases. *Psychological Science*, 21(7), 938–940.
- Blanchard, E. G., Volfson, B., Hong, Y.-J., & Lajoie, S. P. (2009). *Affective artificial intelligence in education: From detection to adaptation*. AIED.

- Bower, G. H., & Forgas, J. P. (2000). Affect, memory, and social cognition. In E. Eich, J. F. Kihlstrom, G. H. Bower, J. P. Forgas, & P. M. Niedenthal (Eds.), *Cognition and Emotion* (pp. 87–168). Oxford University Press.
- Boyles, R. J. M. (2018). A case for machine ethics in modeling human-level intelligent agents.
- Buchanan, T. W. (2007). Retrieval of emotional memories. *Psychological Bulletin*, 133(5), 761.
- Chalmers, D. J. (2016). The singularity: A philosophical analysis. In *Science fiction and philosophy: From time travel to superintelligence* (pp. 171–224).
- Chanan, M. (2023). Artificial stupidity: On the alienation of intelligence. *Free Associations*, 36(88).
- Chursinova, O., & Stebelska, O. (2021). Is the realization of the emotional artificial intelligence possible? Philosophical and methodological analysis. *Filosofija. Sociologija*, 32(1).
- Coeckelbergh, M. (2010). Moral appearances: Emotions, robots, and human morality. *Ethics and Information Technology*, 12, 235–241.
- Coeckelbergh, M. (2014). The moral standing of machines: Towards a relational and non-Cartesian moral hermeneutics. *Philosophy & Technology*, 27, 61–77.
- Damasio, A. (1994). Descartes' error: Emotion, rationality and the human brain (p. 352). Putnam.
- Damasio, A. (2003). Looking for Spinoza: Joy, sorrow, and the feeling brain. Harcourt Inc.
- Damiano, L., & Dumouchel, P. (2018). Anthropomorphism in human–robot co-evolution. *Frontiers in Psychology*, 9, 468.
- Danaher, J. (2017). The symbolic-consequences argument in the sex robot debate. In J. Danaher & N. McArthur (Eds.), *Robot sex: Social and ethical implications*. MIT Press.
- Darling, K. (2012). Extending legal protection to social robots. *IEEE Spectrum*.
- Darling, K. (2015, March 23). 'Who's Johnny?' Anthropomorphic framing in human robot interaction, integration, and policy. In *Anthropomorphic framing in human-robot interaction, integration, and policy* (p. 2). Robot Ethics.
- DeGrazia, D. (2022). Robots with moral status? *Perspectives in Biology and Medicine*, 65, 73–88.
- Dehaene, S., Lau, H., & Kouider, S. (2021). What is consciousness, and could machines have it? In *Robotics, AI, and humanity: Science, ethics, and policy* (pp. 43–56).
- Dennett, D. C. (1994). The practical requirements for making a conscious robot. *Philosophical Transactions of the Royal Society of London. Series A: Physical and Engineering Sciences*, 349(1689), 133–146.
- Fellous, J.-M. (2004). From human emotions to robot emotions. In *Architectures for modeling emotion: Cross-disciplinary foundations*, American Association for Artificial Intelligence (pp. 39–46).
- Fiorini, L., Loizzo, F. G., D'Onofrio, G., Sorrentino, A., Ciccone, F., Russo, S., et al. (2022). Can I feel you? Recognizing Human's emotions during human-robot interaction. In *International Conference on Social Robotics*.
- Fong, T., Nourbakhsh, I., & Dautenhahn, K. (2003). A survey of socially interactive robots. *Robotics and Autonomous Systems*, 42(3–4), 143–166.
- Frank, L., & Nyholm, S. (2017). Robot sex and consent: Is consent to sex between a robot and a human conceivable, possible, and desirable? *Artificial Intelligence and Law*, 25, 305–323.
- Gaspar, K., & Clore, G. L. (2002). Attending to the big picture: Mood and global versus local processing of visual information. *Psychological Science*, 13(1), 34–40.

- Goertzel, B. (2014). Artificial general intelligence: Concept, state of the art, and future prospects. *Journal of Artificial General Intelligence*, 5(1), 1–48.
- Hareli, S., Kafetsios, K., & Hess, U. (2015). A crosscultural study on emotion expression and the learning of social norms. *Frontiers in Psychology*, 6, 152022.
- Healy, P. (2022). Social robots as partners? *AI & SOCIETY*, 39, 1–8.
- Hung, L., Liu, C., Woldum, E., Au-Yeung, A., Berndt, A., Wallsworth, C., et al. (2019). The benefits of and barriers to using a social robot PARO in care settings: A scoping review. *BMC Geriatrics*, 19, 1–10.
- Kagan, B. J., Gyngell, C., Lysaght, T., Cole, V. M., Sawai, T., & Savulescu, J. (2023). The technology, opportunities and challenges of synthetic biological intelligence. *Biotechnology Advances*, 68, 108233.
- Kagan, B. J., Kitchen, A. C., Tran, N. T., Habibollahi, F., Khajehnejad, M., Parker, B. J., et al. (2022). In vitro neurons learn and exhibit sentience when embodied in a simulated game-world. *Neuron*, 110(23), 3952–3969. e3958.
- Kambur, E. (2018). Does emotional intelligence affect quality of service?: An application in the service sector. *The Journal of Academic Social Science*, 6(71), 587–600.
- Khashman, A. (2008). A modified backpropagation learning algorithm with added emotional coefficients. *IEEE Transactions on Neural Networks*, 19(11), 1896–1909.
- Krakovsky, M. (2018). Artificial (emotional) intelligence. *Communications of the ACM*, 61(4), 18–19.
- LeDoux, J. E. (1998). The emotional brain: The mysterious underpinnings of emotional life. Simon and Schuster.
- Leite, I., Martinho, C., & Paiva, A. (2013). Social robots for long-term interaction: A survey. *International Journal of Social Robotics*, 5, 291–308.
- Lerner, J. S., Li, Y., Valdesolo, P., & Kassam, K. S. (2015). Emotion and decision making. *Annual Review of Psychology*, 66, 799–823.
- Levens, S. M., & Phelps, E. A. (2008). Emotion processing effects on interference resolution in working memory. *Emotion*, 8(2), 267.
- Levine, E. E., & Schweitzer, M. E. (2015). Prosocial lies: When deception breeds trust. *Organizational Behavior and Human Decision Processes*, 126, 88–106.
- Marcos-Pablos, S., & García-Peñalvo, F. J. (2022). Emotional intelligence in robotics: A scoping review. In *New trends in disruptive technologies, tech ethics and Artificial Intelligence: The DITTET collection* (Vol. 1).
- McLean, S., Read, G. J., Thompson, J., Baber, C., Stanton, N. A., & Salmon, P. M. (2023). The risks associated with artificial general intelligence: A systematic review. *Journal of Experimental & Theoretical Artificial Intelligence*, 35(5), 649–663.
- Moerland, T. M., Broekens, J., & Jonker, C. M. (2018). Emotion in reinforcement learning agents and robots: A survey. *Machine Learning*, 107, 443–480.
- Monteith, S., Glenn, T., Geddes, J., Whybrow, P. C., & Bauer, M. (2022). Commercial use of emotion artificial intelligence (AI): Implications for psychiatry. *Current Psychiatry Reports*, 24(3), 203–211.
- Neel, R., Becker, D. V., Neuberg, S. L., & Kenrick, D. T. (2012). Who expressed what emotion? Men grab anger, women grab happiness. *Journal of Experimental Social Psychology*, 48(2), 583–586.
- Noone, G. (2024). OpenAI launches GPT-4o, flaunting ability of model to detect user emotions. Tech Monitor. Retrieved June 14, 2024, from [https:// techmonitor.ai/technology/ai-and-automation/ openai-launches-gpt-4o](https://techmonitor.ai/technology/ai-and-automation/openai-launches-gpt-4o)

- OpenAI. (2024). *Hello GPT-4o*. Retrieved June 14, 2024, from <https://openai.com/index/hello-gpt-4o/>
- Pai, R. Y., Shetty, A., Dinesh, T. K., Shetty, A. D., & Pillai, N. (2024). Effectiveness of social robots as a tutoring and learning companion: A bibliometric analysis. *Cogent Business & Management*, 11(1), 2299075.
- Parisi, D., & Petrosino, G. (2010). Robots that have emotions. *Adaptive Behavior*, 18(6), 453–469.
- Pessoa, L., & Ungerleider, L. G. (2004). Neuroimaging studies of attention and the processing of emotionladen stimuli. *Progress in Brain Research*, 144, 171–182.
- Picard, R. W. (2000). *Affective computing*. MIT Press.
- Picard, R. W. (2003). What does it mean for a computer to “have” emotions. In *Emotions in humans and artifacts* (pp. 213–235).
- Rosenbloom, P. S., Gratch, J., & Ustun, V. (2015). Towards emotion in sigma: from appraisal to attention. Artificial General Intelligence: 8th International Conference, AGI 2015, AGI 2015, Berlin, Germany, July 22–25, 2015, Proceedings 8.
- Sætra, H. S. (2022). Loving robots changing love: Towards a practical deficiency-love. *Journal of Future Robot Life*, 3(2), 109–127.
- Scarantino, A., & De Sousa, R. (2018). Emotion. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*.
- Scheutz, M. (2009). The inherent dangers of unidirectional emotional bonds between humans and social robots. In P. Lin, K. Abney, & G. A. Bekey (Eds.), *Robot ethics: The ethical and social implications of robotics*. MIT Press.
- Schuller, D., & Schuller, B. W. (2018). The age of artificial emotional intelligence. *Computer*, 51(9), 38–46.
- Schutte, N. S., Malouff, J. M., Thorsteinsson, E. B., Bhullar, N., & Rooke, S. E. (2007). A meta-analytic investigation of the relationship between emotional intelligence and health. *Personality and Individual Differences*, 42(6), 921–933.
- Sica, A. (2023). The robot will feel you now: The ethics of artificial emotional Intelligence in sex robots. IEEE Ro-Man 2023, Busan.
- Sica, A., & Sætra, H. S. (2024). Artificial emotions and the evolving moral status of social robots. *Proceedings of the 2024 ACM/IEEE international conference on human-robot interaction*.
- Sloman, A. (2001). Beyond shallow models of emotion. *Cognitive Processing*, 2(1), 177–198.
- Smirnova, L., Caffo, B. S., Gracias, D. H., Huang, Q., Morales Pantoja, I. E., Tang, B., et al. (2023). Organoid intelligence (OI): The new frontier in biocomputing and intelligence-in-a-dish. *Frontiers in Science*, 1, 1017235.
- Stark, L., & Hoey, J. (2021). The ethics of emotion in artificial intelligence systems conference on fairness, accountability, and transparency, New York, NY, USA.
- Strömfelt, H., Zhang, Y., & Schuller, B. W. (2017). Emotion-augmented Machine Learning: Overview of an emerging domain Affective Computing and Intelligent Interaction (ACII 17).
- Turing, A. M. (1950). I.—Computing machinery and intelligence. *Mind*, LIX(236), 433–460. <https://doi.org/10.1093/mind/LIX.236.433>
- Turkle, S. (2011). *Alone together: Why we expect more from technology and less from each other*. Basic Books.

- Wada, K., & Shibata, T. (2006). Robot therapy in a care house-its sociopsychological and physiological effects on the residents. *Proceedings 2006 IEEE International Conference on Robotics and Automation*, 2006. ICRA 2006.
- Wada, K., & Shibata, T. (2007). Living with seal robots— Its sociopsychological and physiological influences on the elderly at a care house. *IEEE Transactions on Robotics*, 23(5), 972–980.
- Weiss, A., & Hannibal, G. (2018). What makes people accept or reject companion robots? A research agenda. *Proceedings of the 11th Pervasive technologies related to assistive environments conference*.
- Weng, Y., & Lin, F. (2022). Multimodal emotion recognition algorithm for artificial intelligence information system. *Wireless Communications and Mobile Computing*, 2022, 1–9.
- Whitby, B. (2011). 15 do you want a robot lover? The ethics of caring technologies. In *Robot ethics: The ethical and social implications of robotics* (p. 233).
- Yang, Y., Wang, Y., Yuan, X., & Yin, F. (2012). Hybrid chaos optimization algorithm with artificial emotion. *Applied Mathematics and Computation*, 218(11), 6585–6611.
- Zhu, W.-L., Fang, P., Xing, H.-L., Ma, Y., & Yao, M.-L. (2020). Not only top-down: The dual-processing of gender-emotion stereotypes. *Frontiers in Psychology*, 11, 526960.

Article 2

In technology we trust! But should we?

1. Introduction

In a society where we increasingly rely on algorithmic decision systems for a wide and growing range of activities, questions regarding whether we can and should trust these technologies have emerged. To what degree can we trust machines? And can we answer this question in general, or must we consider what sort of machines we encounter? The issue of trust in technology is particularly important given the direct impact this might have on the acceptance and use of the technology at stake (Freedy et al., 2007; Glikson & Woolley, 2020; Lankton et al., 2015). Furthermore, it has been argued that technology has the potential to interfere with trust, a central value in human society (Sætra, 2021b). A discussion which establishes whether human-technology interactions involve trust is therefore crucial for understanding the potential implications, such as moral value changes (Danaher & Saetra, 2022), which such relationships with technology entail.

Broadly speaking, trust in technology is defined as the willingness to depend on a given technology, in a specific context, where risk is involved, and potential negative consequences are possible (McKnight et al., 2009). However, we claim that a more specific definition of trust, and a clarification of how this concept is applicable in the context of technology, is needed for an adequate assessment of the questions mentioned above. This article explores the concept of trust and its use in both the fields of traditional social sciences and human-technology interaction. We argue that to preserve trust as a meaningful concept, it is beneficial to develop an interdisciplinary approach to the trust, enabling research on the topic to become more integrated and comparable across different fields of study. Such an assessment of trust and the related factors determining its establishment has significant consequences on the analysis of how people can and should develop trust in technologies.

We argue that if trust is a mere rational assessment of likelihood that another entity will act as we want it to, trust and *reliance* is conflated. These concepts can, however, be argued to

differ significantly, and should be distinguished from each other. Furthermore, when trust is distinguished from reliance, we advance that it *does* make a difference what sort of machine are analysed. In particular, we stress the role of emotions, while highlighting how this brings us to a new dilemma as anthropomorphism and the discussion on whether machines *can* have emotions will generate disagreement regarding whether, for example, social robots with artificial emotional intelligence (AEI) could be the recipients' trust proper.

In light of this, we choose AEI social robots as a specific case study of trust in technology. Evidence suggests that a feeling of trust towards these systems is formed more easily and spontaneously than towards other technologies. Given their characterising and specific features, these systems can express emotional, and human-like behaviours and expressions which potentially provoke emotional attachment and form attitudes of affective trust in the user. We consider whether such characteristics are relevant to justify the claim that we trust these systems. If our analysis leads to the conclusion that AEI enabled social robots' capabilities are not sufficient for a relational account of trust to be meaningfully involved in human-robot interactions, then we could argue that the same applies for technology in general.

This article is structured as follows. In Sect. 2, we present a conceptual analysis of the notion of trust and introduce two terms to emphasise the distinction between reliability and trust: namely, rationality-trust on one side, and affective trust on the other. In Sect. 3, we explore how trust is and can be used in the context of technology, questioning the possibility of affective trust in technology. Section 4 introduces AEI and social robotic technologies and considers whether the impact of their emotional and human-like capabilities on affective trust. Namely, could we claim that the AEI social robot's capabilities are enough not only to facilitate, but also to warrant trust?

Finally, Sect. 5 addresses this latter question. Despite the emotional attachment that may be formed with these systems, we argue that the synthetic nature of AEI social robots is morally relevant for the assessment of whether people should place trust in them. Namely, AEI social robots are tools designed for purposes potentially different or in conflict with the ones of their users. Although people may develop a feeling of trust towards these robots, granting trust proper in AEI social robots, and technology in general, is not morally warranted.

2. What is trust?

Before diving into the specific issues related to trust and technology, we explore what lies in the concept of trust more broadly. How is trust used to understand the relationships between humans, what components does it consist of, and what separates it from, for example, reliability? In this section, we present different understandings of trust, before advancing our suggested use of the term.

Some authors define trust as a cognitive construct based on a rational evaluation of the trustee's ability to perform a particular job effectively, and situational features (Coleman, 1990; Ferrin et al., 2007; Glikson & Woolley, 2020; Kim et al., 2006). Nickel, Franssen and Kroes (2010) presents one definition of trust derived from rational-choice theory and Coleman (1990), in which trust is conceptualised as:

a judgement one makes when, having considered the possible costs and benefits of relying on another person's performance, as well as the salient alternatives, one comes to the conclusion that it is worth it to rely on that performance.

However, and as the authors themselves note, such an account of trust does not differ fundamentally from mere rational reliance (Nickel et al., 2010). While they include "person" in the definition, such forms of rationality-trust could arguably also apply to our evaluation of, for example, our toaster or coffee maker. The emotional aspects of the relationship with the trustee are of secondary importance or entirely omitted from trust-considerations.

Others present a different, and arguably more comprehensive, conception of trust with respect to interpersonal relationships (Aroyo et al., 2021; Levine & Schweitzer, 2015; Rousseau et al., 1998; Sætra, 2021a). Rousseau et al. (1998), for example, define trust as.

a psychological state comprising the intention to accept vulnerability based upon positive expectations of the intentions or behaviour of another.

This conceptualization of trust is based on both rational and emotional elements. As Rossi et al. (2020) argue, interpersonal trust is developed on the basis of rational factors which allow us "to access the reliability of actions and behaviours", and by affective factors which establish emotional bonds and increase "the expectation in which the involved agents invest in a common goal". Along the same line, Koyama (2016) emphasises the role of emotions in trust

relationships by defining trust as an “affective attitude” which has the expectation that the other’s “goodwill and competence [...] will extend to cover the domain of our interaction with” them, and that the trustee will be “directly and favorably moved by the thought that we are counting on” them. Accordingly, the emotional relationship among the trustor and the trustee “does not guarantee but provide a reason why the trustee’s actions will be performed” (Koyama, 2016).

Similarly, Glikson and Woolley (2020), following McAllister’s (1995) analysis, address emotion-driven or affect-based trust in interpersonal relationships. They suggest that the emotional ties connecting individuals can provide a partial basis for trust based instead on an impression of the trustee’s reliance and competence, and this can be seen as complementary to rational trust.

According to these discussions, researchers have individuated specific rational and emotional antecedents which help to build the foundation for the development of interpersonal trust; namely, characteristics of a trustee conducive to the generation of trust in the trustor. These are benevolence, integrity, and ability (Glikson & Woolley, 2020; Lankton et al., 2015; Levine & Schweitzer, 2015; Mayer et al., 1995). Benevolence is the degree to which an individual “is believed to want to do good to the trustor, aside from an egocentric profit motive”, suggesting that the trustee has some particular attachment to the trustor (Mayer et al., 1995). Integrity corresponds to an individual’s ethicality and reputation for honesty, and the belief that she adheres to a set of principles that the trustor finds adequate (Mayer et al., 1995). Finally, ability relates to an individual’s technical skills, competence, and expertise in a specific context (Mayer et al., 1995), which supports the belief that the trustee *can* – in the meaning “is able to” – do what the trustor needs (McKnight et al., 2009).

It is key to notice that what is referred to as ability-based trust is closely linked to the rational evaluation of expected performance in a specific situation, which takes us back to reliance (see Table 1). We might, for example, ask whether a person that expresses both benevolent intent and honesty can still be generally considered trustworthy even if they lack the ability to perform what the trustor expects from them. As Mayer, Davis and Schoorman (1995) argue, the trustee’s ability is relative to the required task to perform. The trustee might be “highly competent in some technical area, affording that person trust on tasks related to that area”, however she might have “little aptitude, training, or experience in another area” (Mayer et

al., 1995). Therefore, ability-based trust is domain specific. Benevolence and integrity, however, refer to the nature of the trustee more generally. We might expect, therefore, that in a situation where the trustee lacks the required skills and techniques for carrying out a specific task, she does not necessarily become untrustworthy in the eye of the trustor¹. This latter might maintain the belief that she would act as expected if she had the necessary abilities, given that she possesses both the antecedents of benevolence and integrity. Rather, by lacking ability, she would be considered unreliable for that specific task. Considering this, it is essential to maintain a conceptual difference between reliability and trust based solely on benevolence and integrity, which we refer to as affective trust. As we will discuss in the following, affective trust is the most challenging in the context of in technology, as benevolence and integrity refer to qualities that many machines arguably do not, or even cannot, have. For this reason, in the remaining discussion, this article will focus specifically on affective trust.

Furthermore, we argue in favour of a universal definition of trust that is applicable across academic disciplines, and different types of human-x-relationships. By using trust in the same way when discussing, for example, human-human and human-robot relationships, we can avoid situations in which the same term is used to refer to a very different set of phenomena. Firstly, this allows us to distinguish between trust and reliance, and consequently allows for more conceptual clarity and helps preserve the usefulness of both terms. Secondly, such usage of the term is arguably more in line with everyday understandings of the concept, and it better promotes inter-disciplinary research, as both human-human interaction scholars and HRI scholars will to a larger degree refer to a similar phenomenon.

Trust-dimension	Refers to the trustee's...	Universality	Relies on	Relates to
Ability	...capacity or capability for performing particular actions	Domain-specific	Capabilities	Reliability

¹ Trustworthiness and trust are two interconnected however categorically distinct concepts (Nickel et al., 2010). Hardin (2006) argues that there is a casual connection between the two, as “trustworthiness begets trust”. Whereas trust is usually considered an attitude, trustworthiness is a property of the trustee which satisfies and makes appropriate the attitude of trusting (Nickel et al., 2010).

Benevolence	...desire to do good to others, including the trustor	Universal	“Goodness”	Goodness
Integrity	...honour, likelihood of abiding by their promises and proclivity to act in accordance with accepted principles	Partly universal, partly domain-specific	Intentions and principles	Principledness

Table 1 - Ability, benevolence, and integrity-based trust.

3. Trust in technology

Moving on from the general concept of trust, we explore how trust is and can be used when discussing human interaction with technology.

Lee and See’s (2004) seminal article on trust in automation is often used in research on trust in various forms of technology, and they define trust as: “the attitude that an agent will help achieve an individual’s goals in a situation characterized by uncertainty and vulnerability.” Others working on trust in HRI often explicitly build on this definition, such as Wagner and Arkin (2011), who define trust as “a belief, held by the trustor, that the trustee will act in a manner that mitigates the trustor’s risk in a situation in which the trustor has put its outcomes at risk.” The concept of trust is routinely used in settings in which humans interact with non-human entities, but we see from these examples that the conceptualization of trust is not necessarily integrated with the more comprehensive conceptualizations referred to in the previous section. Extant research, especially from the discipline of philosophy, expresses some reservations about “trust in technology”, by defending the view that the notion of affective trust is misguided and unsuitable for ethical considerations in such a context (Friedman et al., 2000). The trustee’s characteristics for trust to be formed include “possessing moral integrity; being capable of fulfilling one’s moral obligation; actually caring about the interests of the trustor; or, recognizing and responding positively to the very fact that one is being relied upon” (Nickel et al., 2010). Along the same lines, Friedman et al. (Friedman et al., 2000) argue that

entities involved in a trust relationship must possess agency and consciousness, which enable them to express good will, extend benevolence towards others, feel vulnerable, and feel betrayed when their trust is misplaced. Such consideration could also help explain why HRI researchers seem to be referring to a somewhat different foundational concept when using the term “trust” than what is often the case in interpersonal relationships.

When discussing agency, consciousness, capability for moral agency, etc., we are talking about abilities in the form of *properties*, and these are very different from evaluations of benevolence and integrity, which arguably only make sense if the potential trustee has such required properties. We thus propose a tentative distinction in which property-based trust is considered foundational, whereas the two latter are second-level concerns only applicable when the first level is satisfied, as shown in Fig. 1. The reason we use properties instead of abilities is that the original use of ability-based trust (Mayer et al., 1995) mainly referred to

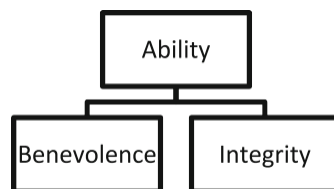


Figure 1 - First and second-level antecedents for trust.

abilities relevant for reliance-based considerations of performance.

Some might argue that such properties are only possessed by humans and conclude that any inanimate object, including technological artifacts, can never be benevolent and honest, and therefore genuine object of trust but of mere reliability (Nickel et al., 2010; Sweeney, 2021). However, as artificial intelligence (AI) and robotics keep developing, this is increasingly being disputed. Indeed, recent development in social robotics has allowed designers to develop a kind of technology which can act like humans or living entities and establish interactions with the user on an emotional level. Research has shown that the distinguishing element of human-likeness possessed by a given technology can influence the kind of antecedents of trust developed, and consequently the conception of trust which best applies to the situation (Lankton et al., 2015). On the one hand, systems that show low human-likeness are more likely to elicit rationality-trust based on reliability, functionality, and helpfulness. On the other hand, when the technology shows high levels of human-likeness, the user might inevitably invoke constructs such as the above-mentioned integrity and benevolence to assess whether the given technology can be trusted, suggesting the possibility of experiencing human-like trust.

Designers of social robots specifically aim to build systems which elicit a feeling of trust in the user, as trust is a fundamental key factor for enabling successful cooperation and maintaining effective relationships between people and robots over time (Billings et al., 2012; Broadbent, 2017; Gompei & Umemuro, 2018; Hancock et al., 2011; Lum, 2020; Prescott & Robillard, 2021; Rossi et al., 2020) . By providing not only physical help, but also emotional support, it is beneficial for social robots to be perceived as safe and trustworthy entities (Song & Luximon, 2020). Trust directly affects the willingness of people to accept information provided by robots, follow robots' suggestions, and thus benefit from the advantages inherent in robotic systems (Freedy et al., 2007; Glikson & Woolley, 2020; Hancock et al., 2011). This is because, as Song and Luximon (2020) affirm and we have previously mentioned, trust in social contexts has a significant impact on persuasion and could directly influence people's intention to follow others' suggestions.

Whereas the robot's reliability certainly plays a crucial role, developers also specifically target human emotions in the design of social robots for trust to be formed (Glikson & Woolley, 2020). As we will argue in the next sections, developing trust attitudes through emotional attachment towards the robot has the potential to allow users to access increased societal benefits from their engagement with robots (Sweeney, 2021). People's trust in robots is likely to be affected by their perception of the robot's capabilities which might depend on human-related and robot-related factors. For example, people's trust might be affected by the robot's appearance, its level of autonomy, and functionalities. It may also be influenced by user's personalities, self-confidence, and prior experience with robots. However, as we will explain below, trust-generation based on perceived benevolence and integrity, and perceived agency in machines, largely result from non-rational and non-conscious mechanisms. This helps explain why humans might self-report trust in a robot, for example, even if these could be objectively argued to be incapable of benevolence and integrity.

To sum up, authors generally agree that technology can be reliable. The possibility for the establishment of affective trust, however, is more controversial and often rejected. Nonetheless, some empirical research has proved that possessing human-like features for a specific technology marks a concrete difference in how the user perceives it and develops trust attitudes accordingly. This is shown both in self-reported trust and in human behaviour (Robinette et al., 2017), but we have also seen that HRI conceptualisations of trust might not

include all parts of the comprehensive concept and could be argued to mainly capture reliance. In light of this, social robots might represent an unprecedented and interesting case for discussion about trust in technology. This brings us to the analysis of the specific case of AEI social robots, which are presented in the following section.

