To be Cared for or Deceived?

Operationalizing Ethics in the Case of Elderly Care Robots under the European AI Act

Gaia Contu

Abstract: *Does the use of social robots in elderly care constitute a form of deception toward the individuals concerned? Focusing on the framework of the European AI Act, the paper examines how such concerns can be practically addressed under Article 5 on prohibited AI practices. Drawing on an analysis of the non-neutrality of technology in the fields of artificial intelligence and robotics, and referring to well-known cases such as COMPAS and the Amazon recruitment algorithm, the study highlights the ethical relevance of design choices in social robotics. It argues that ethical considerations are always present in technological development and must be operationalized from the earliest stages of designing robots for elderly care. Through a theoretical investigation grounded in Theory of Affective Coordination, pragmatism and relational ethics, the contribution shows that social robots in elderly care are not deceptive per se, and that potential risks of deception can be meaningfully addressed through concrete design choices and appropriate normative guidelines.*

Keywords: *Social robots; non-neutrality of technology; value-sensitive design; elderly care robots; ethics of technology; AI act; robot ethics*

Introduction

In a world where emerging technologies such as AI and robotics are rapidly expanding and playing a crucial geopolitical and economic role, the European Artificial Intelligence Act (European Commission 2024) represents one of the first attempts to regulate a field that has typically developed in a legislative vacuum. This effort is significant, given that our relationship with technology is

not occasional but an integral part of daily life. Smartphones can almost be seen as extensions of human bodies; many people use large language models like ChatGPT daily, construct identities and affective relationships through social media, wear smartwatches to monitor vital parameters, and artificial intelligence is being used in socially relevant contexts such as healthcare and public administration.

One sensitive domain where this issue warrants particular attention is elderly care. The field is increasingly integrating digital tools, such as telemedicine devices, to ensure high-quality care at home, particularly in marginalized areas and for patients with limited mobility; at the same time, a growing ethical concern relates to the use of robotic assistants (Sharkey/Sharkey 2012, Anderson/Felzmann 2024). In fact, a rapidly developing area is that of "Elderly Care Robots" (ECRs): robotic devices designed to support older adults across a wide range of tasks, from providing physical assistance and help with daily activities, to offering personalized support and even acting as social companions to alleviate loneliness (Hoppe et al. 2020, Yen et al. 2024, Ahmed [Buruk/Hamari] 2024). Among these, the most ethically controversial use is arguably that of social robots. These machines are designed to simulate human-like interaction and characteristics, including language use, appearance, and behavioral cues. This has led several scholars to question the authenticity, reciprocity, and potential for deception in human-robot relationships — particularly when the users belong to vulnerable populations, such as children or the elderly (Sparrow/Sparrow 2006; Sharkey/Sharkey 2012; Turkle 2006; van Wynsberghe 2016).Thus, an ethical, social, and legal analysis is crucial to protect vulnerable users and to provide clear guidelines and normative boundaries for developers, users, and distributors of emerging technologies.

In this paper, I will explore how to integrate ethics and moral values into technology, with a particular focus on robots designed for the care and companionship of older adults. I will examine the phenomenon of human-robot emotional attachment and the associated risk of deception. The topic will also be addressed from a legal standpoint: according to Article 5 of the European AI Act, AI systems that deploy manipulative or deceptive techniques to impair user autonomy must be prohibited. Therefore, it becomes essential to determine whether social robots and Elderly Companionship Robots fall within this regulatory scope.

To this end, the paper will proceed as follows. First, I will explore whether and how ethics can be operationalized in technological design. This includes demonstrating that technology is never neutral and that law, technoscience,

and ethics are deeply entangled. Drawing on the framework of Science and Technology Studies (STS), the Social Construction of Technology (SCOT), and feminist epistemologies, I will illustrate, through various case studies, how technoscientific artifacts inevitably embed values, worldviews, and contextual assumptions. Building on this, I will argue that it has always been possible to shape technological development in alternative directions, countering the myth of technological determinism. In particular, the framework of Value-Sensitive Design suggests that not only is it possible, but indeed necessary, to integrate specific human values into the design of technological systems. Likewise, user-centered design provides concrete methodological guidelines for moral inquiry, emphasizing the importance of incorporating the lived experiences and perspectives of end users.

Having laid this groundwork, I will then focus on a case study concerning social robots — specifically, Elderly Care and Companionship Robots. I will begin by defining what constitutes a social robot and examining several documented cases of their deployment among vulnerable populations. To this end, I will present a brief literature review of empirical studies evaluating the effects of social robot use among elderly individuals, especially in care home settings. Following this, I will engage with the common critique that social robots, by mimicking social cues and emotional expressions, are inherently deceptive. I will challenge this assumption by offering an ethical and epistemological analysis rooted in a relational understanding of mind and emotions. On this basis, I will propose a revised definition of deception, one that is both theoretically grounded and practically applicable. This reconceptualization will serve as a foundation for outlining regulatory and design-oriented guidelines at three levels, from the abstract to the concrete. I will argue for the need for targeted field studies and suggest pathways for rethinking the design and legal framing of social robots in ways that protect users without necessarily hindering innovation or production.

Background

The issue of guaranteeing a high quality of life for the elderly constitutes a main social concern. This is also stated by the third of the UN Sustainable Development Goals (SDGs) (UN, 2015), which aims to promote social inclusion and to ensure healthy lives and well-being at all ages, with particular attention to the most vulnerable, including older adults. According to the WHO, in 2020 the

number of people aged 60 years and older outnumbered children younger than 5 years, and by 2050 they will nearly double from 12% to 22% (WHO 2024). This wing of population is particularly fragile not only for what concerns physical illnesses, but also with regard to mental health: around 14% of adults