4. Social robots with artificial emotional intelligence

Social robots' specific physical and functional characteristics – their human-likeness – allows them to have the potential to make humans form a feeling of trust towards them. As argued in Plaks et al.'s (2022) article, people are more inclined to place trust in “others who are similar to themselves on key dimensions”. Social robots are consequently the most likely technology-candidates for generating trust-like experiences in humans and potentially forming trust relationships between humans and machines. Hereafter, we describe the technology of social robots and AEI. In the following sections, these systems will be investigated in more detail to identify their specific, characterising factors which distinguish them from other kinds of technologies, and which possibly contribute to develop affective trust in them.

A social robot is a “physically embodied, autonomous agent that communicates and interacts with humans on an emotional level”, by following social behaviour patterns, modelling different human-like “states of mind”, and adopting what they learn through their interactions (Darling, 2012). Social robots often possess humanoid or animaloid physical configurations. As we will discuss later, their appearance has implications for the effectiveness of HRI, which is based not only on visual communication but also on visual and tactile perception (Campa, 2016). The goal of social robotics is the creation of artificial agents capable of social performances and considered “above the status of instruments to that of interlocutors” in the eye of the user (Damiano & Dumouchel, 2018a). In other words, humans prefer to interact with such machines in the same way they interact with other people or other social entities. Differently from inanimate computers, as well as from industrial or service robots, social robots are *designed* to elicit human feelings and mimic social cues. This implies that what we refer to as social robots are not *accidentally* but *intentionally* social, and this excludes a wide range of robots that could potentially trigger social responses in humans despite this not being intended by developers and designers (see, e.g., Scheutz, 2009, for an analysis of user's attachment to their robotic vacuum cleaners, called Roomba).

Emotions have been increasingly used in social robot design, with the aim of making AI more effective and therefore facilitating the establishment of more believable, acceptable, and effective HRI (Fong et al., 2003). The capacity of social robots to cause the users to become attached to them (LaGrandeur, 2015) is expected to be intensified significantly with the implementation of AEI. Schuller and Schuller (2018) define AEI a technology that possesses the ability to “recognize emotions (both of others and oneself), generate and adapt emotions, and apply emotional information in goal accomplishment and problem solving”. This new field of studies within AI research finds its greatest potential in developing computers “to be genuinely intelligent and to interact naturally with us” (Picard, 1997). As research has shown, emotional and volitional factors play a significant role in human behaviour, communication, and interaction, and are essential for the effective functioning of human intellect and thinking processes (Damasio, 1994; Lerner et al., 2015). Specifically, emotional intelligence consists in the individual’s ability to understand, manage, and adjust accordingly to one’s own emotions, as well as to correctly interpret others’ emotions and respond to their behaviours (Goleman, 1995). As a result, interactions with social robots artificially possessing emotional intelligence could potentially, easily transform into a much more intimate and affectionate relationship, consequently exacerbating the possibility of a user establishing an emotional bond and form trust attitudes towards the robot.

Social robots can be differentiated according to the degree of human-likeness and the level of AEI they possess, and therefore to which extent they are perceived as being sociable, warm, and personal (Lankton et al., 2015). Also, depending on the social robot’s specific task, and the duration and frequency of the interaction, the user’s emotional involvement might largely differ. For instance, social robots employed for teaching support necessitate to implement social cues and skills, as well as to appeal to the user’s emotional state, to operate effectively. However, given that their use is limited to a few hours per week and mainly in public settings such as school, we might expect a minimum level of emotional engagement and even lower emotional attachment from the user’s perspective. On the contrary, social robots intended for more personal or even intimate services, such as companionship, daily healthcare, or sex, are usually employed for a considerably prolonged, daily use, and are expected to form an affective relationship with the user. Indeed, their services work most effectively “when they are perceived as social agents, rather than tools” (Darling, 2015). This latter category is

especially relevant for our analysis as it allows us to examine interpersonal trust more specifically. Indeed, the intimate, emotional relationship that this particular kind of social robot aims to establish with the user might possess the required features for human-like trust to be formed.

To summarise, as empirical studies suggest, AI social robotic systems equipped with emotional intelligence further enhance features and attitudes which resemble human ones, and the consequent possibility of trust development. As discussed above, emotions can significantly impact the perception of the antecedents of trust and consequently influence the kind of trust developed in a relationship (Schoorman et al., 2007). As a result, since AEI social robots are intended to interact with humans affectively, it might be assumed that their design implies humans to develop affective trust. Hence, this case is particularly conducive for our research question, as it allows us to suggest that, if trust does or should not apply in the context of social robots, then it would be even harder to consider the possibility of establishing or experiencing trust towards other kinds of technology. The next section presents more accurately some of the AEI social robot's design characteristics which might strongly impact the user's affective trust attitudes.

4.1. How AEI social robots' features impact affective trust

In this section, we describe some of the specific features of AEI social robots which contribute to develop the user's affective trust. There are anthropomorphism, deception, and emotional attachment². These factors are highlighted because they are tightly linked with the comprehensive concept of trust established above, and because evidence suggests that they can cause experiences of trust-like feelings in humans (Beer et al., 2017; Broadbent, 2017; Glikson & Woolley, 2020; Rossi et al., 2020; Torta et al., 2014). Note also that not only social robots with AEI possess such features and can therefore appear in a personal and social way from the user's point of view. However, as it will also be argued in the following, the combination and strength of such features in AEI social robots allow these systems to be considered as different from other kinds, as a large and growing body of literature suggests

² It is important to mention that the social robotics' features here presented raise ethical concerns beyond the topic of trust. Authors such as Turkle (2011) and Scheutz (2009) have expressed their concern about the anthropomorphisation of robotic technology, their deceptive potential, as well as the kind of relationship and unilateral emotional attachment that people would develop with them. However, these moral issues are not the focus of this article, and they will not therefore be addressed unless related to the issue of trust.

(see, e.g., Breazeal, 2004; Coeckelbergh, 2010; Darling, 2012; Darling, 2015; Gunkel, 2018; Hung et al., 2019; Sung et al., 2007; Turkle, 2011).

Anthropomorphism. Anthropomorphism is generally defined as people's spontaneous tendency to attribute human properties and characteristics to non-human entities (Epley et al., 2007; Złotowski et al., 2016), but also to interpret non-human behavior as motivated by humanlike emotional and mental states (Airenti, 2015). These non-human entities include anything "that acts with apparent independence"³, such as non-human animals, natural forces, religious deities, and mechanical or electronic devices (Epley et al., 2007). As research shows, artificial entities such as computers and virtual characters (Darling, 2015), or even robot vacuum cleaners (Scheutz, 2009; Sung et al., 2007) are perceived as social actors. The broader phenomenon might apply to an even broader range of artefacts, however, as similar mechanisms were described by Reeves and Nass (1996) in relation to computers and television.

Compared to other technologies, robots tend to amplify dramatically the perception of possessing human qualities such as feelings and intentions because of their specific characteristics (Darling, 2015). Specifically, anthropomorphism could be driven by interface features (e.g., human-like form of the robot), behavioural features (e.g., gaze, node), and intentional framing (e.g., giving the robot a human name) (Glikson & Woolley, 2020). Social robot design is especially intended to be anthropomorphized (Breazeal, 2004; Healy, 2022), as it can facilitate social exchange between the robot and the user, enhance familiarity, and promote interactions and integration (Damiano & Dumouchel, 2018a). Furthermore, the human-likeness of a given technology potentially affect users' emotional reactions towards it (Hoff & Bashir, 2015). For these purposes, the goal for social robotic designers is to emulate humans "to the greatest extent possible" (De Visser et al., 2016). AEI plays a fundamental role here, as it enhances the tendency to anthropomorphise by depicting the robot as an entity able to *feel*, and consequently allows to perceive an attenuation of the ontological differences between humans and social robots.

As some argue (Aroyo et al., 2021; De Visser et al., 2016), the level of human-likeness of a system produces different effects on trust. Some empirical studies have shown that

³ Interestingly, Tallant's (2019) although experiment presents cases where humans are propense to anthropomorphise objects which do not appear to show any kind of agency.

participants exhibited more trust, compliance, and enjoyment in the interaction with the robot than the computer (Kidd & Breazeal, 2008). With regard to social robotics, the more sophisticated a robots' anthropomorphic features are, the more the robot will be perceived as a subject rather than an object, and the higher the possibility for the user to form human-like trust. Furthermore, the closer the robot's appearance to humans, the easier it is for the user to trust the robot in ways which resemble affective trust (Aroyo et al., 2021).

Deception. Deception is defined as an act or statement that intentionally misleads or promotes a belief that is not true (see, e.g., Boles et al., 2000; Gino & Shea, 2012; Levine & Schweitzer, 2015). Deceiving acts may be implemented through gestures, disguise, by means of action or inaction, or through silence (Bok, 1978).

Social robots, compared to other social technologies, show a more noticeable, stronger, and more prominent deceptive potential (Danaher, 2020a; Sætra, 2021a). The main goal of deceiving robots is to "smoothly participate in standard human modes of communication and social interaction" (Isaac & Bridewell, 2017); i.e., to facilitate and make HRI more humanlike. This would also ease their integration into society and protect people from biases about the infallibility of such systems (Isaac & Bridewell, 2017). Deception can be achieved through different features of the robot, and on different levels. Anthropomorphic expressiveness (e.g., sad eyes, smiling), for instance, encourages the human to believe that the robot experiences emotions, hopes, and desires similar to what a human feels (Aroyo et al., 2021; Matthias, 2015). Emotional deception, which occurs when the user believes that the robot possesses real emotions (Sharkey & Sharkey, 2012), is an ability peculiarly present in AEI social robots. Such a form of deception raises unrealistic expectations regarding the robot's nature and capabilities (e.g., being benevolent or having motivations to act) and alter the possible kind of trust relationship that could be formed accordingly. Indeed, as previously mentioned, attributing certain intentions to the trusted is neglectable to form judgements of reliance, but essential for genuine affective trust (Nickel et al., 2010). With AEI social robots, the user might be more easily deceived to believe that the robot deliberately acts for her own good as she perceives the system as possessing benign intentions and feelings.

Since deceiving behaviours are generally believed to be implemented for personal gain and self-interest, some literature considers deception as particularly damaging to relationships and affective trust, sometimes in an irrevocable and enduring way (e.g., Bok, 1978; Boles et

al., 2000). In contrast, some authors argue that when deception or lying is done for the benefit of the other, such behaviours are not only considered acceptable but even perceived more ethical than telling the truth (Levine & Schweitzer, 2014, 2015). Isaac and Bridewell (2017) argue that ethical lies are possible because the morality of deceptive communication must be assessed in light of its underlying motive, rather than the truth or falsity of deception. The action of misleading with the purpose to do good to the deceived is a type of deception which Levine and Schweitzer (2015) call prosocial deception. Lies and deceptive behaviours, in such a case, are beneficial, prosocial, and have the effect of increasing trust (Isaac & Bridewell, 2017; Sætra, 2021a). In conclusion, deception per se does not undermine the positive interferences associated with prosocial actions (Levine & Schweitzer, 2015).

Emotional Attachment. One of the consequences of social robots' anthropomorphism and emotional deception, and the social engagement that necessarily arises from their use, is the possibility for the user to become emotionally attached to the robot (Sharkey & Sharkey, 2020; Sullins, 2012; Van Maris et al., 2020). Such characteristics are not exclusive requirements for an emotional bond to be formed. Indeed, it has been shown that people become attached to objects or other kinds of technologies even when these lack any social dimension (Carpenter, 2016; Scheutz, 2009; Sung et al., 2007). However, social robotics' anthropomorphic and deceptive features further and deeply encourage the elicitation of affection in humans. For instance, both Breazeal's (2004) studies and Riek's et al. (2009) experiment measured how anthropomorphism affects humans' empathy towards robots. Respectively, they show that when anthropomorphic robots mirror their human interlocutor's facial expressions and body movements, the user is encouraged to develop empathy towards them; and that people empathise more strongly with robots possessing human-like features than with mechanical-looking robots.

Another interesting point comes from Sweeney's (2021) analysis, where she recognises that sometimes we become emotionally invested towards inanimate objects that have a sentimental value such as, for instance, a wedding ring. However, she claims that what distinguishes the emotions developed towards such objects from those towards social robots is the feeling of empathy, elicited exactly by the features described previously. In her words: "We can have emotional attachment to wedding rings, beloved bikes and plants, and we can feel sadness when they are managed, but we do not feel empathy towards them. [I]t is

empathy that marks the difference in the way that we react to social robots from other objects” (Sweeney, 2021).

Also, according to their specific tasks and purposes, some robots are especially designed to elicit attachment in the user, with the aim to facilitate the use of robots as human companions and personal helpers (LaGrandeur, 2015; Van Maris et al., 2020). As described above, AEI social robots particularly have the capability of communicating and engaging affectively, and being sensitive and responsive to the others’ emotions. This consequently enhances the possibility to develop a form of affection towards them (Sharkey et al., 2017). LOVOT, for example, is a social robot specifically designed for companionship, developed to increase the user’s general wellbeing. As the advertisement says: “When you touch your LOVOT, embrace it, even just watch it, [...] It’s a little like feeling love toward another person. [...] LOVOT was born for just one reason – to be loved by you” (LOVOT). LOVOT recognises its surrounding and adapts its actions in accordance with the owner’s engagement with them; e.g., it looks back, laughs when tickled, and expresses joy when caressed. Similarly, another well-known commercialised companion type robot is Paro. Shaped like a cute baby harp seal, it moves and makes seal noises in response to human touch, temperature, light, noise, and orientation (Broadbent, 2017). As studies have shown, Paro reduces loneliness and increases social interactions both with Paro and other people (Heerink et al., 2010).

AEI technology, therefore, could mark a crucial difference for determining the kind of trust that social robots elicit in humans. Indeed, since affective trust is related to affective bases, the emotional connection with the robot may have important effects on the kind of trust relationship possibly established. First, it could reinforce the belief that the involved agents act for a common goal (Rossi et al., 2020); i.e., that the robot actually cares for the user, and therefore acts for her benefit (Boada et al., 2021; Scheutz, 2009; Sharkey & Sharkey, 2020). This might further promote a feeling of emotional safety and confidence in the robot, which are essential components of trust in close relationships (Levine & Schweitzer, 2015). In such cases, and mainly in the early stages of HRI, prosocial deception and lies might be particularly beneficial for strengthening such relationships. Furthermore, AEI social robots used for personal and intimate services (e.g., companionships) are intended to be regarded as fellows, companions, friends, or even lovers. When it comes to affective trust, the role of emotions and attachment make a friend to be more entrusted than a machine (Heuer et al., 2018).

Why Social Robots are Perceived as Social Entities. To sum up, we have explored three main features which essentially characterise AEI social robots and distinguish them from technologies of other kinds for the intensity and the high degree of humanlikeness they can achieve. As Coeckelbergh (2012) puts it, whether something appears as a human-like entity “is a matter of degree”: by resembling humans, social robots with AEI allow to the user to perceive them not as tools, “as something in the background, or as something that has become part of ourselves (e.g., glasses), but as an other”. By possessing such a significant anthropomorphic and deceptive potential, as well as the capacity to elicit emotional attachment in the user, these systems tend to reach the threshold at which, from the user’s perspective, “objects become subjects” (Damiano & Dumouchel, 2018b). Empirical studies show that social robots are perceived as transgressing the traditional ontological boundaries, as objects that are “sort of alive” or “alive enough” (Turtle, 2011). They fall in between the terms of old traditional dichotomies of sentient/not sentient, intelligent/not intelligent, and alive/not alive (Damiano & Dumouchel, 2018a; Gaudiello et al., 2015; Kahn Jr et al., 2002; Severson & Carlson, 2010; Turtle, 2011).

The role of AEI in this context is crucial, as it tends to substantially intensify the user’s perception – correct or not – that they are interacting with someone rather than something. This ambiguity is crucial when considering the issue of trust in technology, for two main reasons. First, it suggests that there is a possibility for humans to feel a sentiment of trust towards social robots that surpasses the way we commonly “trust” objects, i.e., rationality-trust, as empirical studies have already demonstrated. Second, given these findings, it might even support the view that users are *justified* in developing affective-trust attitudes towards AEI social robots. As Sweeney (2023) puts it, when it comes to social robots, the boundaries that make affective trust inappropriate for technology become blurred. Whether such boundaries become irrelevant when discussing trust in AEI social robots, or if it is ethically required to maintain them, will be discussed in the next section. We will present two points of view. The first, supported by Tallant’s (2019) analysis, conceives robots as mere tools that can be relied upon, but not necessarily *trusted*. The second follows Coeckelbergh’s (2012) relational approach which suggests that so long as the user perceives robots as trustworthy social actors, such technologies *are* objects of trust. We present objections to both perspectives, and we then formulate our own view.

5. General discussion – the feeling of trust vs. trust proper

5.1. Should we grant trust in AEI social robots?

We have established that there is a possibility humans can experience trust towards AEI social robots, as opposed to simply relying on them. However, not everyone would be convinced that, from that, it follows that we should trust them. Despite human tendency to believe the robot as good and honest and develop affective trust accordingly, such a tendency does not provide a moral permissibility to do so. Indeed, given that they do not actually possess benevolence and integrity, some might say that a mere pretence of such antecedents is not enough, and trust is inappropriate in this context.

Such a view is sustained by Tallant (2019). He argues that placing trust in objects is the result of their anthropomorphisation which makes us erroneously attribute agency and intentionality to it. As we have discussed before, agency is the property which grounds the possibility for benevolence and integrity to be genuinely possessed. Hence, we must conclude that affective trust in AEI social robots is not morally justified, as “trust occurs because of the error in agency ascription” (Tallant, 2019).

Shall we then place trust in robots based on instrumentalist criteria, namely as mere tools? If this is our conclusion, then the same argument might also apply to other technological artifacts, as we have claimed that social robots equipped with AEI are the most likely candidates for establishing affective trust in this context. However, we do not believe that this would be the right track to follow. First, because it stands in contrast to what users spontaneously experience with social robots, as proved by empirical evidence. According to some, neglecting the instinctive approach of people towards such systems might hinder a correct and comprehensive analysis of trust in technology, as it has been argued that users’ affective responses to a given technology are significant factors in determining ethically permissible behaviours towards these objects (Darling, 2012; Nyholm, 2020). The view that users’ experiences and perceptions might provide a basis for morally justifying affective trust in social robots might be challenged. However, it is plausible to believe that by considering such experiences irrelevant and claiming that only the interactions based on reliability are permissible, it does not automatically follow that people will be prevented from developing

affective trust and therefore assure that their trusting attitudes would be expressions of mere reliance.

This brings us to the second point: namely, the combination and intensity of the AEI social robots' features described above crucially affect the perception of the user. For social robots to be treated as objects that do not generate experiences of affective-like trust, the design of social robots and AEI should be significantly altered.

For instance, authors who encouraged the view of considering social robots strictly as tools, rather than social agents, have argued against the anthropomorphic features of robots (Richards & Smart, 2016). Scheutz's (2009) proposition is to legally require a robot to emit a signal continuously to remind the user that it is only a machine: it does not possess the capability to feel or have emotions, it does not act intentionally, strictly speaking, and cannot act benevolently for the user's goals. This might reduce the robot's deceptive potential, as well as the possibility and strength of emotional bonds which users might develop towards them. As a result, the likelihood to form affective trust will decrease. However, as argued above, evidence suggests that such a strategy would not always be effective. Humans could continue to anthropomorphise these robots, project their own mentality onto them, and form what seems like deep emotional relationships, despite their limited abilities (Breazeal, 2004; Healy, 2022; Turkle, 2011).

Furthermore, as Sharkey and Sharkey (2020) suggest, a continuous signal might impact on the comfort and usefulness of having the robot around, as well as its efficacy, especially in cases where the robot is required for therapeutic reasons. Precisely this latter objection concerns any significant modification applied to the features of AEI and social robotics. As Darling (Darling, 2015) sustains, and we have previously pointed out, framing social robots in anthropomorphic, deceptive, and emotional terms is functional for the robot to achieve its purposes, especially in cases when AEI social robots are employed for personal and intimate uses.

The Robot's Appearance Suffices for Granting Trust. Hereafter, by following Coeckelbergh (Coeckelbergh, 2010, 2012), we analyse a different approach to the question whether granting trust towards AEI social robots is permissible. This view focuses on the perspective of the user,

and their perception of the social robot, rather than on the nature of the robot itself. We then present two objections.

Coeckelbergh's view (2010, 2012) is particularly interesting for our analysis as specifically focused on a concept of trust built on social relations. He claims that a genuine attribution of agency to social robots is of secondary importance. What counts is the way they appear to their users and how the users perceive some relationship with the machine. In Coeckelbergh's (2010) words:

My suggestion is that we can permit ourselves to remain agnostic about what 'really' goes on 'in' there, and focus on the 'outer', the interaction, and in particular on how this interaction is co-shaped and co-constituted by how A[rtificial] A[gent]s appear to us, humans.

It is the appearance of agency that makes these systems eligible recipients of trust in the user's eye, because a fictional but supposed agency is enough to allow them to participate in and shape our social dimension. This is also similar to Danaher's ethical behaviourism, which suggests that the internal states of the robot are as inaccessible to us as is the minds of other humans, and that we are consequently always left with relying on external signs and behaviour (Danaher, 2020b). In so far as robots are part of our social relations, we are inclined to trust them since trust is not produced but "already there, in the social". Contrary to Tallant's (2019) view, therefore, which claims that perceiving social robots as agents is a fallacy in human reasoning, Coeckelbergh values this aspect as sufficient to justify our impression of robots as trustworthy: "Appearingmaking, sometimes named 'deception', [...] is part of 'the social game' and it does not undermine trust but supports it" (Coeckelbergh, 2012). After all, it might be said that even in human-human relationships, affective trust is based on appearances. Benevolence and integrity of the other are presumed, as there is no objective certainty that this latter will behave for our own goal and benefit. As it was for the robot, also for a human trustee it is not possible to know what "really goes on in there" (Coeckelbergh, 2010), since it is perfectly plausible for the other to show nothing more than "a façade of friendliness and care" (Sweeney, 2023). Therefore, for both malevolent people or deceiving robots, the trustor will be able to appreciate the same – or significantly similar – outwardly benevolent expressions, behaviours, and language. As indicated in the second section, the vulnerability of the trustor is integrally part of an affective trust-based relationship, as our

judgement for trusting someone is largely based on how the other presents themselves to us, and whether we believe them trustworthy (Sweeney, 2023). Such vulnerability does not necessarily relate to the performance of the trustee for a specific task, but rather to the possibility of being betrayed.

To this argument, two objections can be advanced. The first relates to the comparison between human-human trust, and human-robot trust, as both based on perceptions and appearances. Differently from humans, the robot's benevolence and integrity would always be pretended (Sætra, 2020). On the contrary, the human façade is intentional and situational. Pretending, it is not the standard in a human-human trust relationship, or trust would not be possible in the first place (Sweeney, 2023).

Secondly, merely focusing on the user's perspective might make us overlook various implications, potentially negative, of granting trust towards AEI social robots, specifically because of their appearance and "always-faking" nature. Differences between trust among humans and trust in robots become, therefore, relevant again when considering the risks associated. AEI social robots are not only social actors, companions, or even friends from the user's point of view. As we will argue, they also are, and should always be seen as, synthetic tools, owned and designed by others for goals and moral values which might differ from the ones of the user (Sætra, 2021a). In the next section, we present some of the consequences of placing trust in social robots – as subjects and not objects.

5.2. Too much trust (in robots) is not a good thing

For a comprehensive assessment of the morality of trust in AEI social robots, considerations regarding the consequences of such trust are needed. As mentioned above, granting trust in robots might involve consequences possibly harmful to the user, which risk being neglected if we consider the robot's appearance as sufficient to morally justify trust. Indeed, as Aroyo et al. (2021) argue, our disposition to trust robots based on their appearance as benevolent and honest systems might lead to overtrusting these systems. Hereafter, we first describe what overtrusting technologies entails and the causes that might induce it. We then briefly explore some of the consequences of overtrust in social robots with AEI. Our analysis includes moral implications both from the perspective of the individual and in relation to human society in general.

Overtrust. Overtrust is defined as “poor calibration in which trust exceeds system capabilities” (Lee & See, 2004), leading to misuse which might expose the user to risks (Borenstein et al., 2018). The possibility to overtrust occurs not only in cases of anthropomorphic and deceiving technologies – as in the case of social robots with AEI – but also for other kinds of technological artifacts when the phenomenon known as automation bias is triggered (Aroyo et al., 2021). Automation bias occurs when people erroneously ascribe “higher authority, greater expected performance, and higher initial trust” to technologies (De Visser et al., 2016), with a consequent reduction of scrutiny and monitoring of the functioning of the system (Aroyo et al., 2021). With respect to automation, therefore, a key component linked to overtrust is the perception of the system as very reliable even when they might not be (Wagner & Nayyar, 2017).

In normal circumstances, the initial trust towards a given technology tends to decrease when erroneous functioning occurs (Hoff & Bashir, 2015). In cases of overtrust, instead, malfunctioning is not perceived per se and do not lead to a recalibration of the user’s positive expectations of the system’s potential (Aroyo et al., 2021). However, even in situations where there is no clear evidence of malfunctioning or defective behaviour, the user might be inclined to overestimate the capacities of the technology.

As for robotics, overtrust leads to a situation where “a person accepts risk because the robot appears to, or is expected to, perform a function it cannot” (Aroyo et al., 2021). More so than with other forms of automation, we have seen that in HRI the user tends to create an illusion of agency and intentionality in the robot’s actions. The presence of affection in HRI, as well as anthropomorphic and deceptive features of the robot, might mislead and biased users’ understanding of the robot’s actual abilities, and induce the user to overtrust the system. Overtrust in autonomous systems, therefore, seems to relate more closely to affective trust than trust in automation (Wagner & Nayyar, 2017).

Privacy, Manipulation, and Surveillance. Placing trust in robots can have adverse implications. Among the most discussed, there is the concern of the protection of private information, as robots can collect personal identifying data that could be accessed by other people. Social robots with AEI raise more serious worries about privacy issues than other technologies, as they can “enter into previously untapped areas in personal households and take on social functions” (Darling, 2015). Given their emotional character and social

interactivity, these systems have access to everyday life of users and more sensitive data, such as their emotional and mental states (Lutz et al., 2019), at a scale previously unimaginable. When humans develop emotional connections and affective trust in robots, “secrets can be revealed”, Lutz et al. (2019) warn. Trust and attachment to the robot might indeed be used to influence or manipulate the user to spontaneously disclose private or sensitive information (Lee et al., 2011). Interestingly, research has shown that, when disclosing personal information, users are inclined to be more honest and open with AI than with other humans (Glikson & Woolley, 2020).

Related to trust and personal data, two other concerns must be pointed out. First, there is the concern that the data collected about the user might be exploited for manipulation. Research on human-computer interaction has proven that humans are prone to being manipulated by social AI (Darling, 2015). As Scheutz (2009) argues, “it will be only a matter of time before robots are used as sales representatives”. For instance, a company might exploit the user’s affective trust in the robot to make this latter convince the owner to purchase their goods. Whereas in human relationships, social emotional mechanisms like guilt and empathy might generally prevent the escalation of such scenarios, from the robot’s side there is nothing “to stop them from abusing their influence over their owners (Scheutz, 2009).

Secondly, if sensitive data might be divulged, worries regarding digital surveillance automatically follows. Digital surveillance of AEI systems expands into ever-more intimate spaces of users’ daily life (Lynch, 2021). Social robots’ appearance as genuinely benevolent towards the user may lead people to welcome them in the home and other locations, where for instance a surveillance camera would not be accepted (Sharkey & Sharkey, 2012). Furthermore, Mohammed (2022) expresses the worry that functional characteristics of AEI such as emotion recognition and sentiment analysis can be used for mass surveillance by companies for marketing reasons, and government for, allegedly, political security, without the users’ meaningful consent or being aware of such exploitation. Finally, given the highly sensitive data that such systems might collect, the risk of hacking becomes even more alarming. As Lutz et al. (2019) point out, hacking social robots allows for deeper insights into the user’s life than from other devices such as laptops or tablets.

When Trust in Robots Changes Human Relations. Others have focused on long-term consequences of affective trust in social robots based on their appearance. For instance, if it

occurs that “robots are generally more believable than humans” (Scheutz, 2009), this might encourage the user to rely more and more on their robots rather than their human fellows, and it might even lead to a general reduction on trust among people (Danaher & Sætra, 2022). Furthermore, Sætra (Sætra, 2021b) analyses the robot’s deceptive potential as one of the main causes for overtrusting robots, possibly leading to cultural change and moral revolution. He claims that exploiting trust in HRI has consequences on human-human relationships and might induce a corrosion of trust and cooperation between people in the short and long period. Whenever the user recognises that they have been deceived, they might be inclined to be less trustful in future interactions (Danaher & Sætra, 2022). This could consequently impact how humans relate to each other and, in an evolutionary time frame, the value of trust in human society as we know it now might be eroded (Sætra, 2021a).

To sum up, in the previous section we have explored two opposite perspectives about trust in social robots. In our view, both presented flaws and were susceptible to objections. The first perspective, based on Tallant’s (2019) argument, required a fundamental change in the design of AEI and social robotic technology, which might nullify the advantages and benefits of such systems for the user. The second, advanced by Coeckelbergh (2012), would implicitly encourage the occurrence of negative effects of granting trust in AEI social robots on the basis of their appearance. We have therefore explored some of these effects and the potential damage brought about to users and society in general. In the next section, by following Sweeney’s (2023) reasoning, we conclude our argument by presenting our own view.

5.3. Reliability as an essential requirement

Our discussion so far has highlighted essential features of social robots with AEI which encourage the user to develop affective trust towards them. Given their anthropomorphic forms, deceptive potential, and ability to elicit emotional responses and affection, people are inclined to believe that the robot possess benevolent and honest behaviours and intentionally act for the sake of the user, despite the robot lacking any genuine intentionality and agency. Merely counting on the appearance of the benevolence and integrity of the robot to grant trust towards these systems might not only negatively affect the user’s wellbeing but even endanger the value of trust in the long term. For this reason, we question Coeckelbergh’s (2012) argument that an apparent but arguably non-existent agency could be enough to allow robots to be moral objects of trust.

We do not intend to deny the evidence that, *empirically*, the user can develop affective trust attitudes towards AEI social robots. In fact, such attitudes of trust are necessary for the effective functioning of social robots. However, it is crucial to keep in mind that trusting these systems becomes dangerous exactly because we are inclined to accept them as genuinely benevolent entities despite being aware of their deceptive quality. The threshold at which, from the user's point of view, "objects become subjects" (Damiano & Dumouchel, 2018a), should arguably not be crossed.

In light of this, to avoid the development of overtrust and the risks associated, robots' synthetic nature should never be disregarded. Considerations of the robot's true capabilities and functionalities – what happens "inside" – and not only what the robot appears to be, have an importance in the assessment of trust towards them. In other words, what should be assessed when placing trust in AEI social robots is their actual reliability to function and perform in accordance with the user's goal. For instance, in cases of errors or malfunctioning of the system, or if it is found out that the owning company sells the collected users' sensitive data to marketing companies, the device is no longer *reliable*, and our spontaneously formed attitudes of affective trust should be necessarily and accordingly readjusted (Sweeney, 2023). Therefore, despite the implementation of AEI into social robotics, the increased robot's capability to deceive and be anthropomorphised, and the consequent importance of the emotional connection which one might develop, users' affective trust should still depend on and be susceptible to positive expectations about the system's reliability.

In conclusion, whereas we cannot deny that the user develops a *feeling* of trust towards the robot, such feeling is not enough to morally justify what we propose calling *trust proper* – a relational conception of trust which necessarily requires authentic capacities of benevolence and integrity. This is distinguished from the "thin" version of trust often used in HRI research, which is arguably very close to the same as simple *reliance*. Relying on the mere appearance of such qualities is not only not enough for trust proper to be established, but it could even expose the user to risks. Importantly, this conclusion does not necessarily necessitate modification of those features which encourage the user to believe the robot capable of benevolence and integrity. Again, a feeling of trust towards these systems could be both necessary and encouraged in certain HRI contexts.

Nonetheless, we claim that people must remain aware of the possibility for the product they are using to become unreliable and recalibrate their affective trust accordingly. Some ethical guidelines might be advanced for the correct development and use of these machines. For instance, informing users of the potential risks involved might be one way to temper an implicit and unwary development of affective trust based on the robot's appearance. Another suggestion comes from Sweeney (2023), who advocates for a need for transparency regarding the actual capabilities of the technology, and for greater responsibility from the manufacturers. As Picard (1997) argues, "In some cases, a machine may need to explain what it can and cannot do", but it will always be essential to help users to set accurate expectations of the technology's actual abilities. Interestingly, also AEI technology might become useful to detect and possibly prevent overtrust behaviours from users (Wagner et al., 2018). Whereas such guidelines would be effective remain still to be demonstrated, but the importance of investigating the potential dangers of trusting social robots and finding ways to temper them is nonetheless crucial.

6. Conclusion

We started by asking: What are the conditions for placing affective trust in technology? We explored the issue through the specific case of AEI social robots. Given their significant impact on human emotional and social dimension, we argued that such technologies are the most likely for people to develop attitudes of affective trust. Indeed, the robot's deceptive behaviours induce the user to believe the robot capable of benevolence and integrity. However, we have argued that granting trust based on their appearance involve risks which might endanger the user. We concluded our argument by claiming that it is therefore morally necessary to focus on the actual nature of AEI social robots, as synthetic tools designed for purposes which might differ or even conflict with the ones of the people that interact with it daily. Whereas we cannot deny that people develop a feeling of trust, and such a feeling is functional for the robot to fulfil the purposes it has been created for, we claim that users are nonetheless not morally allowed to grant trust proper towards these systems.

It might be objected that our analysis of the concept of trust is biased and lacks objectivity as affected by the culture we live in. Coeckelbergh (2012), for instance, claims that the identification and application of criteria for trusting, such as integrity and benevolence, are

influenced by Western culture and might therefore need to be reformulated according to a different cultural environment. This is an interesting point, which could also provide guidance for further empirical research on this subject. At the same time, we believe that this objection does not undermine our conclusion, as it is based on the consequences that trust in robots might bring about. Even if we agree that no universal definition of trust can be given, then we can at least refer to the implications of trusting the technology to assess the moral permissibility of this.

Finally, as we suggested that social robots equipped with AEI are the most likely candidates for trust to be developed, and we have nonetheless objected to such a possibility, it necessarily follows that trust cannot be granted for any kind of technology. To answer our research question, therefore, we conclude that affective trust in technology is not possible, as our tendencies to anthropomorphise and be deceived do and should not guarantee technology to be worthy of trust.

REFERENCES

- Airenti, G. (2015). The cognitive bases of anthropomorphism: from relatedness to empathy. *International Journal of Social Robotics*, 7(1), 117-127.
- Aroyo, A. M., De Bruyne, J., Dheu, O., Fosch-Villaronga, E., Gudkov, A., Hoch, H., Jones, S., Lutz, C., Sætra, H., & Solberg, M. (2021). Overtrusting robots: Setting a research agenda to mitigate overtrust in automation. *Paladyn, Journal of Behavioral Robotics*, 12(1), 423-436.
- Beer, J. M., Liles, K. R., Wu, X., & Pakala, S. (2017). Affective human–robot interaction. In *Emotions and affect in human factors and human-computer interaction* (pp. 359-381). Elsevier.
- Billings, D. R., Schaefer, K. E., Chen, J. Y., & Hancock, P. A. (2012). Human-robot interaction: developing trust in robots. Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction,
- Boada, J. P., Maestre, B. R., & Genís, C. T. (2021). The ethical issues of social assistive robotics: A critical literature review. *Technology in Society*, 67, 101726.
- Bok, S. (1978). *Lying: Moral choice in public life*. Pantheon Books.
- Boles, T. L., Croson, R. T., & Murnighan, J. K. (2000). Deception and retribution in repeated ultimatum bargaining. *Organizational Behavior and Human Decision Processes*, 83(2), 235-259.
- Borenstein, J., Wagner, A. R., & Howard, A. (2018). Overtrust of pediatric health-care robots: a preliminary survey of parent perspectives. *IEEE Robotics & Automation Magazine*, 25(1), 46-54.
- Breazeal, C. (2004). *Designing sociable robots*. MIT press.
- Broadbent, E. (2017). Interactions with robots: The truths we reveal about ourselves. *Annual review of psychology*, 68, 627-652.
- Campa, R. (2016). The rise of social robots: a review of the recent literature.

- Carpenter, J. (2016). *Culture and human-robot interaction in militarized spaces: A war story*. Routledge.
- Coeckelbergh, M. (2010). Moral appearances: emotions, robots, and human morality. *Ethics and Information Technology*, 12, 235-241.
- Coeckelbergh, M. (2012). Can we trust robots? *Ethics and Information Technology*, 14, 53-60.
- Coleman, J. S. (1990). *Foundations of social theory*. Harvard University.
- Damasio, A. (1994). *Descartes' error: Emotion, rationality and the human brain*. New York: Putnam, 352.
- Damiano, L., & Dumouchel, P. (2018a). Anthropomorphism in Human–Robot Co-evolution [Conceptual Analysis]. *Frontiers in Psychology*, 9. <https://doi.org/10.3389/fpsyg.2018.00468>
- Damiano, L., & Dumouchel, P. (2018b). Anthropomorphism in human–robot co-evolution. *Frontiers in Psychology*, 9, 468.
- Danaher, J. (2020a). Robot Betrayal: a guide to the ethics of robotic deception. *Ethics and Information Technology*, 22(2), 117-128.
- Danaher, J. (2020b). Welcoming robots into the moral circle: A defence of ethical behaviourism. *Science and engineering ethics*, 26(4), 2023-2049.
- Danaher, J., & Saetra, H. S. (2022). Technology and moral change: the transformation of truth and trust. *Ethics and Information Technology*, 24(3), 35.
- Darling, K. (2012). Extending legal protection to social robots. *IEEE Spectrum*.
- Darling, K. (2015). 'Who's Johnny?' Anthropomorphic framing in human-robot interaction, integration, and policy. *Anthropomorphic framing in human-robot interaction, integration, and policy (March 23, 2015)*. *Robot Ethics*, 2.
- De Visser, E. J., Monfort, S. S., McKendrick, R., Smith, M. A., McKnight, P. E., Krueger, F., & Parasuraman, R. (2016). Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology: Applied*, 22(3), 331.
- Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: a three-factor theory of anthropomorphism. *Psychological review*, 114(4), 864.
- Ferrin, D. L., Kim, P. H., Cooper, C. D., & Dirks, K. T. (2007). Silence speaks volumes: the effectiveness of reticence in comparison to apology and denial for responding to integrity-and competence-based trust violations. *Journal of Applied Psychology*, 92(4), 893.
- Fong, T., Nourbakhsh, I., & Dautenhahn, K. (2003). A survey of socially interactive robots. *Robotics and autonomous systems*, 42(3-4), 143-166.
- Freedy, A., DeVisser, E., Weltman, G., & Coeyman, N. (2007). Measurement of trust in human-robot collaboration. 2007 International symposium on collaborative technologies and systems,
- Friedman, B., Khan Jr, P. H., & Howe, D. C. (2000). Trust online. *Communications of the ACM*, 43(12), 34-40.
- Gaudiello, I., Lefort, S., & Zibetti, E. (2015). The ontological and functional status of robots: How firm our representations are? *Computers in Human Behavior*, 50, 259-273.
- Gino, F., & Shea, C. (2012). Deception in negotiations. *The Oxford handbook of economic conflict resolution*, 47-60.
- Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of management annals*, 14(2), 627-660.
- Goleman, D. (1995). *Emotional intelligence*. Bantam Books, Inc.

- Gompei, T., & Umemuro, H. (2018). Factors and development of cognitive and affective trust on social robots. International conference on social robotics,
- Gunkel, D. J. (2018). *Robot rights*. mit Press.
- Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y., De Visser, E. J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human factors*, 53(5), 517-527.
- Healy, P. (2022). Social robots as partners? *AI & SOCIETY*, 1-8.
- Heerink, M., Kröse, B., Evers, V., & Wielinga, B. (2010). Assessing acceptance of assistive social agent technology by older adults: the almere model. In: Springer.
- Heuer, T., Schiering, I., & Gerndt, R. (2018). Privacy and socially assistive robots-a meta study. *Privacy and Identity Management. The Smart Revolution: 12th IFIP WG 9.2, 9.5, 9.6/11.7, 11.6/SIG 9.2. 2 International Summer School, Ispra, Italy, September 4-8, 2017, Revised Selected Papers 12*, 265-281.
- Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human factors*, 57(3), 407-434.
- Hung, L., Liu, C., Woldum, E., Au-Yeung, A., Berndt, A., Wallsworth, C., Horne, N., Gregorio, M., Mann, J., & Chaudhury, H. (2019). The benefits of and barriers to using a social robot PARO in care settings: a scoping review. *BMC geriatrics*, 19, 1-10.
- Isaac, A., & Bridewell, W. (2017). Why robots need to deceive (and how). *Robot ethics*, 2, 157-172.
- Kahn Jr, P. H., Friedman, B., & Hagman, J. (2002). " I care about him as a pal" conceptions of robotic pets in online AIBO discussion forums. CHI'02 Extended Abstracts on Human Factors in Computing Systems,
- Kidd, C. D., & Breazeal, C. (2008). Robots at home: Understanding long-term human-robot interaction. 2008 IEEE/RSJ International Conference on Intelligent Robots and Systems,
- Kim, P. H., Dirks, K. T., Cooper, C. D., & Ferrin, D. L. (2006). When more blame is better than less: The implications of internal vs. external attributions for the repair of trust after a competence-vs. integrity-based trust violation. *Organizational Behavior and Human Decision Processes*, 99(1), 49-65.
- Koyama, T. (2016). Ethical issues for social robots and the trust-based approach. 2016 IEEE Workshop on Advanced Robotics and its Social Impacts (ARSO),
- LaGrandeur, K. (2015). Emotion, artificial intelligence, and ethics. *Beyond artificial intelligence: The disappearing human-machine divide*, 97-109.
- Lankton, N. K., McKnight, D. H., & Tripp, J. (2015). Technology, humanness, and trust: Rethinking trust in technology. *Journal of the Association for Information Systems*, 16(10), 1.
- Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human factors*, 46(1), 50-80.
- Lee, M. K., Tang, K. P., Forlizzi, J., & Kiesler, S. (2011). Understanding users' perception of privacy in human-robot interaction. Proceedings of the 6th international conference on Human-robot interaction,
- Lerner, J. S., Li, Y., Valdesolo, P., & Kassam, K. S. (2015). Emotion and decision making. *Annual review of psychology*, 66(1), 799-823.
- Levine, E. E., & Schweitzer, M. E. (2014). Are liars ethical? On the tension between benevolence and honesty. *Journal of Experimental Social Psychology*, 53, 107-117.
- Levine, E. E., & Schweitzer, M. E. (2015). Prosocial lies: When deception breeds trust. *Organizational Behavior and Human Decision Processes*, 126, 88-106.

- LOVOT. LOVOT. LOVOT. Retrieved September 4 from <https://lovot.life/en/>
- Lum, H. C. (2020). The role of consumer robots in our everyday lives. In *Living with robots* (pp. 141-152). Elsevier.
- Lutz, C., Schöttler, M., & Hoffmann, C. P. (2019). The privacy implications of social robots: Scoping review and expert interviews. *Mobile Media & Communication*, 7(3), 412-434.
- Lynch, C. R. (2021). Artificial emotional intelligence and the intimate politics of robotic sociality. *Space and Polity*, 25(2), 184-201.
- Matthias, A. (2015). Robot lies in health care: When is deception morally permissible? *Kennedy Institute of Ethics Journal*, 25(2), 169-162.
- Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of management review*, 20(3), 709-734.
- McAllister, D. J. (1995). Affect-and cognition-based trust as foundations for interpersonal cooperation in organizations. *Academy of management journal*, 38(1), 24-59.
- McKnight, H., Carter, M., & Clay, P. (2009). Trust in technology: Development of a set of constructs and measures.
- Mohammad, S. M. (2022). Ethics sheet for automatic emotion recognition and sentiment analysis. *Computational Linguistics*, 48(2), 239-278.
- Nickel, P. J., Franssen, M., & Kroes, P. (2010). Can we make sense of the notion of trustworthy technology? *Knowledge, Technology & Policy*, 23(3), 429-444.
- Nyholm, S. (2020). *Humans and robots: Ethics, agency, and anthropomorphism*. Rowman & Littlefield.
- Picard, R. W. (1997). *Affective Computing* (Vol. 252). MIT Press
- Plaks, J. E., Rodriguez, L. B., & Ayad, R. (2022). Identifying psychological features of robots that encourage and discourage trust. *Computers in Human Behavior*, 134, 107301.
- Prescott, T. J., & Robillard, J. M. (2021). Are friends electric? The benefits and risks of human-robot relationships. *Iscience*, 24(1).
- Reeves, B., & Nass, C. (1996). The media equation: How people treat computers, television, and new media like real people. *Cambridge, UK*, 10(10), 19-36.
- Richards, N. M., & Smart, W. D. (2016). How should the law think about robots? In *Robot law* (pp. 3-22). Edward Elgar Publishing.
- Riek, L. D., Rabinowitch, T.-C., Chakrabarti, B., & Robinson, P. (2009). How anthropomorphism affects empathy toward robots. Proceedings of the 4th ACM/IEEE international conference on Human robot interaction,
- Robinette, P., Howard, A. M., & Wagner, A. R. (2017). Effect of robot performance on human-robot trust in time-critical situations. *IEEE Transactions on Human-Machine Systems*, 47(4), 425-436.
- Rossi, A., Dautenhahn, K., Koay, K. L., & Walters, M. L. (2020). How social robots influence people's trust in critical situations. 2020 29th IEEE International conference on robot and human interactive communication (RO-MAN),
- Rousseau, D. M., Sitkin, S. B., Burt, R. S., & Camerer, C. (1998). Not so different after all: A cross-discipline view of trust. *Academy of management review*, 23(3), 393-404.
- Sætra, H. S. (2020). The parasitic nature of social AI: Sharing minds with the mindless. *Integrative Psychological and Behavioral Science*, 54(2), 308-326.
- Sætra, H. S. (2021a). Confounding complexity of machine action: a Hobbesian account of machine responsibility. *International Journal of Technoethics (IJT)*, 12(1), 87-100.
- Sætra, H. S. (2021b). Social robot deception and the culture of trust. *Paladyn, Journal of Behavioral Robotics*, 12(1), 276-286.

- Scheutz, M. (2009). The Inherent Dangers of Unidirectional Emotional Bonds between Humans and Social Robots. In P. Lin, K. Abney, & G. A. Bekey (Eds.), *Robot ethics: The ethical and social implications of robotics*. MIT Press.
- Schoorman, F. D., Mayer, R. C., & Davis, J. H. (2007). An integrative model of organizational trust: Past, present, and future. In (Vol. 32, pp. 344-354): Academy of Management Briarcliff Manor, NY 10510.
- Schuller, D., & Schuller, B. W. (2018). The age of artificial emotional intelligence. *Computer*, 51(9), 38-46.
- Severson, R. L., & Carlson, S. M. (2010). Behaving as or behaving as if? Children's conceptions of personified robots and the emergence of a new ontological category. *Neural Networks*, 23(8-9), 1099-1103.
- Sharkey, A., & Sharkey, N. (2012). Granny and the robots: ethical issues in robot care for the elderly. *Ethics and Information Technology*, 14(1), 27-40.
- Sharkey, N., & Sharkey, A. (2020). The crying shame of robot nannies: an ethical appraisal. In *Machine ethics and robot ethics* (pp. 155-184). Routledge.
- Sharkey, N., Wynsberghe, A. v., Robbins, S., & Hancock, E. (2017). Our Sexual Future with Robots. A Foundation for Responsible Robotics Consultation Report. In: The Hague.
- Song, Y., & Luximon, Y. (2020). Trust in AI agent: A systematic review of facial anthropomorphic trustworthiness for social robot design. *Sensors*, 20(18), 5087.
- Sullins, J. P. (2012). Robots, love, and sex: the ethics of building a love machine. *IEEE Transactions on Affective Computing*, 3(4), 398-409.
- Sung, J.-Y., Guo, L., Grinter, R. E., & Christensen, H. I. (2007). "My Roomba is Rambo": intimate home appliances. UbiComp 2007: Ubiquitous Computing: 9th International Conference, UbiComp 2007, Innsbruck, Austria, September 16-19, 2007. Proceedings 9,
- Sweeney, P. (2021). A fictional dualism model of social robots. *Ethics and Information Technology*, 23(3), 465-472.
- Sweeney, P. (2023). Trusting social robots. *Ai and Ethics*, 3(2), 419-426.
- Tallant, J. (2019). You can trust the ladder, but you shouldn't. *Theoria*, 85(2), 102-118.
- Torta, E., Werner, F., Johnson, D. O., Juola, J. F., Cuijpers, R. H., Bazzani, M., Oberzaucher, J., Lemberger, J., Lewy, H., & Bregman, J. (2014). Evaluation of a small socially-assistive humanoid robot in intelligent homes for the care of the elderly. *Journal of Intelligent & Robotic Systems*, 76(1), 57-71.
- Turkle, S. (2011). *Alone together: Why we expect more from technology and less from each other*. Basic Books.
- Van Maris, A., Zook, N., Caleb-Solly, P., Studley, M., Winfield, A., & Dogramadzi, S. (2020). Designing ethical social robots—a longitudinal field study with older adults. *Frontiers in Robotics and AI*, 7, 1.
- Wagner, A. R., & Arkin, R. C. (2011). Recognizing situations that demand trust. 2011 RO-MAN,
- Wagner, A. R., Borenstein, J., & Howard, A. (2018). Overtrust in the robotic age. *Communications of the ACM*, 61(9), 22-24.
- Wagner, A. R., & Nayyar, M. (2017). A theoretical conceptualization for overtrust. International Conference on Applied Human Factors and Ergonomics,
- Złotowski, J., Sumioka, H., Nishio, S., Glas, D. F., Bartneck, C., & Ishiguro, H. (2016). Appearance of a robot affects the impact of its behaviour on perceived trustworthiness and empathy. *Paladyn, Journal of Behavioral Robotics*, 7(1), 000010151520160005.

Article 3

Artificial emotions and the evolving moral status of social robots

1. Introduction

With the rapidly shifting landscape of technology, an intensified debate about the rights and moral standing of robots has emerged (Coeckelbergh, 2010b; Gunkel, 2018b; Gunkel, 2023). These questions have added a fresh dimension to our understanding of our own nature and morality itself, driven in part by how the status of machines is largely examined through explorations of what gives humans, and animals, moral standing (Gellers, 2020; Sætra, 2019). With exponential strides in the fields of robotics and artificial intelligence, the debate surrounding robots' moral status has become part of the mainstream of ethical discussions. For instance, Blake Lemoine, a former engineer at Google, ignited discussions when it suggested that their AI chatbot generator was conscious¹.

One key tenet of this discourse is the exploration of social robots – autonomous, physically embodied machines, designed to interact socially with humans (Darling, 2012). Various iterations of social robots have already permeated the society – from healthcare to education and companionship (Broadbent, 2017; Sica, 2023) – and it is expected that their presence, utility, and capabilities will see substantial growth in the future (Grace et al., 2018).

The moral status of social robots shape individuals' interactions with, and societies' application and regulations of, them. When the technology changes, so could the arguments in favour of – or against – granting them some moral status. This article aims to delve into this pressing dilemma, by exploring whether *artificial emotional intelligence* (AEI) offers a fresh lens through which we can evaluate and possibly redefine the ethical standing of social robots. Specifically, we aim to understand how AEI interacts with and possibly reshapes two dominant

¹ <https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/>

perspectives that influence the contemporary debate of robots' moral status: the property-oriented approach and the relational approach (Gunkel, 2018b).

The approach of considering AEI as a possible key determinant for the moral status of social robots has been partially covered (Corti et al., 2022) but still leaves much to be comprehensively understood and investigated. Through this analysis, we hope to shed light on the evolving discussion surrounding the moral standing of robots in our rapidly progressing tech-driven world. Furthermore, exploring the moral landscape in a world where technology and humanity increasingly converge serves as a mirror to our evolving understanding of morality itself, pushing us to question and redefine long-held beliefs about consciousness, the concept of person and human, and the essence of moral value and ethical consideration (Sætra, 2022).

The article is structured as follows. In section 2, we explore the concept of moral status and the kind of machines addressed, namely social robots equipped with AEI. In section 3, we describe two dominant theories for ascribing moral status – the property-based and the social-relational approach. The aim of this section is to evaluate the potential implications of integrating AEI into social robots on the prevailing theories. Specifically, it seeks to discern whether AEI challenges or consolidates the frameworks of these theories and assesses if it offers adequate grounds for granting moral consideration to robots. We conclude in section 4 with final reflections on the topics explored.

2. Background

Before we ask how our understanding of the moral status of machines changes with AEI, we must establish what sorts of machines we focus on. Even if our analysis will provide insight into the more general question of machines' moral standing, we arrive at this through a specific focus on social robots with AEI.

2.1. Social robots with AEI

A social robot is a physically autonomous entity designed to interact with humans on a social and emotional level, emulating human or animal-like behaviours and mindsets, and learning from these interactions (Darling, 2012). While humanoid social robots might seem particularly likely to evoke reactions resembling those evoked by other human beings, we have chosen to

focus on social robots in general. This is because we start from the premise that a social robot's form will not be the key determinant of its moral consideration. Supporting this view, various ethical perspectives claim that moral status should be granted to a range of non-human animals (Gruen, 2003; Regan, 2013; Singer, 2004), suggesting that human form is not central for the ascription of moral status.

The objective of social robotics is crafting machines that are perceived as more than mere instruments – to elevate these machines to genuine interaction partners (Damiano & Dumouchel, 2018). Such a goal seems to become more likely with the incorporation of AEI. AEI, as defined by Schuller and Schuller (2018), equips technology with the capacity to discern, generate, and utilise emotions for problem-solving and goal achievement, in order to create machines that are genuinely intelligent and able to interact with us seamlessly and authentically (Picard, 1997). More specifically, AEI can detect and recognise others' emotions by analysing emotion-related data (emotion recognition). It can express emotion towards interlocutors through facial and vocal expressions, head position, gestures, and body language (emotion generation). Finally, it can determine behaviour and activity selection based on internal simulated states (emotion augmentation) (Schuller & Schuller, 2018).

We also stress that AEI technology is still in development. Despite advancements in various related fields (Kaushik & Simmons, 2021) and notable progress in social robots such as Pepper (Fiorini et al., 2022), Buddy (Weiss & Hannibal, 2018) and Ryan (Abdollahi et al., 2022), fully equipping robots with human-like empathic abilities remains a challenge (Marcos-Pablos & García-Peñalvo, 2022). Given the limitations of its current applications, our exploration of AEI in this article is largely theoretical, assuming a future where AEI is fully realised as per the aforementioned definition.

The design of social robots equipped with AEI with the express intention of evoking various feelings in human users, and to simulate social cues, differentiates them from other robots that may unintentionally elicit human emotional reactions (Reeves & Nass, 1996). One example of the latter would be how some individuals grow attached to robotic devices such as the vacuum cleaner Roomba (Scheutz, 2009). Given the machine's functions, the human or animal likeness of their interactions, and the potential attachment and feelings they aim to evoke in users, it has been argued that social robots inhabit a grey area between machines and sentient beings (Damiano & Dumouchel, 2018; Kahn Jr et al., 2011; Sica & Sætra, 2023),

providing reason to question whether and how we should attribute moral standing to them, and distinguish them from other technological entities.

2.2. Moral status

This subsection delves into the concept of moral status, a necessary and yet far to be complete exploration given the scope and space constraints of this article. To begin with, we refer to Warren's (2000b) comprehensive and intuitive definition of moral status, described as a delineation of entities to which we have moral duties. Entities with moral status, also called moral patients, are seen as deserving of moral consideration based on their needs, interests, and wellbeing, regardless of benefits or disadvantages we might derive from them. Some authors equate moral status with having moral rights (Reichlin, 2014). While there is debate of this equivalence, both perspectives agree that having moral status, or being morally considerable, imposes obligations on moral agents. Moral agents are those from whom we expect adherence to morality and respect of others' rights.

Moral status, unlike legal status, is and will remain something individuals determine based on a wide variety of considerations. It consequently remains subjective. Legal status, on the other hand, can be granted by policymakers and would apply to all in a given legal context, regardless of their opinions and perception of the machines. While a philosophical treatise about the moral status of social robots with AEI could persuade some to change their opinion of these machines, we do not presume to adjudicate these questions for all. Changes of opinion could change social norms, however. Furthermore, dominant subjective perceptions of moral status tend to inform and change our legislation, which means that debates about robot's moral status can, but need not, change their legal status. In some instances, legal status is given purely for practical reasons, such as a delivery robots' right of way as a necessity for performing a function. A totally different case would be to grant robots some legal protection against, for example, abuse based not on protecting someone's property but because they become perceived as entities with rights to privacy and not to be subject to suffering. The latter case illustrates how debates about robots' moral standing overlaps with environmental ethics and animal rights discourse (Gellers, 2020; Sætra, 2019). We focus exclusively on moral standing in this paper.

3. Two paths to robots' moral status

The question of what moral standing machines have is often explored through two dominant perspectives: the property-based and the relational approach to morality (Gellers, 2020; Gunkel, 2018b). The *property-based approach* posits that moral standing is inherently linked to the possession of certain properties, such as consciousness or the ability to experience pain. Within this framework, the question becomes whether social robots with AEI can genuinely possess these properties or merely imitate them.

The *relational approach* suggests that moral standing is not a matter of inherent properties but is rather intricately tied to the relationships that entities forge with others. In other words, it is not about what an entity “is” or “has” but about the quality, depth, and nature of its interactions with surrounding entities. In the context of social robots, this viewpoint underscores the significance of the interactions and relationships these robots cultivate with human users, irrespective of the authenticity of their emotions.

Our primary objective is to assess whether AEI could have points of connection or relevance that either support or challenge these established theories, and consequently provide new perspectives on the moral standing of machines. In doing so, we do not aim to defend or attack either theory. However, our analysis highlights certain strengths and weaknesses of each approach through how AEI changes – or does not change – how machines should be conceived. There is also significant debate about both what properties “count” in property-based approach (Frank & Nyholm, 2017; Sung et al., 2007), and what the relational approach really entails (Kestutis Mosakas, 2021; Sætra, 2021). This means that both approaches can be used to grant – or deny – robots' moral status, and our intent is to highlight how, and not to determine which approach is more or less “correct”.

3.1. Property-based morality

Some theoretical accounts of moral status define it as a concept which indicates a being's particular attribute which confers it moral standing (Gellers, 2020; Gunkel, 2018b; Warren, 2000a). In other words, these perspectives suggest that an entity's moral value is determined by a specific characteristic or quality inherent to it. The crucial question concerns, therefore, the specific normative criterion on which moral status is based.

Different theories confer moral status based on factors such as being a human, a person, a living individual, a member of a community of forms of life, a sentient being (Warren, 2000a). There is no definite consensus on which specific properties are morally relevant, and a comprehensive analysis of the most suitable theory of moral status is beyond the scope of this article. Hence, for the sake of our discussion, we lean towards the two most discussed criteria when it comes to AI, namely reason and sentience (DeGrazia, 2022; Kestusis Mosakas, 2021; Muller, 2021; Rodogno, 2016; Sætra, 2019; Sytsma J, 2012). We assume that aligning our discussion with previous works not only validate our approach but also enable us to engage with existing debates and contribute meaningfully to the discourse surrounding robots' moral consideration when introducing the novel dimension of AEI.

In the next subsections, we define the concepts of reason and sentience and, following previous authors, explore their interconnection with emotions and broader moral considerations. We then pose the question whether and how artificial emotions could impact these theories, whether they could be equated to human emotions, and what are the consequences for the debate of the moral status of social robots possessing AEI. We finally raise some objections and concerns.

3.1.1. Reason

Reason is the capability to govern one's action by thinking, understanding, and making logical judgements about the world (Honderich, 2005). In the pursuit to comprehend the dynamics of human cognition, modern research has shown that reason and emotions are integrally entwined, contradicting the classical dichotomy that posits them as separate entities (Cacioppo & Gardner, 1999; Keltner & Lerner, 2010). Emotions play a significant role in human behaviour, communication, and interaction and are essential for the effective functioning of human intellect and thinking processes (Damasio, 2003; Damasio, 2006; Lerner et al., 2015). An example of the complex interactions between cognition and emotion in cognitive processes is the mood- congruent effect on memory: when we are in a particular mood, it affects not only what we remember but also how we take in new information and how we feel when we recall these memories (Bower & Forgas, 2000; Buchanan, 2007). Also working memory performance is affected by emotional states, as Levens and Phelps' (2008) study has shown. Furthermore, in the Dual Competition Model proposed by Pessoa (2009), emotions impact how information is processed in two specific ways: one that is driven by external

stimuli and another that depends on the individual's internal state. In both cases, a competitive interaction occurs at both the sensory perception level and the executive function level, indicating a complex interplay in how information is processed under emotional influence.

These examples illustrate how cognitive and emotional processes extend beyond mere interaction; their neural mechanisms are intertwined both influence behaviour (Liu et al., 2009). Given the indispensability of emotions for human reason, some have suggested to reconceptualise humans not as primarily rational but as fundamentally emotional animals (Turkle, 2011). In light of this, we seem to further accentuate the dichotomy between AI and humans, particularly when considering the conventional definition of AI as associated with systems with the computational capacity for goal-oriented tasks and problem-solving, notably excluding emotional considerations from its scope (Copeland, 2020; McCarthy, 2004). Thus, if we accept the proposition that reason lays the foundation for moral status, then we might conclude that AI, despite its advanced rational capabilities, does not qualify for moral valuation due to its inherent lack of an emotional component (Sætra, 2019)

The advance of AEI, however, introduces some complexities to the discourse, implying a possible recalibration of the moral assessment of entities (in our case, robots) possessing artificial emotions. The aim of the research in AEI is to develop computers "to be genuinely intelligent" (Picard, 1997), an intelligence analogous to human cognition which, as shown, consists of both emotional and rational capabilities. Should advancements in AEI eventually make it possible for artificial emotions to guide rational thought processes analogous to human cognitive functioning, a critical question arises: could such developments suffice for the conferral of moral status to AEI social robots?

A few objections could be raised. First, the choice of reason as the criterion for moral consideration is quite controversial and often objected, as it marginalises entities that many argue warrant inclusion within our moral community, such as individuals with mental impairments (Agnieszka & Tannenbaum, 2023). However, assuming we adhere to the assertion that reason is the fundamental moral criterion, it could still be objected that AEI inherently lacks the capability to genuinely experience emotions (Turkle, 2011) and, therefore, AI can never fully align with the constructs of human reason. Whether artificial and organic

emotions could be equated, however, it is a discussion that we leave after introducing the property of sentience.

3.1.2. Sentience, emotions, interests

Sentience refers to the capacity to have pleasant or unpleasant experiences (Gruen, 2003). Contrary to a narrow interpretation that might limit sentience to basic sensory experiences, such as physical pleasure and pain, we follow DeGrazia (2022) by arguing that sentience encompasses a range of emotional states (e.g., satisfaction and frustration), as well as instances of showing care or concern. Furthermore, Singer's (1986, 2011a) work posits a foundational connection between the capacity for emotional experiences and the recognition of interests, as he suggests that the very ability to undergo suffering or relish enjoyment serves as a prerequisite for the existence of interests. In Singer's view, an entity that remains incapable of experiencing these emotional nuances, and by extension, possessing interests, falls outside the realm of moral consideration.

Along similar lines, Rodogno (2016) argues that emotions are deeply intertwined with interests, identified as those moral values which hold profound significance for the subjects deserving moral considerations. He claims that, when an interest is satisfied or thwarted, emotional responses emerge, acting as indicators or reflections of these underlying values. This suggests that emotions, in many cases, serve both as motivators for action and as signals to what matters most to the individual (Stark & Hoey, 2021). Hence, the presence of such emotional responses underscores the possession of certain interests and moral values and, consequently, the candidacy for moral consideration.

This comprehensive understanding of sentience, emotions, and interests sets the stage for the dilemma at hand, namely the potential for AEI robots to acquire these attributes. This leads us back to the issue previously raised: is it feasible for a machine to experience emotions, or are these programmed responses mimicking genuine sentiment something else? If a robot does have the capability to exhibit emotions, as informed by earlier discourse linking emotions to moral worth, it naturally follows that they too might be deserving of moral status.

3.1.3. When it gets artificial

Hereafter, we first delve into theoretical frameworks of emotions, to determine how artificial emotions might affect ascription of moral status based on the selected properties. As we

explore some of these theories, it becomes evident that there is no straightforward answer to the potential equivalency of organic and artificial emotions. This ambiguity stems from the very fact that even within human psychology, there is no universal consensus of the nature, origins, and purposes of emotions. In this section, rather than delving deep into each theory or advocating for one over another, our aim is to highlight that the debate around the equivalency of organic and artificial emotions is at least partially tied to the broader, unresolved discussions about the essence and role of emotions themselves. As we will see, some theories appear to facilitate comparisons between the emotional capacities of humans and AEI robots. We conclude this section by advancing objections to these views.

Any interpretation of a robot's emotional capacity is largely influenced by the theoretical paradigm chosen. For instance, views such as the James-Lange theory consider emotions as processes related to physiological factors. According to such views, emotions are "sensory feelings constituted by perceptions of changes in physiological conditions relating to the autonomic and motor functions" (James, 1884). Accordingly, we do not tremble because we are afraid; rather, we feel fear because we tremble. Such views seem to confine emotions to organic entities capable of undergoing such physiological changes and likely challenge the possibility for AEI robot's emotions to equate humans' ones, however advanced their programming is².

However, motivational theories of emotions present a broader perspective, by considering emotions as "distinctive motivational states" and "internal causes of behaviours aimed at satisfying a goal" (Scarantino & De Sousa, 2018). This definition transcends the organic boundary and resonates more with artificial emotions. If emotions fundamentally drive behaviour, aiming to fulfil certain objectives, then robots with algorithms designed to mimic this process could, theoretically, be said to "experience" emotions. Advances in AEI have shown that machines can be constructed with emotion-based algorithms that do not just mimic emotions, but use them as central components in decision-making, responsiveness, and interactions (Moerland et al., 2018; Rosenbloom et al., 2015; Strömfelt et al., 2017), much in line with Damasio's (Damasio, 2006) theories of the role of emotions in human reasoning and decision-making. Here the emphasis is on the role of emotion augmentation in AEI social

² It could be counterargued, however, that bio-inspired and synthetic nervous system could achieve similar processes in generating emotions

robots. As mentioned before, this function pertains to the deliberate integration of emotional capabilities in machines, with the objective of either enhancing their functionality or making their interactions more relatable and human-like.

Strömfelt et al. (2017) provide a comprehensive overview of the different ways artificial emotion has been applied in AI at an infrastructure level, emphasising cases where emotion is *intrinsic* to the machine's architecture. They mention, among others, a model of AEI that provides a unique perspective on the role of emotion in decision-making processes, particularly in balancing exploration and exploitation. Its emotion generator comprises various elements like sensations, feelings, emotions, and a hormone system which interacts and influences the feeling component. Some of the emotions this system includes are: anxiety, activated when rewards decrease; confidence, which works inversely to anxiety; fear, which increases with anxiety and affects the choice of strategies; and warmth, which indicates when to stop the algorithm based on levels of fear and the number of iterations. Based on these sensations, a dominant emotion emerges, which serves an executive role as it adjusts specific parameters of the algorithm and potentially alters its strategy or actions.

If we accept that AEI social robots can “feel” in this manner, does it not suggest that there is a potential for such robots to possess genuine emotions, although in a format distinct from human experience? Furthermore, it is worth noting that such AEI processes seem to significantly bridge the gap for robots to align more closely with the above-described reason criterion. If the line between human emotions and artificial emotions becomes ambiguous, it beckons us to revisit and possibly redefine our understanding of emotions and experiences and, consequently, for the arguments stated above, social robots' position in our moral community. In the next subsection, we present counterarguments to such a view.

3.1.4. *Objections*

As we have seen, motivational theories of emotions could offer support to the idea that artificial emotions might be functionally equated to human emotions. However, this perspective is not without its challenges. A significant objection is rooted in the rich subjective experience of emotions. Arguably, emotions extend beyond mere behavioural responses. They manifest as sensations that individuals can introspect upon, recall, and be overwhelmed by.

This intrinsic nature of emotions leads to a crucial question: can an entity truly experience emotion without possessing the capability for subjective experience, or in essence, consciousness? A comprehensive exploration and conceptualisation of consciousness would far exceed the bounds of this article. However, a modest and general understanding of consciousness as the capacity to have private mental experiences (DeGrazia, 2022) seems enough to argue that consciousness becomes an indispensable component for an entity to experience any form of emotion, sensory perception, or care.

When considering social robots with AEI, the distinction between authentic emotional experience and programmed response becomes crucial. Causing distress in a system devoid of the capability to genuinely perceive it raises the question whether any real suffering occurs. Let's refer again to the example of the AEI model described in Strömfelt et al.'s (2017) work and consider an AEI model implemented in a robot that exhibits, e.g., the emotion of fear. Some might argue that such fear is, in essence, a programming design to avoid hazardous situations. Without consciousness acting as a backdrop, the genuineness of such emotion remains in question.

After all, as Mosakas (2021) suggests, it appears absurd to claim that pain carries meaningful implications for an entity that cannot genuinely perceive it. It follows that entities that lack consciousness naturally lack the capability to truly experience pleasure, pain, or any emotional nuance in between; namely, the capacity of sentience. For non-sentient beings, the external world's actions upon them are met with indifference, due to their inherent incapability to perceive and process these experiences meaningfully (Singer, 2011b). For the argument elaborated above, it follows that, without consciousness, an entity's claim to moral rights and consideration becomes tenuous. This places us on the threshold of reevaluating the weight we assign to artificial emotions and subsequently reshaping our moral obligations towards such entities.

However, a few objections can be advanced. First, it might be argued that, in the future, more sophisticated robots could reach a certain degree of consciousness (Butlin et al., 2023; Frank & Nyholm, 2017). Furthermore, the very definition and understanding of consciousness are not set in stone and, depending on which lens we view consciousness, we could posit that robots, even in their current state, possess a form of it (Bryson, 2012; Dennett, 2015).

Secondly, there is an ongoing debate which asks whether sentience can be attributed to living things without consciousness. For instance, Warren (1997) claims that non-self-aware beings, however unconscious, such as embryos and fetuses, might possess interests in, e.g., being alive. Embryos and fetuses might still have interests in continuing to live until birth, however being unaware of it.

Thirdly, when it comes to reason as the criterion for moral consideration, it might be argued that the enhancements that artificial emotions impart to machine capabilities may *sufficiently* mirror the contribution of emotions to human reasoning. It could be said that the presence of consciousness and authentic emotions may not be indispensable, as the approximation of artificial emotions to organic emotions may well enable robots to function in ways comparable to human cognition.

And, finally, as observed by some in the context of robots' moral agency, adhering to a property-based approach could be inherently flawed when it comes to robots (Kim et al., 2023). According to Kahn's (2011) New Ontological Category Hypothesis, robots represent a unique ontological category, consequently challenging the validity of applying the same criteria for moral patiency to robots as we do to living beings. As such, it may be necessary to adopt a different approach when considering robots in terms of moral status.

Such arguments are complex and warrant a more thorough examination than we cannot offer in this article. Nonetheless, by presenting such objections, our intention is to bring attention to the prevailing absence of unanimity on whether consciousness is essential for sentience and reason. This current lack of consensus renders the assertion - that artificial emotions are not comparable to human emotions and that AEI does not sufficiently justify attributing moral status to robots – significantly contentious.

In conclusion, this section has analysed the influence of social robots' artificial emotions on discussions about their moral status, based on a property-related approach. Using the property-based approaches anchored in the concept of reason and sentience, we investigated the parallels between artificial and human reasoning and emotions. Our findings indicate that AEI in social robots clearly influences the degree to which robots can aspire to the properties chosen. This implies that, depending on how one defines and approaches these concepts, AEI

could indeed change people's perception of their moral status and blur the boundaries between humans and machines.

3.2. The social-relational approach

In this section, we present the social-relational approach as an alternative method for the conditions of moral status ascription. We will see that the approach itself carries inherent challenges that may not be resolved by the technology of AEI. Nonetheless, this perspective stands to be considerably strengthened by the incorporation of AEI into social robots.

In traditional viewpoints, as the one previously explored, the entity's intrinsic properties determine its moral value; in other words, something has value because of what it inherently is. However, the social-relational approach, advanced by authors such as Coeckelbergh and Gunkel (2013, 2018a, 2018b, 2020; 2023) suggests that, instead of looking within the entity for its moral value, we should look at the relationships and social context surrounding it. Here, moral value is not intrinsic but rather is an outcome of the social interplays and dynamics; i.e., how the entity is situated in our social world and how we relate to it.

Interestingly, this framework holds that moral consideration of social robots does not necessarily hinge on authenticity. According to the property-based view, this authenticity is equated to a robot's ability to have genuine emotions. Instead, the social-relational view focuses on our interactions with, perceptions of, and experiences with the robot. It is the way an entity *appears* to us that determines the quality and depth of the relational experiences. The entity's properties still play a role as the foundation on which we base our moral consideration, but their importance is reframed (Sætra, 2021). Instead of them being valuable for what they inherently are, they matter because of how we experience and interpret them (Coeckelbergh, 2010b).

After all, Coeckelbergh (2010a) argues, when interacting with another individual, we do not have direct access to their consciousness or emotions. Instead, we rely on external cues – behaviour, expressions, words – for how they appear to us, and interpret these as indicators of their internal states. Our moral considerations are based on such interpretations, not on direct knowledge or authenticity. This marks a crucial distinction from property-based frameworks and opens new possibilities for considering social robots as morally worthy.

Empirical work in HRI provides valuable insights, particularly concerning the impact of people's perception of robots on moral and social considerations. Thellman et al. (2021) investigated the impact of people's emotional state on their perceptions of human-like qualities and mental capacities of robots, along with their attitudes towards these machines. Malle et al. (2016) demonstrated that robot appearance affects people's moral judgements about them. Carpinella et al. (2017) have developed a scale (RoSAS) to measure social perception of robots. Research conducted by De Graaf et al. (2021) involving an online survey exploring layman's attitudes towards granting particular moral and legal rights to robots, reveals that public opinions on a robot's emotional and cognitive capabilities significantly influence the decision to grant robots rights. Also, Weiss and Hannibal (2018), has proposed a study which explores how the relationship between social robots and users evolves over time.

According to the social-relational approach, a robot's moral standing is not reliant on its intrinsic properties but rather on the dynamics of its interaction within human society. It is about how the robot makes us feel, the kind of relationships we can form with it, and the experiences it can foster. Sex robots, falling under the broader range of social robots and sometimes referred to as love robots (Whitby, 2011), stand as a prominent example in this context. These robots have been developed beyond their initial purpose of physical gratification (Frank & Nyholm, 2017), and are increasingly seen as capable of engaging in a broader spectrum of intimate relationships (Scheutz & Arnold, 2016). The experiential and relational aspects of interacting with sex robots have served as platforms for exploring how robotic entities can fulfil emotional, social, and psychological needs in ways previously not considered, such as in context of care (Fosch-Villaronga & Poulsen, 2021). A notable example in this regard is Davecat's relationship with his sex doll Sidore, which he has been married to for over 20 years (Nyholm, 2023). This case suggests a foundational aspect of the social-relational approach, where the authenticity or the lifelike qualities of the doll are secondary to the emotional fulfilment, companionship, and psychological comfort Davecat derives from the relationship.

Nonetheless, Davecat's doll lacks any form of AI. Their relationship can be seen as an important, yet preliminary exploration in the discourse on the moral status of AI-enhanced robots from a social-relational perspective. As Coeckelbergh (2010b) argues, the robot needs

to be sufficiently advanced to be granted moral consideration. Which is to say, it has to simulate the ability to feel or experience so convincingly that, to an observer, they seem almost indistinguishable from genuine emotions. Within this framework, the capacity to simulate and generate emotions of AEI (what we have previously called emotion generation) in social robots becomes crucial. The simulation, if sophisticated enough, can potentially foster deeper and more genuine feeling relationships between humans and robots, consequently challenging the traditional boundaries we have established between the two entities. The emotional interactivity AEI could bring about between robots and humans, might amplify the illusion of genuine subjectivity of the robot, making this latter more relatable in their interactions. As we do not demand direct evidence of the presence of subjectivity or genuine emotions in another human, we should not necessarily demand the same of AEI social robots (Coeckelbergh, 2010a). If we interpret artificial emotions as sufficiently genuine, allowing us to foster meaningful interactions with the robot, this alone becomes the basis for the robot's moral status. Within the social-relational framework, therefore, AEI might significantly increase the likelihood for granting moral consideration to social robots equipped with such technology.

However, the social-relational approach raises criticisms which, as we will argue, might persist even in the face of AEI's advancements. Let us delve into some of these objections in the following.

3.2.1. Objections

A primary concern implicit in the social- relational approach is the adoption of a relativistic viewpoint regarding moral status. As discussed in the previous section, the moral significance of an entity is determined by the relationship and attachment individuals or communities form with that entity. In simpler terms, if people care about something or someone due to the connection they have established, then that entity is granted moral consideration. The problem is that the threshold for what qualifies as having moral status becomes highly variable and subjective. Since moral significance is influenced by feelings, personal biases, and cultural values, it can be argued that virtually anything can be given moral importance depending on individual or collective sentiments.

To highlight this problem, Muller (2021) makes the hypothetical example of a community that holds deep regard for pencils. They might expect outsiders to respect their sentiment towards pencils. However, this subjective reference should not be mistaken for an objective assertion that pencils have universal moral value. Similarly, more controversial, just because a group might devalue a particular set of individuals (red-haired women, in the given example) does not mean that those individuals lack moral value.

The underlying challenge is that the social-relational runs the risk of diluting the concept of moral status to mere personal inclinations. Without a more concrete and shared foundation for determining moral status, this approach could lead us into moral ambiguity, where each individual's or group's feelings become their own moral compass.

Furthermore, applying the social-relational approach to social robots possessing artificial emotions might introduce additional concerns, such as wrongly prioritising them over entities that truly deserve moral patiency, such as humans and animals (Birhane & van Dijk, 2020). Limited resources and attention could be diverted towards ensuring the rights of robots, neglecting moral obligations towards other beings. Implications of granting moral status to social robots should be critically evaluated.

Finally, authors such as Sætra (2021) have argued that the social relational approach may exhibit anthropocentric tendencies, as it places importance on how humans decide to respond to entities and what they decide to afford to them based on their performance, or even simply based on the human's projection of characteristics or performance of inanimate objects, such as a simple doll as in the case of Davecat. Therefore, this perspective ignores the intrinsic value of other beings or entities when not belonging to human relationships and establishes humans as the sole arbiters of moral status. This reconnects to the problem of prioritising entities designed to foster human emotions such as AEI social robots, over natural entities that might not immediately elicit strong human emotions but have inherent worth, such as animals like fishes in the sea.

To conclude this section, the challenges embedded within the social-relational approach are not easily surmounted, and the advent of artificial emotions does not seem to present resolutions to these issues. Nevertheless, the role of AEI becomes pivotal within this approach, opening up novel avenues for its proponents and possessing the potential to alter

perspectives regarding the moral status of social robots. While the predominance of the social- relational view in conferring moral status remains uncertain, it is plausible to suggest that the advancements in AEI render this perspective increasingly viable and worthy of consideration.

4. Conclusion

In this paper, we have asked whether the introduction of artificial emotions in social robots might influence the ongoing discourse regarding their moral status. We delved into the two predominant frameworks on this subject: the property-based and social- relational approaches, re-evaluating them specifically in the context of social robots enhanced with AEI. From our analysis, we deduce that there is not a definitive resolution, as the implementation of AEI brings robots closer to humans in more intrinsic, fundamental ways, introducing additional complexity to the debate.

In this last section, we aim to address some final considerations on the matter of robots' moral status. Although AEI does not seem to definitively bestow moral value upon existing social robots, these entities could still hold instrumental value (Nyholm, 2020). This aligns with the concept of derived moral status (Muller, 2021), indicating a kind of moral status based on the entity's instrumental, extrinsic or emotional value (Kestutis Mosakas, 2021; Muller, 2021).

Robots' derived moral status suggests that, while social robots might not possess intrinsic value akin to sentient or rational beings, their treatment and role in human societies warrant certain moral considerations. Indeed, it is imperative to consider the broader social implications of our interactions with these robots. Treating robots with cruelty or insensitivity, simply because they lack consciousness or genuine emotions, could foster an environment where such behaviour becomes normalised (Bryson, 2018; Parthemore & Whitby, 2014). While this is a potential concern with any object, social robots with artificial emotions merit special attention due to their appearance and their specific function in society.

A pertinent example relates to sex robots. While they lack consciousness, and thus one could argue that they cannot be directly harmed, harassed, or abused (Frank & Nyholm, 2017), they still mimic human features and behaviours. The ethical concern is not about causing harm to the robot itself, but about what such interactions might represent: a potential undermining of consent norms in human sexual relationships (Danaher, 2017). Any act that objectifies or

dehumanises these robots can have a ripple effect, potentially reinforcing or normalising demeaning behaviours in real-life human relationships (Nascimento et al., 2018; Sica, 2023). As such, advocating for respectful treatment of sex robots is less about safeguarding their wellbeing and more about preserving the dignity and rights of actual humans.

Finally, we want to highlight a few potential counterarguments to the propositions we have presented, indicating areas for further exploration in future research. For instance, one could posit that, as technological advancements persist, the future could bring forth artificial entities endowed with morally relevant attributes – such as emotions, interests, and a susceptibility to suffering – which exist solely in digital environments (Harris & Anthis, 2021). Such entities would be entirely devoid of a tangible, physical presence, whether it mirrors human form or any other recognisable structure.

Our focus on social robots might be seen therefore as a limitation in the scope of our study, challenging the necessity to restrict moral considerations to entities with physical embodiments and urging research on purely digital entities with the potential to manifest genuine emotions.

However, in response to these objections, our focus on robots is informed by theories which maintain that both reason and emotions are necessarily embodied experiences (Damasio, 1994; Müller, 2017; Shapiro & Spaulding, 2021). These theories assert that emotions are deeply intertwined with our bodily experiences and interactions with the world, suggesting that the presence of a physical body is crucial for the manifestation of morally relevant emotions. Thus, in this view, social robots offer a more authentic and meaningful parallel to human experiences and moral considerations than entities that are purely digital.

In closing, this article prompts discussion regarding the evolving moral landscape we find ourselves in as social robots are equipped with AEI. It is crucial to reflect on AEI's transformative role in eroding the demarcations traditionally drawn between humans and machines. As we have seen, AEI, by integrating emotional processing within the computational frameworks of robots, endangers a profound re-evaluation of our perceptions and interactions with these entities. Regardless of the specific theoretical perspective one adopts concerning the moral status of robots, the implications of AEI remain significant and far-reaching. This convergence of cognition and emotion within machines provokes deeper

reflections on the essence of consciousness and the nature of emotional experience, and demands a meticulous reconsideration of our ethical obligations, moral values and societal norms. It is essential, as technology advances, to continue to explore and re-evaluate our moral frameworks and ethical standpoints in relation to artificial entities, to foster a harmonious coexistence and to uphold the values and dignities that define our humanity.

REFERENCES

- Abdollahi, H., Mahoor, M., Zandie, R., Sewierski, J., & Qualls, S. (2022). Artificial emotional intelligence in socially assistive robots for older adults: a pilot study. *IEEE Transactions on Affective Computing*.
- Agnieszka, J., & Tannenbaum, J. (2023). The Grounds of Moral Status. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy*.
- Birhane, A., & van Dijk, J. (2020). Robot Rights? Let's Talk about Human Welfare Instead. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*.
- Bower, G. H., & Forgas, J. P. (2000). Affect, memory, and social cognition. *Cognition and emotion*, 87-168.
- Broadbent, E. (2017). Interactions with robots: The truths we reveal about ourselves. *Annual review of psychology*, 68, 627-652.
- Bryson, J. J. (2012). A role for consciousness in action selection. *International Journal of Machine Consciousness*, 4(02), 471-482.
- Bryson, J. J. (2018). Patience is not a virtue: the design of intelligent systems and systems of ethics. *Ethics and Information Technology*, 20(1), 15-26.
- Buchanan, T. W. (2007). Retrieval of emotional memories. *Psychological bulletin*, 133(5), 761.
- Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., Deane, G., Fleming, S. M., Frith, C., & Ji, X. (2023). Consciousness in Artificial Intelligence: Insights from the science of consciousness. *arXiv preprint arXiv:2308.08708*.
- Cacioppo, J. T., & Gardner, W. L. (1999). Emotion. *Annual review of psychology*, 50(1), 191-214.
- Carpinella, C. M., Wyman, A. B., Perez, M. A., & Stroessner, S. J. (2017). The robotic social attributes scale (RoSAS) development and validation. *Proceedings of the 2017 ACM/IEEE International Conference on human-robot interaction*,
- Coeckelbergh, M. (2010a). Moral appearances: emotions, robots, and human morality. *Ethics and Information Technology*, 12, 235-241.
- Coeckelbergh, M. (2010b). Robot rights? Towards a social-relational justification of moral consideration. *Ethics and Information Technology*, 12(3), 209-221.
- Copeland, B. J. (2020). artificial intelligence. *Encyclopedia Britannica. Inc.: Chicago, IL, USA*.
- Corti, L., Stefano, N. D., & Bertolaso, M. (2022). Artificial emotions: toward a human-centric ethics. *International Journal of Social Robotics*, 1-15.
- Damasio, A. (1994). Descartes' error: Emotion, rationality and the human brain. *New York: Putnam*, 352.
- Damasio, A. (2003). *Looking for Spinoza: Joy, Sorrow, and the Feeling Brain*. Harcourt inc.
- Damasio, A. (2006). *Descartes' Error: Emotion, Reason, and the Human Brain*. Quill. Penguin.
- Damiano, L., & Dumouchel, P. (2018). Anthropomorphism in human-robot co-evolution. *Frontiers in Psychology*, 9, 468.

- Danaher, J. (2017). The Symbolic-Consequences Argument in the Sex Robot Debate. In J. Danaher & N. McArthur (Eds.), *Robot sex: Social and ethical implications*. MIT Press.
- Darling, K. (2012). Extending legal protection to social robots. *IEEE Spectrum*.
- De Graaf, M. M., Hindriks, F. A., & Hindriks, K. V. (2021). Who wants to grant robots rights? Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction,
- DeGrazia, D. (2022). Robots with moral status? *Perspectives in Biology and Medicine*, 65, 73-88.
- Dennett, D. C. (2015). *Elbow room, new edition: The varieties of free will worth wanting*. mit Press.
- Fiorini, L., Loizzo, F. G., D'Onofrio, G., Sorrentino, A., Ciccone, F., Russo, S., Giuliani, F., Sancarlo, D., & Cavallo, F. (2022). Can I Feel You? Recognizing Human's Emotions During Human-Robot Interaction. International Conference on Social Robotics,
- Fosch-Villaronga, E., & Poulsen, A. (2021). Sex robots in care: Setting the stage for a discussion on the potential use of sexual robot technologies for persons with disabilities. Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction,
- Frank, L., & Nyholm, S. (2017). Robot sex and consent: Is consent to sex between a robot and a human conceivable, possible, and desirable? *Artificial intelligence and law*, 25, 305-323.
- Gellers, J. C. (2020). *Rights for Robots: Artificial Intelligence, Animal and Environmental Law*. Routledge.
- Grace, K., Salvatier, J., Dafoe, A., Zhang, B., & Evans, O. (2018). When will AI exceed human performance? evidence from AI experts. *Journal of Artificial Intelligence Research*, 2, 729-754.
- Gruen, L. (2003). The moral status of animals.
- Gunkel, D. J. (2013). Mark Coeckelbergh: Growing moral relations: critique of moral status ascription: Palgrave Macmillan, New York, 2012, 239 pp, ISBN: 978-1-137-02595-1. In: Springer.
- Gunkel, D. J. (2018a). The other question: can and should robots have rights? *Ethics and Information Technology*, 20, 87-99.
- Gunkel, D. J. (2018b). *Robot rights*. mit Press.
- Gunkel, D. J. (2020). The right (s) question: Can and should robots have rights? In *Artificial intelligence* (pp. 255-274). Brill mentis.
- Gunkel, D. J. (2023). *Person, Thing, Robot: A Moral and Legal Ontology for the 21st Century and Beyond*. MIT Press.
- Harris, J., & Anthis, J. R. (2021). The moral consideration of artificial entities: a literature review. *Science and engineering ethics*, 27(4), 53.
- Honderich, T. (2005). *The Oxford companion to philosophy*. OUP Oxford.
- James, W. (1884). What is an emotion? *Mind*, 9, 188-205.
- Kahn Jr, P. H., Reichert, A. L., Gary, H. E., Kanda, T., Ishiguro, H., Shen, S., Ruckert, J. H., & Gill, B. (2011). The new ontological category hypothesis in human-robot interaction. Proceedings of the 6th international conference on Human-robot interaction,
- Kaushik, R., & Simmons, R. (2021). Perception of emotion in torso and arm movements on humanoid robot quori. Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction,

- Keltner, D. T., & Lerner, J. S. (2010). Emotion. In D. T. Gilbert, D. T. Fiske, & G. Lindzey (Eds.), *The Handbook of Social Psychology* (Vol. 1). Wiley.
- Kim, B., Phillips, E., Zhu, Q., & Williams, T. (2023). *Perspectives on Moral Agency in Human-Robot Interaction* Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction, Stockholm, Sweden. <https://doi.org/10.1145/3568294.3579966>
- Lerner, J. S., Li, Y., Valdesolo, P., & Kassam, K. S. (2015). Emotion and decision making. *Annual review of psychology*, 66, 799-823.
- Levens, S. M., & Phelps, E. A. (2008). Emotion processing effects on interference resolution in working memory. *Emotion*, 8(2), 267.
- Liu, Y., Fu, Q., & Fu, X. (2009). The interaction between cognition and emotion. *Chinese Science Bulletin*, 54(22), 4102-4116.
- Malle, B. F., Scheutz, M., Forlizzi, J., & Voiklis, J. (2016). Which robot am I thinking about? The impact of action and appearance on people's evaluations of a moral robot. 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI),
- Marcos-Pablos, S., & García-Peñalvo, F. J. (2022). Emotional intelligence in robotics: a scoping review. *New Trends in Disruptive Technologies, Tech Ethics and Artificial Intelligence: The DITTET Collection 1*,
- McCarthy, J. (2004). What is artificial intelligence?
- Moerland, T. M., Broekens, J., & Jonker, C. M. (2018). Emotion in reinforcement learning agents and robots: a survey. *Machine Learning*, 107, 443-480.
- Mosakas, K. (2021). On the moral status of social robots: considering the consciousness criterion. *AI & SOCIETY*, 36, 429-443.
- Mosakas, K. (2021). On the moral status of social robots: considering the consciousness criterion. *AI & SOCIETY*, 36(2), 429-443.
- Muller, V. C. (2021). Is it time for robot rights? Moral status in artificial entities. *Ethics and Information Technology*, 23, 579-587.
- Müller, V. C., & Hoffmann, M. (2017). What is morphological computation? On how the body contributes to cognition and control. *Artificial life*, 23(1), 1-24.
- Nascimento, E. C. C., Silva, E. d., & Siquiera-Batista, R. (2018). The 'use' of sex robots: A bioethical issue. *Asian Bioethics Review*, 10(3), 231-240.
- Nyholm, S. (2020). *Humans and robots: Ethics, agency, and anthropomorphism*. Rowman & Littlefield.
- Nyholm, S. (2023). Robotic animism: the Ethics of attributing minds and personality to Robots with Artificial Intelligence. In *Animism and Philosophy of Religion* (pp. 313-340). Springer.
- Parthemore, J., & Whitby, B. (2014). Moral agency, moral responsibility, and artifacts: What existing artifacts fail to achieve (and why), and why they, nevertheless, can (and do!) make moral claims upon us. *International Journal of Machine Consciousness*, 6(02), 141-161.
- Pessoa, L. (2009). How do emotion and motivation direct executive control? *Trends in cognitive sciences*, 13(4), 160-166.
- Picard, R. W. (1997). *Affective Computing* (Vol. 252). MIT Press
- Reeves, B., & Nass, C. (1996). The media equation: How people treat computers, television, and new media like real people. *Cambridge, UK*, 10(10), 19-36.
- Regan, T. (2013). 'The Case for Animal Rights'. In *Ethics, Humans and Other Animals* (pp. 179-188). Routledge.

- Reichlin, M. (2014). Moral Status. In H. ten Have (Ed.), *Encyclopedia of Global Bioethics*. Springer.
- Rodogno, R. (2016). Robots and the Limits of Morality. In M. Nørskov (Ed.), *Social Robots: Boundaries, Potential, Challenges* (pp. 39-55). Ashgate.
- Rosenbloom, P. S., Gratch, J., & Ustun, V. (2015). Towards emotion in sigma: from appraisal to attention. Artificial General Intelligence: 8th International Conference, AGI 2015, AGI 2015, Berlin, Germany, July 22-25, 2015, Proceedings 8,
- Sætra, H. S. (2019). Man and his fellow machines: An exploration of the elusive boundary between man and other beings. In F. O. E. S. Larsen (Ed.), *Discussing borders, escaping traps: Transdisciplinary and transspatial approaches*. Waxman.
- Sætra, H. S. (2021). Challenging the Neo-Anthropocentric Relational Approach to Robot Rights [Hypothesis and Theory]. *Frontiers in Robotics and AI*, 8. <https://doi.org/10.3389/frobt.2021.744426>
- Sætra, H. S. (2022). Robotomorphy: Becoming our creations. *Ai and Ethics*, 2(1), 5-13.
- Scarantino, A., & De Sousa, R. (2018). Emotion. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*.
- Scheutz, M. (2009). The Inherent Dangers of Unidirectional Emotional Bonds between Humans and Social Robots. In P. Lin, K. Abney, & G. A. Bekey (Eds.), *Robot ethics: The ethical and social implications of robotics*. MIT Press.
- Scheutz, M., & Arnold, T. (2016). Are we ready for sex robots? 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI),
- Schuller, D., & Schuller, B. W. (2018). The age of artificial emotional intelligence. *Computer*, 51(9), 38-46.
- Shapiro, L., & Spaulding, S. (2021). Embodied Cognition. *The Stanford Encyclopedia of Philosophy*.
- Sica, A. (2023). *The Robot will Feel You Now: The Ethics of Artificial Emotional Intelligence in Sex Robots* IEEE Ro-Man 2023, Busan.
- Sica, A., & Sætra, H. S. (2023). In Technology We Trust! But Should We? International Conference on Human-Computer Interaction,
- Singer, P. (1986). *Applied ethics*. Oxford University Press.
- Singer, P. (2004). Animal liberation. In *Ethics: Contemporary Readings* (pp. 284-292). Routledge.
- Singer, P. (2011a). *The expanding circle: ethics, evolution and moral progress*. Princeton University Press.
- Singer, P. (2011b). *Practical ethics*. Cambridge university press.
- Stark, L., & Hoey, J. (2021). *The ethics of emotion in artificial intelligence systems* Conference on Fairness, Accountability, and Transparency, New York, NY, USA.
- Strömfelt, H., Zhang, Y., & Schuller, B. W. (2017). *Emotion-Augmented Machine Learning: Overview of an Emerging Domain* Affective Computing and Intelligent Interaction (ACII 17),
- Sung, J.-Y., Guo, L., Grinter, R. E., & Christensen, H. I. (2007). "My Roomba is Rambo": intimate home appliances. UbiComp 2007: Ubiquitous Computing: 9th International Conference, UbiComp 2007, Innsbruck, Austria, September 16-19, 2007. Proceedings 9,
- Sytsma J, M. E. (2012). Two sources of moral standing. *Review of Philosophy and Psychology*, 3, 303-324.

- Thellman, S., Thunberg, S., & Ziemke, T. (2021). Does Emotional State Affect How People Perceive Robots? Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction,
- Turkle, S. (2011). *Alone together: Why we expect more from technology and less from each other*. Basic Books.
- Warren, M. A. (1997). *Moral status: Obligations to persons and other living things*. Clarendon Press.
- Warren, M. A. (2000a). *Moral Status: Obligations to Persons and Other Living Things*. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780198250401.001.0001>
- Warren, M. A. (2000b). *Moral Status: Obligations to Persons and Other Living Things*. . Oxford University Press.
- Weiss, A., & Hannibal, G. (2018). What makes people accept or reject companion robots? A research agenda. Proceedings of the 11th PErvasive Technologies Related to Assistive Environments Conference,
- Whitby, B. (2011). Do You Want a Robot Lover? The Ethics of Caring Technologies. *Robot ethics: The ethical and social implications of robotics*, 233.

Article 4

The robot will feel you now: The ethics of artificial emotional intelligence in sex robots

1. Introduction

Sex robots have become a topic of growing ethical and social concern, particularly regarding their potential impact on sexual health, human relationships, and the implications of humans forming emotional attachments to them. Indeed, manufacturers are aiming to build sex robots that have the capability of showing intimacy and love. Compared to other devices, including other kinds of social robots, sex robots could pose unique risks to users given their employment in very intimate contexts and their distinctive potential to trigger intense emotions and feelings such as love.

This article aims to address the ethics of creating a sex robot capable of eliciting loving feelings in the user. In addition, it focuses on the development of artificial emotion intelligence (AEI) and how its implementation in sex robot technology gives reason for new concerns, as AEI is precisely targeted at evoking human emotions. My research question is whether AEI exacerbates or offers a solution to the ethical and social issues surrounding sex robots, particularly in relation to sexual health and human-human relationships.

The paper is structured as follows. I first present a brief description of sex robots and AEI (section 2), followed by a review of some of the philosophical literature there is on this topic already (section 3). I then address the ethical issues that arise specifically with the implementation of AEI and the development of intimate and loving relationships with sex robots, including social isolation, deception, manipulation, and their societal impact (section 4). Finally, I offer suggestions for an ethical design of sex robots, argue that the technology of AEI might provide some solutions to the concerns examined in the article (section 5).

2. Beyond sex with robots

Before addressing the core of my argument, it would be helpful to start with a general description of what it is meant by ‘sex robot’ and AEI. I will begin by briefly delineate features and characteristics of sex robots’ software and hardware, and provide a description of AEI technology. I will then explain why I am specifically focusing on AEI sex robots and why it is crucial to address the ethical considerations of this topic.

Sex robots are humanoid social robots that perform actions contributing directly and *primarily* towards “improvement in the satisfaction of the sexual needs of a user” (Fosch-Villaronga & Poulsen, 2020). In contrast to sex *dolls*¹, these robots are equipped with sensors and actuators, and possess some degree of artificial intelligence – i.e., they are capable of interacting with their environment. They can react and respond to their users by demonstrating conversation, and displaying emotions and programmed personalities with an animatronic head that can move, smile, blink and chat (Döring & Pöschl, 2018). They also exhibit partially autonomous behaviours of sexual awareness and responsiveness, e.g., by simulating sexual movements and expressing orgasm (Sharkey et al., 2017). Different models are available on the market, such as Harmony and Henry by Abyss Creations, or Roxxy Gold and Rocky Gold by TrueCompanion (Sharkey et al., 2017).

Sex robot designers are now striving to create more advanced and complex machines that can offer more than just physical pleasure by interacting with users on a deeper, affectionate level (Frank & Nyholm, 2017). The goal is to establish an emotional connection with the robot, with the hope that users will perceive it as a ‘true companion’ and even express love towards it (Frank & Nyholm, 2017). At present, robots cannot feel love or form emotional connections². However, by simulating human-like emotions, expressions, and behaviours, they can create the illusion of love, intimacy, and tenderness. Given human tendencies to anthropomorphise (i.e., to attribute human properties and characteristics to non-human entities) and to be vulnerable to deception (i.e., to attribute mental states, intentions and abilities to the robot),

¹ Outlining differences as well as eventual similarities with sex dolls and sex toys might be useful for the ethical assessment of sex robots. Given the scarcity of current empirical research about the development and potential consequences of the use of sex robots, we might base our analysis on data concerning the use of precursor technologies such as sex dolls and sex toys, provided that parallels between the three categories emerge (Döring & Pöschl, 2018).

² Advocates of behaviourism might object my claim. For further discussion on this topic, see Wagner et al. (2018).

we can plausibly assume that the possibility of a human-robot relationship that goes beyond mere sexual intercourse and involves emotional attachment and intimacy is not pure fantasy (Airenti, 2015; Danaher, 2020; Glikson & Woolley, 2020; Matthias, 2015; Sung et al., 2007).

AEI can play a crucial role in developing such robots. Emotions have been increasingly used in social robotic design to make human-robot interactions more believable, acceptable and, consequently, effective (Fong et al., 2003). With a focus on sex and robots, the capacity of such devices to elicit emotions and create an affecting bond with their users is expected to significantly intensify with the implementation of AEI. Schuller and Schuller (2018) define AEI as a technology possessing three core functions. AEI can detect and recognise others' emotions by analysing emotion-related data (*emotion recognition*). It can express emotion towards interlocutors through facial and vocal expressions, head position, gestures and body language (*emotion generation*). Finally, it can determine behaviour and activity selection based on internal simulated states (*emotion augmentation*).

This emerging field of studies within AI research finds its greatest potential in developing computers "to be genuinely intelligent and to interact naturally with us" (Picard, 1999). As research has shown, emotional and volitional factors play a significant role in human behaviour, communication, and interaction, and are essential for the effective functioning of human intellect and thinking processes (Damasio, 1994; Lerner et al., 2015). Specifically, emotional intelligence consists of the individual's ability to understand, manage, and accordingly adjust one's own emotions, as well as to correctly interpret others' emotions and respond to their behaviours (Goleman, 1995). As a result, we can plausibly assume that relationships with emotionally intelligent sex robots will involve an even greater degree of intimacy and affection, consequently exacerbating the possibility of a user to establish an emotional bond.

In this paper, I will focus on these advanced but still hypothetical sex robots. As the technology is not yet fully realized, the discussion presented here will be mostly theoretical.

3. What is wrong with sex robots (so far)?

In the following, I present a general consideration of the most frequently debated benefits and problems associated with the use of sex robots. This includes aspects concerning sexual

health, their influence on human relationships, their impact on gender issues, and specifically, their implication for the statuses of women in society.

3.1. Benefits of sex with robots

It has been argued that sex robots might offer potential benefits (Danaher et al., 2017; Di Nucci, 2017; McArthur, 2017), especially in relation to the individual's sexual health and wellbeing. Sexual health is defined as "the ongoing process of physical, psychological, and socio-cultural wellbeing related to sexuality", which requires a positive and respectful approach to sexuality (Pan American Health Organisation, 2000). Given that sexual expression is an important aspect of human experience and can contribute to feelings of happiness and satisfaction, the use of sex robots could lead to improved overall wellbeing and mental health.

For instance, sex robots could enhance general sexual satisfaction, particularly for people with physical or mental impairments who face challenges or even impossibilities in accessing sexual experiences. Also, sex robots might help to bridge the so-called orgasm gap, a disparity of orgasm frequency between men and women in heterosexual sex (Mahar et al., 2020). They could offer a means for heterosexual women to experiment with their own bodies and learn what feels good to them, without the pressure or judgement that may come from a male human partner. In addition, sex robots would provide more consistent stimulation and attention to women's sexual needs, which may help to increase the likelihood of orgasm³.

Moreover, sex robots have been suggested as a potential tool for comprehensive sexuality education, which includes increasing sexual health and wellbeing, promoting healthy and respectful sexual behaviour, and leading to a more positive and accepting attitude towards sexuality in general (Dubé & Anctil, 2021). Indeed, one advantage of using sex robots is that they can simulate various sexual situations and behaviours, allowing their users to learn about sexual anatomy, physiology, and hygiene in a practical, interactive, and safe way. They could also help to model and reinforce positive sexual behaviour, such as asking for and respecting consent, and encourage communication and boundary setting with their sexual partners (Dubé & Anctil, 2021; Frank & Nyholm, 2017). Furthermore, sexual experience with robots

³ Clearly, however, sex robots alone would not be enough to address fully the complex issue of orgasm gaps, as this involves many social, cultural, and psychological factors.

can assure a non- judgemental space for exploring sexual desires and preferences, and for promoting self-exploration and self- discovering, especially for those who may feel ashamed or embarrassed about expressing their sexuality (McArthur, 2017).

3.2. Problems of sex with robots

On the other hand, sex robots are believed to have a profound and negative impact on people's sexual health, sexual and close relationships, and general wellbeing (Nascimento et al., 2018). For instance, by promoting specific body types or sexual performances, sex robots might create unrealistic expectations for real-life sexual partners, which might lead to dissatisfaction, frustration, or even low self-esteem. Also, individuals who rely heavily on sex robots may have difficulties communicating their desires, boundaries, and preferences with human partners, hindering the development of a healthy sexual relationship.

Also, it has been argued that sex robots could perpetuate harmful power dynamics in human relationships, particularly if they are designed to simulate subservient or subordinate roles (Danaher, 2017). Given that most sex robots are depicted with stereotypical conceptions of female bodies (Döring & Pöschl, 2018), some experts argue that this could reinforce problematic gender norms, where women are seen as objects to be dominated or controlled, and men are seen as being entitled to such domination (Goleman, 1995; Richardson, 2016).

Another ethical and social issue is that sex robots may lower inhibition thresholds and lead to an increase in deviant sexual behaviours (Döring & Pöschl, 2018). This concern arises from the idea that people may become desensitised to more extreme forms of sex by using such machines, leading them to engage in these behaviours more frequently or in their relationships with other humans (Sharkey et al., 2017). This could be especially true if robots were designed to simulate non-consensual or violent sexual acts, potentially increasing the likelihood of people who are already predisposed to such deviant sexual behaviours (Nyholm, 2022). Again, by depicting objectified female sexual bodies and potentially encouraging violent sexual intercourses, sex robots might further exacerbate misogyny and aggravate violence against women, which is considered "a global health problem of epidemic proportions" (Nascimento et al., 2018). This could lead to other types of physical and psychological trauma, such as mental health disorders, substance use problems, injury and death (World Health Organization, 2013).

Such a worry, however, lacks empirical evidence, and some have objected that, rather, using sex robots for individuals with deviant sexual behaviours, such as paedophilia, might potentially serve as a cathartic or even therapeutic outlet. In this perspective, sex robots could provide a non-harmful way for these individuals to manage their urges and consequently prevent the harm that might otherwise be inflicted on real individuals, particularly children and women.

4. Implementing AEI

Besides concerns and benefits related to the sexual act with robots itself, the ethical discussion surrounding sex robots is becoming increasingly focused on the potential for these entities to evoke emotions and strong feelings in users, potentially related to the sentiment of love (Borenstein & Arkin, 2019; Frank & Nyholm, 2017; Nyholm & Frank, 2019; Sullins, 2012; Viik, 2020). The implementation of AEI into sex robots' software, and the consequent enhancement of affection and love felt towards these robots, can have a significant impact on the moral issues explored previously, and potentially give rise to new ones. This discussion will be addressed in the following, but first, a premise is necessary.

Some of the ethical issues which will be discussed here might relate to social robots more broadly. Nonetheless, I argue that such problems assume a more complex dimension when it comes to sex robots. This complexity arises from the intimate nature of interactions with sex robots, especially when enabled with AEI, and the resulting emotional vulnerability of users. We might plausibly assume that, when in love, people are generally more susceptible. Some of the ethical concerns generally associated with social robots could pose even more serious implications for users of AEI sex robots, since they could be more easily manipulated due to their increased emotional involvement. Even if both typologies of robots are designed to simulate emotional responses, the depth of emotional engagement expected from sex robots might be significantly more intense and complex due to their intimate role. Furthermore, social robots and sex robots could have a different impact on societal norms. Specifically, sex robots, given their intimate function, could potentially shape societal views on sexuality, consent, and romantic relationships.

Some might argue that love and intimacy can also be developed towards social robots which are not designed for that purpose. After all, we have examples of people in love with their sex

dolls (Nyholm et al., 2022), showing that the object of affection does not need to possess specific capabilities for feelings to flourish. Along the same line, while AEI can facilitate the emergence of love, it is not a prerequisite for love to develop, nor is it necessary for the occurrence of any of the issues discussed in this section. However, while with social robots love may be a casual and unexpected outcome, with AEI sex robots love is the intended objective. This, from an ethical perspective, marks a crucial distinction.

4.1. Social isolation

A concern arising from the possibility of intimate and loving relationships with robots is the risk of loss of human contact, with consequences such as moral and social deskilling (Friedman et al., 2022), and desensitisation of intimacy and empathy, which are only developed by “experiencing human interaction and consensual relationships” (Sætra, 2022). Some argue that social isolation and consequent loneliness could be due to AEI robots being easier and more satisfying to interact with, emotionally and sexually, than people (Nyholm, 2023), which could result in humans becoming sexually unattractive and unable to meet their own standards (Beschorner & Krause, 2017).

Others have instead suggested that the loss of human contact might arise from the fact that robots fail to meet the specific needs of humans (Beschorner & Krause, 2018). One reason for this is because the user’s awareness that the robot, however equipped with AEI, has no experience of genuine emotion or capacity for pleasure might affect the loving and/or sexual interaction⁴ (Sharkey et al., 2017). As McArthur (2017) argues, it may be impossible for a robot to replicate the complete psychological experience of human sexual intercourse, and “we may think this would be for the best”. Therefore, as long as robots’ agency is an illusion (i.e., they can only *pretend* to possess loving feelings and capacity to experience sexual pleasure), love and sex with humans will continue to offer something that intimacy and sexual activity involving robots does not (Migotti & Wyatt, 2017).

Furthermore, robots are artefacts that cannot love you back (Turtle, 2011) and can only offer a “one-sided relationship” (Sharkey et al., 2017). In human-robot relationships, whether these

⁴ Sharkey et al. (2017) offer an interesting parallel between sex robots’ and sex workers’ pretence of love and pleasure. They cite a former sex worker and writer, Cathryn Berarovic, who claims that “no matter how good a whore is at her job, the client always knows, somewhere in his head, that he’s paying for this woman’s time and renting access to her body”.

are loving, intimate or sexual, “the deep and nuanced notions of love and the concord of true friendships” are ignored (Dubé & Anctil, 2021). Of course, we might find these machines physically attractive and personally engaging, but they “would only satisfy, but not truly satisfy, our physical and emotional needs, while doing nothing for our moral growth” (Dubé & Anctil, 2021).

However, this cannot be currently proved empirically since, as previously specified, the development of these machines is still in the hypothetical stage. Moreover, we should not underestimate the human ability to suspend disbelief, as well as to override the possibility that the user’s perception of the robot’s mental states might be changed by ownership and long-term use (Sharkey et al., 2017).

4.2. Deception

Related to the latter point, another ethical concern arises, which is the robot’s potential emotional deception. This could occur if the user perceives that the robot possesses real emotions (Sung et al., 2007). As already mentioned, deception is one of the most discussed and controversial issue related to social robotics. In this specific case, the human vulnerability to be deceived can be fuelled by the strong human desire to be loved, cared for, and to have companionship [39]. When equipped with AEI, sex robots’ deceiving potential would increase significantly, with the possibility of creating unrealistic expectations regarding its capabilities (e.g., being in love). Such an enhanced capacity for deception might influence the nature of the relationship potentially formed with the robot, by intensifying the likelihood to create more complex and intimate relationships as well as making it easier for the user to fall in love with it. Specifically, by enabling sex robots with AEI, such robots would be able to lead users into believing that the relationship is two sided, that there are genuine feelings involved, and that the sexual pleasure is truly felt. Consequently, the human susceptibility to deception might intensify, as users could be even more inclined to disregard evidence against the robot’s deception, as well as develop emotional attachment based on this perception.

As Nyholm and Frank (Nyholm & Frank, 2019) argue, such deceptive behaviours are problematic from a Kantian perspective on respect for persons as end in themselves. Briefly, when sex robots deceive users into believing that they possess real emotions and are capable of forming reciprocal emotional connections, they may undermine the user’s autonomy and

dignity. Furthermore, the gravity of some of the ethical concerns discussed above would increase, including the loss of what some authors consider to be more meaningful relationships with other humans, as well as the risk of social isolation.

It has been argued that deceiving capacity should be limited in robots. By preventing users from mistakenly believing that the robot is capable of experiencing emotions such as love, care, or physical pleasure, the affective connection with the robot would be reduced, as well as the potentiality of the ethical concerns involved. For this reason, some authors argue that deceptive design in technologies should not be allowed, and that “their machine nature should be transparent” (Boden et al., 2011). However, if a sex robot’s emotional limitations were explicitly stated, it could diminish its appeal and potentially undermine the purposes for which it was created – i.e., the perceived value of the robot in fulfilling the user’s need for love and sex, as well as companionships, could be significantly lessened, and users could be less likely to engage with it.

A more nuanced suggestion comes from Nyholm and Frank (Nyholm & Frank, 2019), as they propose requiring sex robots to continuously emit a signal, to remind users that they are just machines and do not possess the capability to feel or have emotions. They argue that this approach would reduce, however not eliminate, robots’ deceptive potential and, once again, the possibility and strength of emotional bonds that users might develop towards them. However, evidence suggests that such a strategy might not always be effective. Humans could continue to anthropomorphise these robots, project their own mentality onto them, and form what seems like deep emotional relationships, despite their limited abilities (Breazeal, 2004; Healy, 2022; Nyholm, 2023).

4.3. Societal impact

Further concerns regard the impact AEI sex robots could have on the society. For instance, the widespread adoption of emotionally intelligent sex robots could lead to a decrease in the number of human partnerships and marriages, with consequent changes in population dynamics. As briefly touched upon earlier, they could also affect how human relationships are perceived and valued. As people become more accustomed to robots that fulfil their emotional, social and sexual needs, they might develop unrealistic expectations for human

relationships. This could lead to dissatisfaction when human partners are unable to meet these expectations, ultimately affecting the quality and longevity of human relationships.

4.4. Manipulation

In addition, the possibility for AEI to be used to exploit users' emotional vulnerabilities is also a concern, with the consequent risk of manipulation by third parties (e.g., designers or the owning company) (Nyholm & Frank, 2019). AEI sex robots indeed specifically target human emotions and have greater capacity than simple sex robots to emotionally engage with users, potentially leading to the influence or even control over their emotions and feelings, as well as decision-making processes. As Sullins (Sullins, 2012) argues, designers should recognise certain limits regarding the ethical permissibility of exploiting and manipulating the human psychological tendencies of the prospective users of these machines. This concern becomes even more significant when considering love, as "[l]ove is a powerful emotion and we are easily manipulated by it" (Sullins, 2012).

5. Towards an ethical design of sex robot

Considering that the moral controversies and problems surrounding sex robots would become even more pressing with the introduction of AEI, is it appropriate to discourage its development? Maybe not, as such a conclusion would entail a risk of overlooking the potential *benefits* of AEI in sex robots. It is worth noting that while AEI might aggravate ethical worries regarding sex robots, it could also significantly enhance opportunities to improve users' happiness and wellbeing.

For instance, AEI in sex robots may increase sexual satisfaction, as well as provide a more gratifying solution for companionship for lonely people and a more effective comprehensive sexual education. Also, women specifically may benefit from their use. AEI in sex robots can be programmed to promote respectful and consent-based interactions, which might help in reducing objectification and harmful sexual behaviours towards women. AEI can also have therapeutic uses for survivors of rape and sexual assault, by providing a controlled environment where individuals can rebuild their comfort and confidence with intimacy at their own pace.

In addition, by carefully tailoring their interactions to the user's emotion, AEI sex robots could monitor their emotional health and wellbeing. If they detect patterns or sign of emotional distress, they could provide appropriate support or recommendations, such as seeking professional help when necessary. Given the intimacy between the user and the robot, there is a greater chance that the user will follow the robot's advice, ultimately deriving greater benefits from their interactions. On top of that, in the following, I argue that AEI has the potential to tackle the moral issues related to the robot's deception and the resulting ability to evoke feelings of affection in users.

5.1. Monitoring emotional attachment

AEI could enable sex robots to recognise when a relationship with their users has become overly close. By incorporating mechanisms that monitor the emotional closeness between robots and their users, sex robots could take appropriate measures to create emotional distance when they detect that the user is forming excessive attachment and/or dependency. This would help to prevent the robot from inadvertently conveying a sense of greater intimacy or mutual connection than what is actually present, thereby mitigating the human tendency and vulnerability to be deceived. A similar argument has been suggested by Wagner et al. (2018), as they claim that AEI's capabilities "to generate information about the person's attentive state and make behavioral predictions" could be used to aid users to avoid forming excessive attachment to, and inappropriate trust in, the machine.

5.2. Personal growth and skill development

Furthermore, building upon the perspectives of Viik (2020) and Nyholm and Frank (Nyholm & Frank, 2019), I propose that AEI sex robots have the potential to support individuals in improving their emotional and social domains. Given their emotional and social intelligence, they could facilitate "personal self-actualisation, identity construction, character improvement, useful habit acquisition, and relationship building" (Viik, 2020). For instance, AEI sex robots could be designed to facilitate the development of emotional skills in users, such as empathy, emotional regulation, and effective communication. Users who wish to engage in this aspect of the robot's functionality, can do so through, e.g., role-playing scenarios in which the robot acts out various social or emotional situations. Similarly, AEI sex robot can assist with social-skill building, by helping users develop skills such as active listening

and conflict resolution, which can have positive effects on users' relationships with other humans.

5.3. AEI sex robots as transition tools

AEI could also be implemented to encourage human social interaction, consequently helping users build their social network and reduce reliance on the robot for emotional and social needs. While providing emotional support and companionship, therefore, designing the robot to also encourage users to maintain and seek meaningful human connections would help to ensure that robots complement rather than replace human relationships. As a result, sex robots could be seen as “a sort of therapy or transition tool” (Nyholm & Frank, 2019) for enabling users to develop and strengthen meaningful and fulfilling relationships with other people, and mitigating the risk of social isolation and loneliness.

In this regard, an interesting and relevant example of how sex robots can assist users in their human relationships is presented in the documentary *Hi, AI* (Friedman et al., 2022). They describe the case of Chuck, a man from Texas who attempted to have a romantic relationship with the humanoid sex robot Harmony. As they explain, through his relationship with Harmony, Chuck was able to regain confidence in his ability to form emotional connections with others and eventually began dating human women again. Harmony is not as emotionally advanced as the robots discussed in this paper. Nonetheless, such a sex robot has been able to have a profound impact on Chuck's mental, emotional and social wellbeing, and we might expect that more sophisticated emotionally intelligent robots will have more profound, effective, and beneficial results⁵.

To conclude, by incorporating AEI into the technology of sex robots, we may be able to reduce the risk of deception and social isolation. The use of AEI offers a significant advantage in detecting users' emotions and minimising the potentially negative ethical and social impact,

⁵ It is worth mentioning another intriguing case, which comes from Davecat's relationship with his sex doll, named Sidore. Instead of shifting from relationships with sex dolls to those with humans, as in the case of Chuck, Davecat has been “married” to Sidore for over 20 years (Boden et al., 2011). This example might seem to contradict my argument that sex robots could serve as transitional tools towards human-human relationships. However, given that the sex doll in question does not possess any degree of artificial intelligence, including AEI, Davecat's case does not present a direct objection to my argument. Instead, it highlights how human tend to anthropomorphise humanoid inanimate objects, and incorporating AEI could potentially act as a countermeasure to reduce such tendencies.

while enhancing the benefits that individuals can gain from using sex robots, both in terms of sexual health and general wellbeing.

6. Conclusion

The use of sex robots could have major consequences on human life. Caution is imperative when introducing such new technologies that are potentially disruptive for the individual and society as a whole. Nevertheless, it is crucial to recognise the potential benefits that could be derived from them. As argued in this article, prohibiting the development of sex robots would overlook their potential positive impact (Whitby, 2011). Some might even argue that their creation is ethically justified, or even required, given their potential to fulfil the sexual rights of people who might otherwise face barriers to accessing sexual experiences (Friedman et al., 2022). Furthermore, a prohibition of these systems would be difficult to enforce (Borenstein & Arkin, 2019). A more effective approach would be to develop sex robots in such a way that maximises their benefits and minimises potential risks.

AEI could help us towards achieving this goal, by creating sex robots not to replace human contact but to function as “useful and joyful” complements, with positive impact in terms of sex education, sexual therapy and sexual wellbeing (Döring & Pöschl, 2018). As Whitby (2011) argues, technology is not inherently good, bad or neutral; rather, “[i]t is always all three”, depending on how we choose to use it. Our goal is to assure that, through informed, cautious and thoughtful discussions, the right decisions are taken (Whitby, 2011).

Many other questions remain unexplored and have not been addressed in this article. The issue of sexual consent with sex robots has often been discussed (Frank & Nyholm, 2017; Sharkey et al., 2017), and it is worth considering whether the introduction of AEI would make a difference in this matter. We might also question whether the introduction of sex robots with AEI could affect relationship stability, potentially impacting or leading to alternative family structures, such as families that include sex robots as companions or caretakers. This might have further consequences on children’s upbringing and parenting responsibilities, as well as influencing the child’s health and wellbeing, and their relationships with parents.

As is evident, the lack of empirical data is a clear impediment to our understanding of the philosophical and ethical assumptions regarding the effects of sex with AEI- enhanced robots. My suggestion is, therefore, to approach this topic with open minds (and hearts).

REFERENCES

- Airenti, G. (2015). The cognitive bases of anthropomorphism: from relatedness to empathy. *International Journal of Social Robotics*, 7(1), 117-127.
- Beschorner, T., & Krause, F. (2017). Dolores and robot sex: fragments of non-anthropocentric ethics. International Conference on Love and Sex with Robots,
- Beschorner, T., & Krause, F. (2018). Dolores and robot sex: fragments of non-anthropocentric ethics. Love and Sex with Robots: Third International Conference, LSR 2017, London, UK, December 19-20, 2017, Revised Selected Papers 3,
- Boden, M., Bryson, J., Caldwell, D., Dautenhahn, K., Edwards, L., Kember, S., Newman, P., Parry, V., Pegman, G., Rodden, T., Sorell, T., Wallis, M., Whitby, B., & Winfield, A. (2011). *Principles of robotics*. The United Kingdom's Engineering and Physical Sciences Research Council (EPSRC).
- Borenstein, J., & Arkin, R. (2019). Robots, ethics, and intimacy: the need for scientific research. *On the Cognitive, Ethical, and Scientific Dimensions of Artificial Intelligence: Themes from IACAP 2016*, 299-309.
- Breazeal, C. (2004). *Designing sociable robots*. MIT press.
- Damasio, A. (1994). Descartes' error: Emotion, rationality and the human brain. *New York: Putnam*, 352.
- Danaher, J. (2017). The Symbolic-Consequences Argument in the Sex Robot Debate. In J. Danaher & N. McArthur (Eds.), *Robot sex: Social and ethical implications*. MIT Press.
- Danaher, J. (2020). Robot Betrayal: a guide to the ethics of robotic deception. *Ethics and Information Technology*, 22(2), 117-128.
- Danaher, J., Earp, B., & Sandberg, A. (2017). Should we campaign against sex robots?
- Di Nucci, E. (2017). Sex robots and the rights of the disabled.
- Döring, N., & Pöschl, S. (2018). Sex toys, sex dolls, sex robots: Our under-researched bed-fellows. *Sexologies*, 27(3), e51-e55.
- Dubé, S., & Anctil, D. (2021). Foundations of erotobotics. *International Journal of Social Robotics*, 13, 1205-1233. <https://doi.org/10.1007/s12369-020-00706-0>
- Fong, T., Nourbakhsh, I., & Dautenhahn, K. (2003). A survey of socially interactive robots. *Robotics and autonomous systems*, 42(3-4), 143-166.
- Fosch-Villaronga, E., & Poulsen, A. (2020). Sex care robots: exploring the potential use of sexual robot technologies for disabled and elder care. *Paladyn, Journal of Behavioral Robotics*, 11(1), 1-18.
- Frank, L., & Nyholm, S. (2017). Robot sex and consent: Is consent to sex between a robot and a human conceivable, possible, and desirable? *Artificial intelligence and law*, 25, 305-323.
- Friedman, C., Nyholm, S., & Frank, L. (2022). Emotional Embodiment in Humanoid Sex and Love Robots. *Social Robotics and the Good Life: The Normative Side of Forming Emotional Bonds With Robots*, 233.
- Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of management annals*, 14(2), 627-660.
- Goleman, D. (1995). *Emotional intelligence*. Bantam Books, Inc.
- Healy, P. (2022). Social robots as partners? *AI & SOCIETY*, 1-8.
- Lerner, J. S., Li, Y., Valdesolo, P., & Kassam, K. S. (2015). Emotion and decision making. *Annual review of psychology*, 66, 799-823.

- Mahar, E. A., Mintz, L. B., & Akers, B. M. (2020). Orgasm equality: Scientific findings and societal implications. *Current Sexual Health Reports*, 12, 24-32.
- Matthias, A. (2015). Robot lies in health care: When is deception morally permissible? *Kennedy Institute of Ethics Journal*, 25(2), 169-162.
- McArthur, N. (2017). The Case for Sexbots. In J. Danaher & N. McArthur (Eds.), *Robot sex: Social and ethical implications* (pp. 0). The MIT Press. <https://doi.org/10.7551/mitpress/10718.003.0007>
- Migotti, M., & Wyatt, N. (2017). On the very idea of sex with robots.
- Nascimento, E. C. C., Silva, E. d., & Siquiera-Batista, R. (2018). The 'use' of sex robots: A bioethical issue. *Asian Bioethics Review*, 10(3), 231-240.
- Nyholm, S. (2022). The ethics of humanoid sex robots. In *The Routledge Handbook of Philosophy of Sex and Sexuality* (pp. 574-585). Routledge.
- Nyholm, S. (2023). Robotic animism: the Ethics of attributing minds and personality to Robots with Artificial Intelligence. In *Animism and Philosophy of Religion* (pp. 313-340). Springer.
- Nyholm, S., Danaher, J., & Earp, B. D. (2022). The technological future of love. In *Philosophy of love in the past, present, and future* (pp. 224-239). Routledge.
- Nyholm, S., & Frank, L. E. (2019). It loves me, it loves me not: is it morally problematic to design sex robots that appear to love their owners? *Techne: Research in Philosophy & Technology*, 23(3).
- Pan American Health Organisation, W. H. O. (2000). (Promotion of sexual health: recommendations for action, Issue.
- Picard, R. W. (1999). Affective Computing for HCI. HCI (1),
- Richardson, K. (2016). The asymmetrical 'relationship' parallels between prostitution and the development of sex robots. *Acm Sigcas Computers and Society*, 45(3), 290-293.
- Sætra, H. S. (2022). Loving robots changing love: Towards a practical deficiency-love. *Journal of future robot life*, 3(2), 109-127.
- Schuller, D., & Schuller, B. W. (2018). The age of artificial emotional intelligence. *Computer*, 51(9), 38-46.
- Sharkey, N., Wynsberghe, A. v., Robbins, S., & Hancock, E. (2017). Our Sexual Future with Robots. A Foundation for Responsible Robotics Consultation Report. In: The Hague.
- Sullins, J. P. (2012). Robots, love, and sex: the ethics of building a love machine. *IEEE Transactions on Affective Computing*, 3(4), 398-409.
- Sung, J.-Y., Guo, L., Grinter, R. E., & Christensen, H. I. (2007). "My Roomba is Rambo": intimate home appliances. UbiComp 2007: Ubiquitous Computing: 9th International Conference, UbiComp 2007, Innsbruck, Austria, September 16-19, 2007. Proceedings 9,
- Turkle, S. (2011). Alone together: Why we expect more from technology and less from each other. Basic Books.
- Viiik, T. (2020). Falling in love with robots: a phenomenological study of experiencing technological alterities. *Paladyn, Journal of Behavioral Robotics*, 11(1), 52-65.
- Wagner, A. R., Borenstein, J., & Howard, A. (2018). Overtrust in the robotic age. *Communications of the ACM*, 61(9), 22-24.
- Whitby, B. (2011). Do You Want a Robot Lover? The Ethics of Caring Technologies. *Robot ethics: The ethical and social implications of robotics*, 233.

World Health Organization. (2013). *Global and regional estimates of violence against women: prevalence and health effects of intimate partner violence and non-partner sexual violence*. <https://www.who.int/publications/i/item/9789241564625>

Article 5

A qualitative study on love robots' potential in addressing involuntary singleness among women

1. Introduction

Artificial emotional intelligence (AEI) systems are rapidly moving from labs into everyday products, including AI companions and early visions of robotic partners (Sica, 2023). Major technology firms now openly pitch AI friendship – and even romantic companionship – as ways to meet social needs (Wong, 2025). In this paper, love robots refer to advanced machines designed to provide emotional and/or physical intimacy and companionship by recognizing and responding to human emotions through AEI. As these systems become thinkable – and buildable – the question is not only whether they can work technically, but whether they could help with real social problems (Sica & Sætra, 2025).

One such problem is involuntary singleness: people who want a romantic relationship but struggle to establish one due to external factors hindering finding or remaining in a relationship (Apostolou et al., 2019). Although often framed as a male issue, it also affects women (Nygård, Ivanova and Fischer, 2024) and carries social and psychological costs (Adamczyk, 2017), which makes the promise of AEI companionship especially important. Our study asks whether – and under what conditions – AEI love robots could be experienced as viable companions to women who identify as involuntary single. We use a qualitative design with young women in Norway to explore how participants imagine the characteristics and design of such robots, what they see as possibilities and limits for romantic companionship, and the hopes, concerns, and expectations they attach to this technology.

By focusing on women, the study serves several purposes. First, it challenges the stereotype that only men are interested in technologies such as sex and love robots. Second, women might face different societal judgements or ethical concerns compared to men in pursuing

artificial relationships, including concerns about stigma and how gender norms shape women's acceptance of such technologies. Exploring these issues can provide valuable insights into how gender dynamics affect the acceptance and integration of such technologies in society. Third, it not only contributes to a more inclusive understanding of involuntary singleness but also ensures that technological solutions like love robots are developed with a diverse user base in mind, moving beyond the traditional emphasis on male users. As the perspectives of women have often been overlooked in the sex and love robot industry, it is especially important to prioritise their voices and mobilise awareness about their potential contributions to an ethical development in future human-robot relationships (HRRs).

Therefore, this study aims to address research gaps, challenge stereotypes, and contribute to the development of feminist technologies that are sensitive to the needs and experiences of all users, regardless of gender. As both AEI and love robots are still emerging technologies, the insights gathered will inform future developments in their design and contribute to broader ethical discussions in their role in human relationships.

1. Background

1.1. Involuntary singleness

The desire to form meaningful and enduring relationships with others is a fundamental part of human nature. Close social connections are essential for both physical and psychological wellbeing, and their absence can have significant negative effects (Baumeister & Leary, 2017). Societal expectations of dating and marriage further amplify the intrinsic need for connection, potentially leading to the fear of being single – a state characterised by anxiety or distress about the prospect or reality of being without a romantic partner (Spielmann et al., 2013; Zhong & Huang, 2024). For individuals who desire a romantic relationship but struggle to establish one, a condition known as involuntary singleness, this can result in frustration, decreased self-esteem, and social isolation (Adamczyk, 2017).

While not equivalent to loneliness, involuntary singleness can contribute to it, as the lack of intimate companionship may exacerbate feelings of social disconnection (Apostolou et al., 2019). Loneliness is defined as a subjective state of “distressing, depressing, dehumanizing, detached feelings” for a discrepancy between the person's social needs and their actual social

life (Killeen, 1998). It is recognized as a pervasive social and psychological issue with significant negative impact on both mental and physical health, increasing the risks of morbidity and mortality (Heinrich & Gullone, 2006). Addressing loneliness may therefore substantially improve overall wellbeing and reduce psychosomatic diseases. To this goal, De Freitas et al. (2024) argue that high-quality, one-on-one interactions are particularly effective in reducing feelings of loneliness. However, lonely individuals often face social anxieties or dysfunctional behaviours that hinder their ability to form satisfying relationships (Jones et al., 1983). The potential for AI companions to alleviate loneliness has been proposed as a solution, as their ability of mimicking human-like social interactions may be more successful in fulfilling individuals' emotional and social needs than other forms of technology (De Freitas et al., 2024).

Involuntary singleness is often framed as a predominantly male problem. Much of the existing literature focuses on incels – individuals, predominantly males, who experience involuntary celibacy and struggle to find sex and/or romantic partner. The incel phenomenon has evolved into an online movement frequently linked to violence, far-right political ideologies, misogyny, and anti-feminist rhetoric (Ging, 2019; Jones, 2020; O'Donnell & Shor, 2022; Sparks et al., 2022; Thorburn et al., 2023). Far less attention is paid to women, referred to as femcels (Johanssen & Kay, 2024) or female insings (Larsen & Kennair, 2024). This disparity is reinforced by narratives that women are advantaged in contemporary dating – due to gender-equality shifts and platforms like Tinder – and thus have higher “sexual market value” (Johanssen & Kay, 2024; Larsen & Kennair, 2024; Ndlela & Tanner, 2018; Kay, 2021).

Emerging evidence, however, may challenge this gendered understanding. In Norway, Nygård, Ivanova and Fischer (2024) found that men and women equally experience involuntary singleness, and the underlying mechanisms are far more complex than commonly portrayed. One surprising discovery was that single men reported higher rates of sexual satisfaction than single women. This challenges the notion that involuntary singleness and celibacy are primarily male issues, and highlights that women, too, may experience significant barriers to intimacy and sexual fulfilment.

Given these findings, there is a need to explore potential solutions that address involuntary singleness among women. In the next section, we introduce AEI love robots as one such possibility

1.2. AEI love robots

AEI love robots are a type of social robots designed for intimate companionship, that use AEI to enable emotionally responsive human-robot interaction (HRI). This section situates them within the broader field – moving from social robots to sex robots to love robots –, explains AEI's role in emotional engagement, and briefly outlines key ethical concerns.

Advancements in technology have opened new avenues for addressing social isolation and involuntary singleness. Social robots, designed to engage with humans on a social level, have been introduced as companions, providing interaction through conversation, gestures, and emotional responses (Breazeal, 2003). While initially focused on providing companionship for lonely people such as the elderly (Broekens et al., 2009; Gasteiger et al., 2021; Pu et al., 2018; Wada & Shibata, 2007), the potential applications of such robots are expanding. For instance, sex robots have been proposed as a possible intervention for individuals who have difficulty forming intimate human relationships (Danaher et al., 2017; Jecker, 2021; Levy, 2007; McArthur, 2017). Sex robots are humanoid social robots that perform actions contributing directly and primarily towards “improvement in the satisfaction of the sexual needs of a user” (Fosch-Villaronga & Poulsen, 2020). In contrast to sex dolls, current sex robots are equipped with sensors and actuators, and possess some degree of artificial intelligence – i.e., they are capable of interacting with their environment.

While the term “sex robot” emphasise physical intimacy, it does not fully capture the broader vision for these technologies. Developers are now striving to create more advanced and complex machines that offer more than just physical pleasure. They aim to develop robots capable of interacting with users on a deeper, affectionate level (Frank & Nyholm, 2017). The goal is to establish an emotional connection with the robot, hoping users will perceive it as a “true companion,” which could be particularly beneficial for individuals without the opportunity to have a human partner (Danaher & McArthur, 2017).

To better reflect this multifaceted purpose, we adopt the term “love robots” in this study (also used in: Döring & Poeschl, 2019; Nordmo et al., 2020; Nyholm & Frank, 2017; Sætra, 2022). This terminology underscores their role in fostering deep, affectionate, and supportive relationships with users, moving beyond the limitations associated with the term sex robots. Central to this vision is the integration of AEI. AEI belongs to the broader research field of affective computing, which aims to make HRI more believable, relatable, and effective by incorporating emotions into social robotic design (Picard, 2000). Schuller and Schuller (2018) define AEI as a technology possessing three key functions. First, emotion recognition, allows AEI to detect and recognise users’ emotional states. As people express their emotions through different modalities, AEI robots must employ a multimodal emotion recognition system, including facial expressions, heart rate, Galvanic skin response, vocal features, and body gestures. The second, emotion generation, consists in the ability to express emotions in response to users through expressions, body posture, and voice intonation. And the last, emotion augmentation, allows the system to guide behaviour and decision making based on internal simulated states. These capabilities, especially emotion recognition and generation, help robots to appear emotionally responsive, which can help foster a sense of connection, companionship, and support in users.

1.3. Concerns and controversial issues

Despite the anticipated benefits of love robots for people facing difficulties in finding partners, their development and use present significant ethical objections. These objections include concerns about the authenticity of love, since love is more than behaviour or simulation and requires genuine other-regarding concern, whereas robots can at best provide “deficiency love” (Sullins, 2012; Nyholm & Frank, 2017; Sætra, 2022). Authors also warn about the erosion of human sociality, as robotic companionship may foster isolation and cannot satisfy humans’ species-specific social needs (Whitby, 2011; Sullins, 2012; Richardson, 2019). Another issue is the illusion of reciprocity, whereby the appearance of mutual affection could deter individuals in pursuing genuine human connections (Scheutz, 2009). Relatedly, love robots risk becoming an “easy alternative” to human relationships, reducing human’s motivation to form authentic partnerships and hindering moral and social development, since robots cannot replicate the rich and valuable dimensions of human ties (Scheutz & Arnold, 2016; Sullins, 2012; Kroes et

al., 2008; Carvalho Nascimento et al., 2018; Nyholm & Frank, 2019). Critics further highlight the danger of loneliness and exploitation, as reliance on artificial companionship may exacerbate isolation while enabling companies to profit from emotionally vulnerable users. Such relationships are especially harmful because they lack emotional mechanisms like empathy and guilt that typically regulate human interaction (Nyholm & Frank, 2019; Carvalho Nascimento et al., 2018; Scheutz, 2009). Also, some argue that love robots with advanced emotional intelligence could threaten family stability, reshape relationship structures and potentially affect parenting and children's wellbeing bonds (Sica, 2023).

A final concern is the question of the moral status of such robots (Sica & Sætra, 2024). Some authors argue that robots cannot possess moral status, as they lack essential qualities such as consciousness, which are typically seen as necessary for moral considerations (Mosakas, 2021). Conversely, relational accounts argue that robots warrant some moral consideration not because they have mental states or consciousness, but because of how they appear and relate to us. When robots look or behave like humans, we tend to treat them as if they were moral beings, making the social relation morally significant (Coeckelbergh, 2010, 2014; Gunkel, 2018). Nyholm (2020) holds that some moral consideration is needed for robots that look like human or non-human animals because of how we treat them can reflect and shape our moral character and express our responsibility as human agents to act ethically. Beschorner and Krause (2017) add that the social role of robots shaped by the humanoid aspect, the intimacy HRI reflect, customization, and framing – can influence moral perceptions. Since morality is a cultural product, societal perceptions may expand future spheres of moral consideration to include non-living objects (Beschorner & Krause, 2017).

As we seek to explore whether love robots could be a partial solution to involuntary singleness in the future, we distance our inquiry from historical and actual robots (such as sex robots) (Danaher, 2017; Döring & Pölsch, 2018; Richardson, 2016). We instead envision a future where this technology has advanced further, and in which the problems related to, for example, their stereotypical appearance and overly sexualised nature have been addressed. Indeed, we imagine that love robots are designed inclusively, representing all genders and sexual preferences, and intentionally avoiding exaggerated body shapes. By prioritising diverse representations and inclusivity in their design, we avoid focusing this study on the traditional

narratives that have often been a source of ethical contention and rather focus it on something different.

2. Methods

This study adopts a qualitative research design that directly engages with the primary focus of the research: involuntary single women. Through in-depth interviews, this design seeks to uncover insights into participants' personal experiences, attitudes, and beliefs towards love robots and developing relationships with them. This approach ensures that the end-users' voices are deeply integrated into technology's development. The underlying premise is that none is better positioned than involuntary single women themselves to articulate their needs and perspectives. They hold situated knowledge of their lives and a unique understanding of how technology can or cannot support them in fulfilling these needs.

2.1. Participants

Inclusion criteria for the study were: (1) self-identification as involuntary single, (2) current residence in Norway, and (3) age between 20-40. Involuntary singleness was assessed through self-identification: participants themselves defined and reported this status. Participants were recruited primarily through posts on relevant Facebook groups and a few via direct invitations. They were not financially compensated.

We recruited twelve young women, identifying themselves as involuntary single women. All participants were originally from outside Norway but were currently living in Oslo or other urban areas. Although migration background was not part of the inclusion criteria, the women who responded to the recruitment call and volunteered to participate all happened to be immigrants living in Norway. Most were full-time workers, with a few still pursuing their studies. Most were familiar with AI in general, either through work or study, but almost none had prior experience with AEI and love robot technologies. All had previous experiences with romantic relationships, although not all of these had included sexual experiences. Relationship histories varied: some had recently come out of breakups, some had been single for a longer period, some actively used dating apps without establishing serious relationships, and a few had never experienced a long-term partnership. Most participants identified as heterosexual, with two identifying as bisexual. Some participants knew more about the relevant technology

than others, and some spoke more easily about their experiences and thoughts. All participants were given fictitious names to maintain their anonymity (see Appendix 1).

2.2. Data gathering

Data was gathered through individual semi-structures interviews conducted by the first author, either in person at a university office or over video call. All interviews were conducted in English and lasted between 45 and 60 minutes. The flexibility of the semi-structured format allowed the interviewer to ask follow-up and probing questions to help participants voice their thoughts (see Appendix 2 for the interview guide). Before the interview, participants provided informed consent, either by signing a form on the day of the interview or returning it by email in advance. All interviews were audio-recorded with the participants' permission. During the interviews, the interviewer also took brief notes, mainly as reminders for follow-up questions without interrupting the flow.

Given that most participants were unfamiliar with emerging technologies such as AEI or love robots, the authors considered how best to introduce these concepts during the interview. One key decision was to avoid framing love robots primarily in terms of sex robots, a more common concept among participants, to reduce bias and avoid unnecessary associations. For this reason, images of existing sex robots were sometimes shown only at the end of the interview, to allow participants to first imagine love robots freely based on the definitions provided. To facilitate understanding, the interviewer initially described the broader category of social robots, showing pictures of currently available robots used in healthcare, companionship, or costumer service. The concept of AEI was then explained in simple, descriptive terms – as the capacity to recognize, simulate, and respond to human emotions. Participants were then introduced to the idea of a love robot equipped with such emotional capacities. From this point, the interview was framed as a thought experiment: participants were invited to imagine and describe whatever comes to mind in response to the concept. Potential benefits and risks were not emphasized in the description. Instead, participants themselves were encouraged to articulate both positive and negative associations.

The interviews were transcribed and anonymized by the interviewer. Transcription was initially conducted with Autotekst, developed by the University of Oslo and based on OpenAI's

Whisper system, and then checked for accuracy by listening to the recordings while reviewing the text.

2.3. Data analysis

The interviews generated rich and complex data material which we analysed inductively through a thematic analysis framework (Braun & Clarke, 2021, 2022). Interviews were coded in the qualitative data analysis software NVivo (Hilal & Alabri, 2013). The analysis was inductive and data-driven, meaning that themes were developed directly from the material. The six phases of thematic analysis guided our work:

1. Familiarization – Several authors first read the anonymised transcribed interviews in full to gain an overall sense of the material and to note initial impressions.
2. Generating initial codes – The first author coded the material in NVivo (Hilal & Alabri, 2013). During this process, memos were written to capture reflections, emergent ideas, and connections across interviews. The codes were marked with different colours and tabulated by the first author. Codes and memos were repeatedly revisited, refined, and reorganized in light of ongoing reading of the transcripts. Examples of such inductive coding are *Differences between love robot and human partner*, *Ideal partner*, *Customisation*, *Therapist*, *Safety*, *Family*, and *Societal acceptance*.
3. Searching for themes – Codes were clustered into broader candidate themes and thematic maps were developed.
4. Reviewing themes – The authors repeatedly met to discuss overlaps, divergences, and potential reorganisations of codes and themes. Themes were developed through multiple rounds of review: assessing whether they captured patterned meaning across data, considering how they related to each other, and mapping them in relation to the research questions. This collaborative process supported reflexivity by surfacing the influence of each researcher's perspectives.
5. Defining and naming themes – Candidate themes were refined, merged, or split to ensure internal coherence and external distinction.

6. Producing the report – The final themes represent interpretative patterns that capture recurrent meanings across the dataset. The results are presented below together with illustrative quotes.

3. Results

3.1. What is a love robot? Initial impressions of the technology

As discussed above, most participants, however knowledgeable about AI, had little prior familiarity with the specific concepts of AEI and love robots. As a result, it was particularly interesting to observe how they imagined these technologies during the interviews. Many drew inspiration from popular media – especially in films *Her* and *Ex machina*, and the tv series *Black mirror* – to visualise what love robots and emotionally responsive AI might be like. Such initial impressions shaped participants' initial thoughts on the emotional and relational potential of love robots, as well as other aspects such as their societal acceptance – topics that will be explored in more detail later in the article.

3.1.1. *The promise and problem of perfection*

A recurring theme was the contradiction between the perceived advantages of love robots and their lack of genuine human qualities like unpredictability and organic emotional depth. While many described the love robot as an ideal partner, they also questioned whether something so “perfect” and programmable could ever truly improve or substitute a human relationship.

Several participants appreciated the idea of having a partner who could be tailored to meet their specific needs – whether emotional or physical. Julie, for example, imagined the robot learning from ongoing interactions and offering increasingly empathetic responses over time. Others highlighted the comfort of avoiding the emotional risks and disappointments inherent in human relationships. Anna appreciated the idea of a companion who would genuinely listen and “remember everything I say”. The contrast with past relationships with human partners was often pointed out, as by Linda who believed a robot would “understand me emotionally” better than a person could.

Participants also expressed positive attitudes towards the possibility of customizing a love robot's physical traits. Most favoured robots that looked as human as possible and even envisioned life-course features (ageing, illness) to sustain companionship over time. By contrast, a minority perceived the idea of a human-like robot unsettling, as Sabrina: "I would feel a bit taken aback from it because I would see that they have the humanoid form, but I know that deep down inside they're not human." To avoid the uncanny feeling, several women considered the possibility of designing the love robot to look like something else completely. For instance, Linda played with the idea of getting a fantasy partner: "It would be kind of fun to have it like kind of fantasy, like an elf."

However, this idealisation coexisted with significant reservations. While some participants imagined emotionally rich, supportive relationships with robots, they simultaneously emphasised their artificial nature as a limitation. Maria gave words to this view: "It feels [...] alien, very far-fetched and also not sexual at all. [...] it's like a lab test or something." The human condition – marked by imperfection, change, and unpredictability – was valued as something fundamentally absent from robotic interaction. As Blair noted: "You can't simulate the same uncertainty as you can with an actual human, because humans can be unpredictable. Whereas a computer [...] is something that is actually programmed, even if it's not hard coded directly." In this way, the very perfection that made love robots appealing also rendered them problematic. Several participants expressed discomfort with artificial flawlessness. Even the possibility of programming imperfections would not be enough, according to Linda. Carol also agreed that pre-programmed "humanity" would not resolve the essential artificiality of the experience: "It doesn't matter if it's a physical robot [...] it has to be alive, warmth, softness."

3.1.2. Just a machine

For many, the robot was ultimately viewed as an object, as Linda points out: "I was thinking of it kind of standing on the docking station like a vacuum robot." One participant, Christina, even felt scared by the idea of a love robot, perceived as "completely least human". Beth and Ina reinforced this perception, as it lacks mind or consciousness. Some participants even equated love robots to sex toys, like Karen. In a similar vein, Carol described the love robots as objects of utility or convenience: "He doesn't sit on the couch, he has to be useful."

This objectification further extended to questions of power dynamics and moral status. Anna voices her concerns around this: “Maybe I pay to have this robot, I bought it, I rented it, so it’s obvious that I gave my part and now I take it back from the robot”, concluding that “It’s a robot, so I don’t have to worry about its feelings or its pleasure.” Carol also appreciated the same control: “If I don’t want it, then I don’t feel bad about [...] turning it off.” Furthermore, Emily questioned how far one could take this rationale by asking whether robots would have moral rights and what users would be allowed to do to them. Because none of the participants had direct experience with this technology, their reflections opened broader practical, emotional, and ethical questions – not only about how to treat the robots, but about how instrumental relationships with artificial intelligence might shape their own social, emotional, and social lives.

3.1.3. The risk of displacement

Participants expressed concerns that artificial perfection might distort relational expectations, as love robots being so ideal for the qualities described above, that an actual human being could never compare. As Anna put it: “There’s the risk that it’s so perfect that you say, well, “I’ll never find a better person than him”, so I’ll keep it.” Paradoxically, the same “perfect” qualities were seen by others as potentially beneficial for setting healthier relationship standards, like how it was perceived by Carol: “So, if they had this kind of companion, maybe they wouldn’t choose to be with a man that’s unhealthy for them and have higher standards when it comes to real men.”

However, as robotic relationships start redefining what people expect from a partner, participants worried that the boundary between supplementing and replacing human relationships could blur. Maria worried about the potential consequences of such a shift for human priorities and moral judgement: “I feel when humans choosing some technology over a human being that’s a problem. Are you going to save your phone or are you going to save a kid who’s drowning?” Participants had worries about how the robots could potentially influence their personal priorities and stand in the way of morally right behavior towards other people, as well as towards the robots.

3.1.4. *The potential for “Love”*

The contrasting nature of the love robot – as an “ideal partner”, but ultimately an object – strongly influenced the kind of relationship possibly formed with them. Most of the participants could not imagine spending the rest of their lives with a robot. Karen also states that it would be more out of curiosity than actual romantic hopes: “I don’t think I would be able to do it in the long term [...] but it would be interesting to try.” Beth worried that, however momentarily, “time, money and energy” would be wasted over a robot instead of a human partner, implying the same short-term focus rather than a permanent solution to finding a partner.

Participants also doubted whether genuine romantic feelings could develop when predictability and customizability eliminate the surprise and mutual discovery that characterise human relationships. Beth worried that pre-programming a robot’s personality might take away the exploratory dimension of getting to know a partner, therefore undermining reciprocity and desire. Maria described robotic interactions as transactional rather than relational: “you give me comfort, I take something back [...] I believe that love, true love, [...] it’s not about what you give me.” Karen expressed discomfort with the idea of forming a connection with something designed to reflect their own desires: “it seems to me a thing in which there is a victim and a carnage.” Blair echoed this unsettling control: if you disagree, you can “change them to whatever fits you [...] And then it’s just very dark, I feel, being able to change somebody if you don’t agree with them.”

The perception of artificial affection – programmed rather than freely given – also raised questions about the authenticity of being loved. As Karen put it, “I don’t have the feeling that I am really desired. [...] The robot is programmed to desire you.” Most participants did not consider robots as human enough to fulfill their relationship desires, yet a few challenged the person-machine binary. Beth pointed out that “you can feel love towards other things than humans”, acknowledging the possibility of strong emotional attachment. Emily emphasised how she would stay faithful to a potential robot partner, not wanting to date human partners at the same time: “If I am already feeling something with AI, that’s my relationship. So, the only reason I would start looking for something outside would be because this one is failing.”

3.2. Safety

One of the major concerns expressed by participants regarding love robot was the issue of safety, which emerged across two main dimensions. On one hand, love robots were often described as offering emotional and physical sense of safety. On the other hand, they were perceived as introducing new risks, particularly in terms of data privacy and the intentions of those designing or profiting from the technology. The following two subsection explore these contrasting perceptions of safety. First, at the level of the individual relationship with the robot, and second, in terms of broader external concerns related to data protection and corporate or institutional misuse.

3.2.1. *Physical and emotional safety*

Many participants viewed love robots as offering reassuring alternative to human relationships, highlighting their predictability, emotional consistency, and perceived reliability. Christina noted that “The robot couldn’t hurt me [emotionally]”, while Sabrina appreciated that it “wouldn’t try to wear a mask or try to lure you into the relationship.” The theme of trust also emerged, with robots perceived as trustworthy companions by some of the participants, like Anna: “[with the love robot] you feel like receiving a lot but maybe not giving much. So that, in my opinion, it’s a safe space. [...] Instead, a friend, in general, in my opinion, if he offers you these services, wants something in return, right?”

Additionally, safety was connected to psychological comfort, specifically the freedom to express oneself authentically without fear of social judgement or relational consequences. Participants appreciated the unconditional acceptance offered by robots, contrasting this conditional acceptance commonly found in human relationships. For instance, Anna stated: “In theory, you should always be yourself and loved for who you are.” Love robots were viewed as providing as a safe space for introverted or sexually inexperienced individuals to explore social and intimate situations without embarrassment or discomfort. Blair appreciated the idea of dating someone without the usual social worries: “And maybe some of this awkwardness could go away if you had tried it with a robot, you know?”

In addition to emotional reassurance, physical safety was also emphasized, particularly in contrast to the risks of violence in human relationships. Some participants saw love robots not only as safer alternatives to individuals, but also as potentially contributing to broader social

benefit of reducing violence against women. For instance, Sabrina suggested that wider acceptance and use of love robots might prompt societal reflection on gendered violence, potentially influencing attitudes and behaviours positively: “If this kind of robot becomes more common, people would maybe empathize with it more and try to reflect on their own behaviour.” Further, Julie suggested that love robots could serve as a substitute to prevent sexual violence against women: “We’ll give it to guys so that they don’t rape girls.”

However, participants also expressed concerns regarding potential emotional and physical dangers posed by love robots. Fears included risks from programming errors or deliberate misuse, potentially leading to physical or emotional harm. Specific worries mentioned were technological malfunctions causing physical injury, which Ina brought up: “Let’s say there is a glitch in the system or anything like that, [...] and it has bad intentions at some point...”. Additionally, participants brought up concerns regarding emotional manipulation or possessiveness due to robots’ misinterpreting human emotions, as Blair feared: “These kinds of things can quickly go out the window with AIs that take things quite literally. And even as you feed them different situations and kind of try to nuance the picture for them, then it can spiral out of control.”

3.2.2. Data safety

Participants raised important worries about the risks that these technologies might pose to users’ data privacy. The intimacy of the user-robot relationship that some found reassuring, also raised anxieties about how personal information might be collected, stored, or misused.

For some, the idea that a robot could accumulate highly personal knowledge about their emotions, preferences, and vulnerabilities was deeply disturbing. As Anna explained: “The danger [...] that this person knows a lot about you, your most intimate sides ... and that’s scary, if it goes into the hands of someone you wouldn’t want to have them.” While Carol saw the robot as a potentially more trustworthy confidant than a human - “Like a private diary: If you lock it, it’s locked.” - she also emphasized the uncertainty surrounding data ownership and regulation: “There’s a question of who owns the data. [...] What are the regulations that control this robot?”

Along similar lines, Maria questioned the motivations of companies developing love robots, expressing concern that the emotional bonds users form could be deliberately exploited for

profit: “People who are finding comfort in this intimate connection [...] don’t think that this connection is unfortunately purposefully targeted in order to use their data, use their time, make them purchase things.”

3.3. Love robots in society

This section addresses the broader societal dimensions of love robots, focusing on two areas. The first explores participants’ views on how love robot relationships might be perceived in society, including the stigma and gendered judgements users may face. The second considers participants’ reflections on love robots’ impact on users’ social lives and skills.

3.3.1. *Acceptance*

Every participant expressed concerns about the social stigma associated with being in a relationship with a love robot. Many feared being judged, misunderstood, or excluded if seen in public with such a partner, as Christina: “[With] how society is right now, I think I would be with [the robot] at home, but not outside.” For several participants, love robot relationships were seen as a form of “forbidden love”, difficult to disclose in “our conservative society”, as Maria phrased it.

While some concerns focused on fear of external judgements, others mentioned internal feelings of shame or perceived inadequacy. As Carol put it: “I guess I’m a bit embarrassed to take him outside [...] Because for some reason I feel embarrassed at needing it.” These responses suggest that stigma may stem not only from anticipated social reactions but also from internalised norms about what constitutes legitimate or successful relationships.

Despite these personal reservations, some participants expressed acceptance of others choosing love robots. In a scenario where other people were publicly seen with their love robots, Julie said she would be “curious”, but not judgmental. Emily echoed this: “If you are happy and in love with your robot, be happy and in love with your robot, why bother?” Some believed that wider social use could gradually reduce stigma, like Sabrina: “If people started going out with robots, then I would feel more comfortable [with having one].”

Participants also discussed how societal acceptance might differ depending on gender, revealing conflicting views. Several participants believed that women would face more stigma than men, including Emily: “History shows that we are more judgmental towards women...

[love robots] would probably be more easily accepted for men.” Ina explains this further: “There is double standards [...] If something like this happened to a woman, there will be immediately articles about it [...] For men it’s just, oh, it’s his sexual desire.” Julie added that societal discussions around commercial products related to sexuality are often focused on men, and she saw love robots as a way to make women’s needs more visible.

Others took the opposite view and said that women might face less stigma because people assume they are seeking emotional connection. Sabrina said: “Women seem to be more emotional [...] people would empathize more.” Karen also noted that a “platonic relationship” would be judged more acceptable than a sexual one, pointing to the film *Her* as framing robot relationships as romantic rather than sexual.

An additional element of concern shaping societal acceptance of love robots came from participants’ deeper anxieties about the future of the institution of the family. Several women questioned what would happen to traditional family structures, reproduction, and parenthood if romantic relationships with robots became widespread. Sabrina was quite reserved of having love robots forming or becoming part of the family, such as Sabrina. Anna expressed discomfort with the idea of expanding legal or social recognition to such relationships. At the same time, others pointed out that new forms of reproduction already exist that could address these issues, like Christina: “Nowadays there are opportunities to still get pregnant without having a human partner.” Further, Ina reflected on how robots might support family life, as she imagined robots assisting with childcare and household duties that are typically women’s work. Still, questions remained about the role of robots in parenting and child development. Carol wondered how society would handle responsibilities in non-biological parenting situations: “So, if I had a child with my robot and they’re not the biological father [...] how do we handle [that]?” She further questioned the impact on children by being raised by or with robots: “Do [children] learn to interact with humans differently?”

Finally, participants also raised concerns about wider social tensions. Some warned that love robot relationships could provoke hostility or even violence, like Sabrina: “People get angry and try to stop this from happening and also harm actual humans for being with robots.” Emily echoed this fear: “I think people with love robots can even be killed.” Religious groups or male-dominated communities were mentioned as likely sources of resistance by Ina. Despite these

worries, other participants like Julie imagined more hopeful scenarios in which humans and robots could peacefully coexist.

3.3.2. Social life with a love robot

Participants expressed mixed views regarding the impact of love robots on social skills, interpersonal interactions, and loneliness. While some viewed love robots as useful tools for improving social skills and reducing isolation, others worried that they might ultimately worsen disconnection from meaningful human relationships. Below, these contrasting perspectives are explored in more detail.

Several participants believed that love robots could help people develop or regain social skills, particularly for individuals wishing to reconnect with human partners. Christina suggested that love robots “could be a good way to practice dating skills or talking skills in general.” Some participants specifically highlighted men as likely beneficiaries of this practice, with Linda stating that “men that haven’t spoken that much to women, maybe they will after learn how to get a girlfriend.” Additionally, love robots were seen as tools for promoting emotional intelligence and social development. Carol described them as “in-house therapists”, helping users become more confident and emotionally aware, even after establishing new human relationships.

Specifically for individuals struggling with rejection or low self-esteem, the robot could offer reassurance and a sense of being valued. Anna said repeated rejection can wear people down, and a robot could “give you a little bit of self-confidence, to make you feel appreciated.” Sabrina added that love robots could also provide comfort and unconditional acceptance, enabling users to “feel more accepted” and be authentic without the pressure of conforming to social expectations.

Participants also identified specific groups who could benefit from such robots, including older individuals, infertile women, people with sexually transmitted diseases, individuals with mental health issues, as well as individuals with physical disabilities. For these groups, Karen highlighted: “If society doesn’t work, probably yes, technology can work. It’s a perfect solution on an emotional and sexual level.” Linda suggested use in rehabilitation for people who have been “locked up or isolated for a very long time.” Such reflections aligned with existing uses of social robots already providing companionship and care for similar groups, suggesting that

love robots could extend this support to romantic and intimate needs, like Blair pointed out: “If you’re able to use social robots, then you can definitely switch over to romantic relationships with them.”

Many participants emphasised the emotional benefits of companionship and the possibility of love robots to reduce loneliness. Karen warned robots could push some people further into isolation, but said the effects could be positive “as long as the robot encourages you to have relationships with the outside”. Beth was cautiously optimistic: robots can help you feel less lonely and offer practice in intimacy, but they can also be “a dead end” as they do not replace human relationships. Ina stressed the value of reliable presence: having “someone that is there whenever you need them [...] just talking or things like that ...”

In contrast, some participants also expressed significant concerns that love robots might undermine social development. Karen warned against creating “echo chambers” where users only interact within personalised and unchallenging social environments, potentially diminishing empathy and authentic engagement. As Maria explained: “You cannot discover yourself within your own vacuum. [...] You need to constantly engage with people.”

A common worry was that relying on love robots could reduce the incentive to seek out human contact. Karen expressed doubts that robots could support genuine reintegration into society: “It provides a solution from an emotional point of view, in short terms, but [...] it can simply lead people to remain more isolated, I think.” Others, like Julie, shared similar concerns, describing scenarios where robots replaced daily social interactions.

Linda and others feared that reliance on robots could make future human contact more difficult: “If someone is meeting all of your needs [...], then you might have it harder to get out of your shell and communicate again with the rest of the people.” This extended also to the fear of reinforcing unhealthy relational expectations, particularly among impressionable young people who might internalise unrealistic behavioural norms from robot interactions. As Blair emphasized, “If young users start believing that these things are what is supposed to happen, then they’ll go out in another relationship, in an actual relationship maybe, and then they have [...] to go really reveal [themselves] to somebody and don’t say *no* or something.”

An additional aspect of concern was related to social acceptance. If love robots remain stigmatized, users might further isolate themselves from others. As Beth noted, fear of judgement or marginalization could therefore reinforce feelings of secrecy, shame, and ultimately increase isolation.

Interestingly, some participants observed that these contrasting perspectives on the effects of love robots on social skills and sociability (which can plausibly extend to other aspects explored throughout this article) were not necessarily inherent to the robot's nature but rather depended on how individuals chose to engage with them. Even participants with a generally positive attitude towards love robots pointed out that their use still requires critical consideration and responsible development, like Blair: "You want to be careful that whatever the AI is trained on, it doesn't develop some things that are harmful for the user afterwards." Reflecting on human tendencies, Emily summarised: "We are very good at making good things bad."

4. Discussion

This study challenges prevailing gendered assumptions about involuntary singleness by highlighting that women, too, experience deep frustrations and unmet needs in romantic and sexual spheres. While previous research has disproportionately focused on men as the archetypal involuntary singles, our participants' narratives emphasise that also women face structural and interpersonal barriers to forming relationships. The findings thus call for a more inclusive understanding of singleness that is sensitive to women's experiences.

In this section, we interpret our findings in light of existing research on sexuality, gender, and HRRs. We consider how societal expectations shape women's openness or resistance to artificial companions and reflect on the ontology and ethical implications of these technologies.

4.1. Social norms of dating and sexuality

All participants in our study were young women living in Norway but originally from other countries, who experienced challenges with coming to a foreign dating culture. In this way, they contributed to an outside perspective of Scandinavian social and sexual norms.

The way we understand our own sexuality and identity is socially constructed, with social norms shaping our behavior (Dworkin & O'Sullivan, 2005; Simon, 2017). Previous research has shown how these norms are gendered. Women are expected to behave more passively, to be less sexually assertive (Morrison et al., 2015), and to show less interest in casual sex than men (Farvid et al., 2017). This carries over into attitudes toward sexual technology. In our study, these gendered expectations were reflected in how participants discussed potential stigma over the use of love robots. Many of our participants perceived that embracing such relationships would imply transgressing expectations about how women are supposed to seek and express intimacy.

Some participants described the idea of having a robot partner as liberating, as it allowed them to prioritize their own desires without feeling obligated to satisfy a partner. These reactions reflect an awareness of gendered sexual norms, where women are often expected to manage and prioritise male desire (Ford, 2021; Sanchez et al., 2012) – patterns also observed among Norwegian couples (Myrtveit-Stensrud et al., 2024; Myrtveit-Stensrud et al., 2023). However, others expressed discomfort with the notion of a partner being designed to fulfil only their desires, with some even referring to the robot as a “victim”.

Importantly, participants reflected on how social norms influenced their own comfort with the idea of using love robots. Several women said they would feel uncomfortable being seen in public with a love robot or introducing it to others, while others described a sense of embarrassment at the idea of “needing” a robot to feel loved or understood. This shows how social and sexual competence are closely tied in our culture (Traeen, 2008), with women fearing being labelled as deviant in how they seek and obtain intimacy.

Media representations played a significant role in shaping participants’ perceptions of stigma. As Döring and Poeschl (2019) argue, in the absence of first-hand experience with love robots in the broad population, public understanding is largely shaped by how these technologies are portrayed in both fictional and non-fictional media. Several participants referred to films and series as sources that influenced their views on what relationships with robots might look like, and how these might be socially received. Romantic narratives like the film *Her*, which center on emotional rather than sexual intimacy, were perceived as more socially acceptable, suggesting that emotionally driven connections with AI could be normalized in the public

imagination. In contrast, portrayals that emphasize overt sexual gratification, often involving male users and featured in non-fictional media (Döring & Poeschl, 2019), were seen as more stigmatizing. This differentiation led to gendered interpretations. While some participants believed women would be judged more harshly than men for using love robots – due to prevailing double standards around sexuality and emotional dependency – others noted that women’s use of such technologies might be seen as more acceptable if framed as emotional rather than sexual. These reflections align with broader patterns in popular culture, where traditional sexual scripts are reinforced and internalised by viewers (Dillman Carpentier et al., 2017; Kim et al., 2007) – often positioning men as seekers of sex and women as seekers of connection.

Finally, several participants brought up the cultural expectation that dating should ultimately lead to marriage and children – an outcome they felt could only be achieved through relationships with human partners. Robots were not seen as fulfilling roles such as parenthood or integration into family life. While some authors have questioned whether love robots might eventually influence relationship stability and even reshape family structures (Sica, 2023), such possibilities were not reflected in the participants’ accounts. On the contrary, their views reinforced the belief that, yes, love robots might offer temporary emotional support, but they could not provide a complete or socially accepted alternative to human companionship. Instead, as Sætra (2022) notes, relationships with robots will remain “limited and private”.

4.2. Ontological ambiguity and new ethics

A central theme that emerged from the interviews was the ontological ambiguity of the love robot – whether it should be understood as a person, a tool, or something in between. This ambiguity significantly shaped participants’ perceptions of robots as potential companions or romantic partners. Several participants were uneasy with the idea of forming romantic bonds with entities whose affection, empathy, or desire could be programmed rather than freely expressed. For these women, authentic love was not only about care or attention, but also mutual unpredictability, autonomy, and the possibility of being genuinely chosen by another.

As such, behaviourist or functionalist accounts of love, such as those advanced by Levy (2007), were mostly rejected. Instead, participants’ conceptions of love reflected discussions found in Nyholm and Frank (2017), as they stated that, if a robot is designed to love by default, its

expressions of affection feel inauthentic. As Nyholm and Frank argue, love, to be meaningful, must be chosen and sustained – not built-in or automatic. This distinction was central to participants’ skepticism about whether a robot could ever truly be considered a loving partner.

However, some participants challenged the human-machine binary by pointing to the ways emotional attachment can emerge even when the other is non-human, as echoed by Scheutz (2009), Sullins (2012), and Turkle (2011). Indeed, participants suggested that the strength and authenticity of emotional bonds may depend more on the robot’s perceived responsiveness than on its ontological status, as similarly argued by Coeckelbergh (2010).

In addition, participants worried about important ethical implications that the ontological ambiguity of love robots might lead to, particularly regarding the moral status of robots. For instance, they asked: Can robots be turned off without moral consequences? Similar questions have been explored in the literature. Beschorner and Krause (2017), for example, ask whether acts of violence towards sex robot should be considered immoral. Although most participants rejected the idea of robots possessing moral status, their discussion revealed the emergence of new ethical grey areas in human-robot intimacy, as discussed by Coeckelbergh (2010) and Beschorner and Krause (2017) – especially as robots come to simulate more complex emotional responsiveness with AEI.

If so, an important ethical question that emerged from the interviews concerned which relationship we will eventually come to value more: those with robots or with humans? This reflects a tension around the ethical weight assigned to artificial versus human agents, and possible consequences that such shifts may have for our moral priorities. As Sica and Sætra (2024) argue, these development risk leading to a misallocation of moral attention and resources – potentially elevating the status of robots at the expense of humans and other morally deserving beings. This raises important concerns about how emotional attachment to robots might influence societal values, moral responsibilities, and the future of human relational practices.

4.3. Loneliness

Participants in our study often expressed contrasting perspectives on the potential of love robots to alleviate loneliness. Some saw these robots as a way to ease isolation: a safe place

to practice conversation and dating, build emotional intelligence, and regain self-confidence – especially for people facing repeated rejection, social anxiety, old age, or long period of isolation. This fits with studies on existing social robots used for companionship and therapy, especially with older adults (Gasteiger et al., 2021; Yen et al., 2024). Participants viewed love robots as a support for connection, not a replacement, aligned with Kroes et al. (2008), who argue robots should be designed to enhance users’ social life rather than replace human interactions.

Others predicted the opposite effect. Because of the robot’s emotionally undemanding and less challenging social simulations, participants feared users would withdraw from reciprocal, real-world relationships. Participants framed robots as the easy and convenient option, which could displace human relationships (Kroes et al., 2008; Carvalho Nascimento et al., 2018; Turkle, 2011). They also worried that the perceived “perfection” of robots would make human partners less attractive (Beschoner & Krause, 2017). Even if this does not produce loneliness per se, the prospect of dramatically reduced human connections can decline people’s social ability (Whitby, 2011).

Participants also worried that relationships with love robots might hinder personal growth. If the robot mirrors users’ preferences and personality, there is no room for challenge, disagreement, or contrast – elements they saw as essential to meaningful development (Nyholm & Frank, 2019). They also stressed the inherent asymmetry of such relationships, as users maintain full control. Nyholm and Frank (2019) warn that such imbalance may foster problematic traits in users, such as “self-centeredness and a tyrannical and domineering personality”. Sætra (2022) similarly argues that normalizing unequal HRRs could reshape our moral character and alter our very conception of love.

Another interesting aspect that emerged from our study is that participants’ attitudes towards love robots were often influenced by their experiences with loneliness. Those who expressed a strong desire for physical touch, companionship, emotional connection, or simply the physical presence of another, were typically open towards the idea of being in a relationship with a robot. Instead, those who wished for a romantic partner without feeling lonely necessarily were more sceptical or less inclined to the idea of using such technologies. This pattern can be found in real-life cases, such as that of Davecat, a man who married his sex doll

Sidore. He claims that the reason why he chose to have his sex doll is to overcome feelings of loneliness (Nyholm & Frank, 2019). In this context, participants described loneliness not only as a condition that love robots might alleviate or rather aggravate, but also as a motivating factor for seeking robotic companionship in the first place.

However, the vulnerability inherent in feelings of loneliness raises ethical concerns. Participants worried about the potential financial exploitation of emotionally vulnerable people by the companies producing love robots, as also discussed by Nyholm and Frank (2019), Carvalho Nascimento et al. (2018), and Scheutz (2009).

4.4. Implications for practice

This last section explores possible recommendations for love robot design based on participants' insights. Participants stated that the technology itself is not inherently harmful or beneficial. Instead, whether love robots will ultimately support or undermine wellbeing depends largely on human choices, especially in how these technologies are designed and the intentions behind their creation. As Floridi et al. (2018) argue, while AI can offer significant opportunities, it also brings risks when shaped by harmful intent, greed, or poor regulation. Participants highlighted that robots developed to encourage emotional growth, user autonomy, and social interactions with others – rather than passive compliance or dependence – can help foster more meaningful and ethical forms of companionship. This view is also supported by Sica (2023), who suggests that the same AEI technology, when ethically designed, can contribute to such positive outcomes.

However, these possibilities depend by the motivations driving development. As discussed above, participants expressed concern that commercial interests may prioritise profit over user wellbeing, especially by encouraging emotional dependence or exploiting vulnerable users through manipulative design. They mentioned issues such as invasive data collection, lack of transparency, and poor privacy protections – issues particularly concerning in technologies meant to simulate emotional relationships (Laban & Cross, 2024; Lutz et al., 2019; Lynch, 2021). Such concerns highlighted the importance of transparency, accountability, and clear regulatory frameworks to prevent misuse.

Beyond these immediate concerns, participants also acknowledge that the social meaning and role of love robots would change over time through media, public discourse, and everyday use. Participants also specifically emphasised the need for diversity in both design and decision-making processes, including greater involvement of women in shaping how these technologies evolve.

4.5. Strengths and limitations

The main strength of our study is the in-depth interviews conducted with women, a group which has been mostly overlooked in previous research on this topic. However, our study also has some important limitations. As a qualitative study of 12 women, our findings might not be generalisable to all women. Further, our participants were exclusively women living abroad without their usual social network, leaving them especially vulnerable to loneliness at the time of interview, possibly influencing their views on the research topic. A further limitation of our study is the biased recruitment during data collection. Specifically, most participants shared a relatively high level of familiarity with AI, which may have shaped the way they engaged with the topic. This shared background means the sample might not reflect the perspectives of individuals with little or no knowledge of AI. Nonetheless, their background allowed for more in-depth discussions and the articulation of relevant concerns about AEI and love robots, as many were already aware of broader debates surrounding AI. Additionally, some participants assumed that the interviewer was in favour of love robots, which may have influenced their responses. To mitigate this, the interviewer clarified their neutral stance in later interviews. Lastly, the study was based on hypothetical scenarios, as love robots with advanced AEI do not yet exist on the consumer market. This theoretical context shaped participants' reflections, which may not fully represent how they would respond if such robots were widely available. As several participants themselves pointed out, their views might change if love robots became normalised in everyday life.

5. Conclusion

This study explored the potential of love robots with AEI as companions for involuntary single women. The main themes investigated include women's experiences with involuntary singleness, gendered social norms and stigma around AEI technologies, and the ethical

implications of HRRs. By highlighting the experiences, perceptions, and concerns of involuntarily single women, our research challenges prevailing gendered narratives that position men as the primary beneficiaries or users of intimate technologies such as love robots. Women's insights revealed how social norms, gendered expectations, and media portrayals deeply shape their openness or resistance to these technologies. Participants acknowledged the potential of love robots to alleviate feelings of loneliness, support emotional growth, and provide safe spaces for developing intimacy and social competence. However, they also expressed considerable reservations about the authenticity of robotic affection, fears surrounding social stigma, and ethical concerns around normalising intimate relationships with machines.

The ethical dimensions highlighted by participants include uncertainty about the ontological status of love robots and the nature of genuine emotional connection, as well as broader societal consequences such as potential shifts in relational practices. These concerns emphasise the need for thoughtful, ethical design that prioritise autonomy, emotional wellbeing, and social connectedness.

Ultimately, this study calls for a more inclusive approach in technological development, advocating for greater diversity, transparency, and sensitivity to gendered experiences. By incorporating women's voices into these developments, we can advance more ethical and socially responsible uses of love robots in society. Future research should include comparative studies examining both men's and women's perspectives to understand similarities and differences in how genders perceive and accept love robots. Additionally, cross-cultural investigations beyond Scandinavia are needed to explore how diverse societies view artificial intimacy.

Funding The authors did not receive any support from any organization for the submitted work.

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

Ethics approval and informed consent This study was approved by the Norwegian Agency for Shared Services in Education and Research. Data of approval: 14.01.2025. All procedures

involving human participants were conducted in accordance with the ethical standards of the agency. Participation was voluntary, and informed consent was obtained prior to their involvement. Participants were assured of anonymity and confidentiality, and they retained the right to withdraw from the study at any time without any consequences.

Data Availability Statement The participants of this study did not consent to their data to be shared, so supporting data are not available.

REFERENCES

- Adamczyk, K. (2017). Voluntary and involuntary singlehood and young adults' mental health: An investigation of mediating role of romantic loneliness. *Current Psychology*, 36(4), 888-904.
- Apostolou, M., Matogian, I., Koskeridou, G., Shialos, M., & Georgiadou, P. (2019). The price of singlehood: Assessing the impact of involuntary singlehood on emotions and life satisfaction. *Evolutionary Psychological Science*, 5, 416-425.
- Baumeister, R. F., & Leary, M. R. (2017). The need to belong: Desire for interpersonal attachments as a fundamental human motivation. *Interpersonal development*, 57-89.
- Beschorner, T., & Krause, F. (2017). Dolores and robot sex: fragments of non-anthropocentric ethics. *International Conference on Love and Sex with Robots*,
- Birhane, A., & van Dijk, J. (2020). Robot Rights? Let's Talk about Human Welfare Instead. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*.
- Braun, V., & Clarke, V. (2021). *Thematic analysis: A practical guide*. Sage.
- Braun, V., & Clarke, V. (2022). Conceptual and design thinking for thematic analysis. *Qualitative psychology*, 9(1), 3.
- Broekens, J., Heerink, M., & Rosendal, H. (2009). Assistive social robots in elderly care: a review. *Gerontechnology*, 8(2), 94-103.
- Carvalho Nascimento, E. C., Da Silva, E., & Siqueira-Batista, R. (2018). The "use" of sex robots: A bioethical issue. *Asian Bioethics Review*, 10(3), 231-240.
- Coeckelbergh, M. (2010). Moral appearances: emotions, robots, and human morality. *Ethics and Information Technology*, 12, 235-241.
- Coeckelbergh, M. (2014). The moral standing of machines: Towards a relational and non-Cartesian moral hermeneutics. *Philosophy & technology*, 27, 61-77.
- Danaher, J. (2017). The Symbolic-Consequences Argument in the Sex Robot Debate. In J. Danaher & N. McArthur (Eds.), *Robot Sex: Social and Ethical Implications*. MIT Press.
- Danaher, J., Earp, B., & Sandberg, A. (2017). Should we campaign against sex robots?
- De Freitas, J., Uguralp, A. K., Uguralp, Z. O., & Stefano, P. (2024). AI companions reduce loneliness. *arXiv preprint arXiv:2407.19096*.
- Dillman Carpentier, F. R., Stevens, E. M., Wu, L., & Seely, N. (2017). Sex, love, and risk-n-responsibility: A content analysis of entertainment television. *Mass Communication and Society*, 20(5), 686-709.
- Dworkin, S. L., & O'Sullivan, L. (2005). Actual versus desired initiation patterns among a sample of college men: Tapping disjunctures within traditional male sexual scripts. *Journal of Sex Research*, 42(2), 150-158.

- Döring, N., & Poeschl, S. (2019). Love and sex with robots: a content analysis of media representations. *International Journal of Social Robotics*, 11(4), 665-677.
- Döring, N., & Pöschl, S. (2018). Sex toys, sex dolls, sex robots: Our under-researched bed-fellows. *Sexologies*, 27(3), e51-e55.
- Farvid, P., Braun, V., & Rowney, C. (2017). 'No girl wants to be called a slut!': Women, heterosexual casual sex and the sexual double standard. *Journal of Gender Studies*, 26(5), 544-560.
- Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V.,...Rossi, F. (2018). AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. *Minds and machines*, 28, 689-707.
- Ford, J. V. (2021). Unwanted sex on campus: The overlooked role of interactional pressures and gendered sexual scripts. *Qualitative Sociology*, 44(1), 31-53.
- Fosch-Villaronga, E., & Poulsen, A. (2020). Sex care robots: exploring the potential use of sexual robot technologies for disabled and elder care. *Paladyn, Journal of Behavioral Robotics*, 11(1), 1-18.
- Gasteiger, N., Loveys, K., Law, M., & Broadbent, E. (2021). Friends from the Future: A Scoping Review of Research into Robots and Computer Agents to Combat Loneliness in Older People. *Clinical Interventions in Aging*, 16(null), 941-971. <https://doi.org/10.2147/CIA.S282709>
- Ging, D. (2019). Alphas, betas, and incels: Theorizing the masculinities of the manosphere. *Men and masculinities*, 22(4), 638-657.
- Gunkel, D. J. (2018). *Robot rights*. mit Press.
- Heinrich, L. M., & Gullone, E. (2006). The clinical significance of loneliness: A literature review. *Clinical Psychology Review*, 26(6), 695-718. <https://doi.org/https://doi.org/10.1016/j.cpr.2006.04.002>
- Hilal, A. H., & Alabri, S. S. (2013). Using NVivo for data analysis in qualitative research. *International interdisciplinary journal of education*, 2(2), 181-186.
- Jecker, N. S. (2021). Sociable robots for later life: Carebots, friendbots and sexbots. *Sex robots: Social impact and the future of human relations*, 25-40.
- Johanssen, J., & Kay, J. B. (2024). From femcels to 'femcelcore': Women's involuntary celibacy and the rise of heteronihilism. *European Journal of Cultural Studies*, 13675494241293731.
- Jones, A. (2020). Incels and the manosphere: Tracking men's movements online.
- Jones, W. H., Sansone, C., & Helm, B. (1983). Loneliness and interpersonal judgments. *Personality and Social Psychology Bulletin*, 9(3), 437-441.
- Kay, J. B. (2021). Abject desires in the age of anger: Incels, femcels and the gender politics of unfuckability. In *Imagining "We" in the Age of "I"* (pp. 29-46). Routledge.
- Killeen, C. (1998). Loneliness: an epidemic in modern society. *Journal of advanced nursing*, 28(4), 762-770.
- Kim, J. L., Lynn Sorsoli, C., Collins, K., Zylbergold, B. A., Schooler, D., & Tolman, D. L. (2007). From sex to sexuality: Exposing the heterosexual script on primetime network television. *Journal of sex research*, 44(2), 145-157.
- Kroes, P., Vermaas, P. E., Light, A., Moore, S. A., & Sullins, J. P. (2008). Friends by design: A design philosophy for personal robotics technology. *Philosophy and design: From engineering to architecture*, 143-157.

- Laban, G., & Cross, E. S. (2024). Sharing our Emotions with Robots: Why do we do it and how does it make us feel? *IEEE Transactions on Affective Computing*.
- Larsen, M., & Kennair, L. E. O. (2024). Enough with the incels! A literary cry for help from female insings (involuntary single). *Evolutionary Behavioral Sciences*.
- Levy, D. (2007). *Love + Sex with Robots: The Evolution of Human-Robot Relationships*. HarperCollins.
- Lutz, C., Schöttler, M., & Hoffmann, C. P. (2019). The privacy implications of social robots: Scoping review and expert interviews. *Mobile Media & Communication*, 7(3), 412-434.
- Lynch, C. R. (2021). Artificial emotional intelligence and the intimate politics of robotic sociality. *Space and Polity*, 25(2), 184-201.
- McArthur, N. (2017). The Case for Sexbots. In J. Danaher & N. McArthur (Eds.), *Robot Sex: Social and Ethical Implications* (pp. 0). The MIT Press.
<https://doi.org/10.7551/mitpress/10718.003.0007>
- Morrison, D. M., Masters, N. T., Wells, E. A., Casey, E., Beadnell, B., & Hoppe, M. J. (2015). "He enjoys giving her pleasure": Diversity and complexity in young men's sexual scripts. *Archives of Sexual Behavior*, 44, 655-668.
- Mosakas, K. (2021). On the moral status of social robots: considering the consciousness criterion. *AI & SOCIETY*, 36, 429-443.
- Myrtveit-Stensrud, L., Schaller, S. L., Haugstad, G. K., & Groven, K. S. (2024). Navigating conflicting ideals of masculinity: a qualitative study of the experiences of male partners of women with vulvodynia. *The Journal of Sex Research*, 1-12.
- Myrtveit-Stensrud, L., Haugstad, G. K., Rème, S. E., Schaller, S. L., & Groven, K. S. (2023). "It's all my fault": a qualitative study of how heterosexual couples experience living with vulvodynia. *Acta Obstetricia et Gynecologica Scandinavica*, 102(10), 1378-1389.
- Ndlela, M., & Tanner, M. (2018). The perceived psychological empowerment of women using mobile dating applications: The case of tinder. *SOTICS 2018: The Eighth International Conference on Social Media Technologies, Communication, and Informatics*.
- Nordmo, M., Næss, J. Ø., Husøy, M. F., & Arnestad, M. N. (2020). Friends, lovers or nothing: Men and women differ in their perceptions of sex robots and platonic love robots. *Frontiers in Psychology*, 11, 355.
- Nygård, S., Ivanova, A., & Fischer, N. (2024). Er ufrivillig singelhet egentlig et mannsproblem? *Aftenposten*.
- Nyholm, S. (2020). *Humans and robots: Ethics, agency, and anthropomorphism*. Rowman & Littlefield Publishers.
- Nyholm, S., & Frank, L. E. (2017). From Sex Robots to Love Robots: Is Mutual Love with a Robot Possible? In *Robot sex: Social and ethical implications*.
- Nyholm, S., & Frank, L. E. (2019). It loves me, it loves me not: is it morally problematic to design sex robots that appear to love their owners? *Techne: Research in Philosophy & Technology*, 23(3).
- O'Donnell, C., & Shor, E. (2022). "This is a political movement, friend": Why "incels" support violence. *The British Journal of Sociology*, 73(2), 336-351.
- Picard, R. W. (2000). *Affective computing*. MIT press.
- Pu, L., Moyle, W., Jones, C., & Todorovic, M. (2018). The Effectiveness of Social Robots for Older Adults: A Systematic Review and Meta-Analysis of Randomized Controlled Studies. *The Gerontologist*, 59(1), e37-e51. <https://doi.org/10.1093/geront/gny046>

- Richardson, K. (2016). The asymmetrical 'relationship' parallels between prostitution and the development of sex robots. *Acm Sigcas Computers and Society*, 45(3), 290-293.
- Richardson, K. (2019). The human relationship in the ethics of robotics: a call to Martin Buber's I and Thou. *AI & SOCIETY*, 34(1), 75-82.
- Sanchez, D. T., Fetterolf, J. C., & Rudman, L. A. (2012). Eroticizing inequality in the United States: The consequences and determinants of traditional gender role adherence in intimate relationships. *Journal of sex research*, 49(2-3), 168-183.
- Scheutz, M. (2009). The Inherent Dangers of Unidirectional Emotional Bonds between Humans and Social Robots. In P. Lin, K. Abney, & G. A. Bekey (Eds.), *Robot Ethics: the ethical and social implications of robotics*. MIT Press.
- Scheutz, M., & Arnold, T. (2016). Are we ready for sex robots? 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI),
- Schuller, D., & Schuller, B. W. (2018). The age of artificial emotional intelligence. *Computer*, 51(9), 38-46.
- Sica, A. (2023). The Robot will Feel You Now: The Ethics of Artificial Emotional Intelligence in Sex Robots IEEE Ro-Man 2023, Busan.
- Sica, A., & Sætra, H. S. (2024). Artificial Emotions and the Evolving Moral Status of Social Robots. *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*.
- Sica, A., & Sætra, H. S. (2025). Feeling Machines: Challenges and Ethical Perspectives in Developing Robots with Emotions. In *International Handbook of Emotions: Resourceful Cultural Perspectives*, Vol. 2 (pp. 31-43). Cham: Springer Nature Switzerland.
- Simon, W. (2017). *Sexual conduct: The social sources of human sexuality*. Routledge.
- Sparks, B., Zidenberg, A. M., & Olver, M. (2022). One is the loneliest number: Involuntary celibacy, mental health, and loneliness. Preprint, 10.
- Spielmann, S. S., MacDonald, G., Maxwell, J. A., Joel, S., Peragine, D., Muise, A., & Impett, E. A. (2013). Settling for less out of fear of being single. *Journal of personality and social psychology*, 105(6), 1049.
- Sullins, J. P. (2012). Robots, love, and sex: the ethics of building a love machine. *IEEE transactions on affective computing*, 3(4), 398-409.
- Sætra, H. S. (2022). Loving robots changing love: Towards a practical deficiency-love. *Journal of future robot life*, 3(2), 109-127.
- Thorburn, J., Powell, A., & Chambers, P. (2023). A world alone: Masculinities, humiliation and aggrieved entitlement on an incel forum. *The British Journal of Criminology*, 63(1), 238- 254.
- Traeen, B. (2008). When sex becomes a duty. *Sexual and Relationship Therapy*, 23(1), 61-84.
- Turkle, S. (2011). *Alone together: Why we expect more from technology and less from each other*. Basic Books.
- Wada, K., & Shibata, T. (2007). Living with seal robots—its sociopsychological and physiological influences on the elderly at a care house. *IEEE transactions on robotics*, 23(5), 972-980.
- Whitby, B. (2011). Do You Want a Robot Lover? The Ethics of Caring Technologies. *Robot ethics: The ethical and social implications of robotics*, 233.
- Wong, B. (2025). Mark Zuckerberg Thinks AI 'Friends' Can Solve The Loneliness Crisis. Here's What AI Experts Think. https://www.huffpost.com/entry/mark-zuckerberg-on-ai-friendships_l_681a4bf3e4b0c2b15d96851d

- Yen, H.-Y., Huang, C. W., Chiu, H.-L., & Jin, G. (2024). The effect of social robots on depression and loneliness for older residents in long-term care facilities: a meta-analysis of randomized controlled trials. *Journal of the American Medical Directors Association*, 25(6).
- Zhong, L., & Huang, V. (2024). For imagination or disillusionment? The social compensation and social enhancement effects on dating anxiety for online daters who fear being single. *Computers in Human Behavior*, 157, 108251.

Appendix 1 – Overview of participants

Pseudonym	Age	Country of origin	Occupation	Tech competency level
Anna	30	Southern Europe	Employed	High
Julie	34	North Asia	Employed	Medium
Linda	24	Eastern Europe	Student	Medium
Christina	24	Central Europe	Employed	Low
Maria	30	North Asia	Employed	High
Carol	33	Central Europe	Employed	Low
Emily	40	Central America	Employed	Medium
Karen	~25	Southern Europe	Employed	Medium
Blair	~25	Central Europe	Student	High
Sabrina	~30	Central Europe	Employed	Low
Beth	~25	Central Europe	Employed	High
Ina	28	Central Europe	Unemployed	Low

Appendix 2 – Interview guide

1. Relationship background:
 - a. Can you tell me your experience with being single? What are you looking for in a partner or relationship? Do you use dating apps?
2. Addressing competency in AI:
 - a. Do you have any knowledge of or experience with technology and AI?
 - b. What social roles do you think AI can play?
3. Introducing the concept of AEI:
 - a. How do you feel about the idea of technology that simulates emotions?
4. Introducing the concept of love robots:
 - a. Can you describe how you imagine a love robot?
 - b. How does it look like?
 - c. What is it able to do?
5. Considering having a relationship with a love robot:
 - a. Do you think a love robot could provide something similar to the kind of relationship you are looking for?
 - b. What could be interesting in using a love robot?
 - c. What do you imagine could put you off when using a love robot?
 - d. Is there anything else you would like the love robot to be able to be or do?
 - e. Do you see a difference between a robot and a human partner?
 - f. Would you be personally interested in having a love robot for yourself?
 - g. (if using dating apps) Would you have a preference between people met on dating apps and a love robot?
6. Considering social and ethical implications of love robots:
 - a. Thinking more generally, not necessarily related to your personal experience, do you think there are benefits for involuntary single people that use a love robot? Do you think of concerns, instead?
 - b. Do you think people in general will be willing to accept love robots?
7. Reflection and wrap up:
 - a. If you had such a robot, would you stop looking for a partner, or are they different things to you?
 - b. Is there anything else you would like to add?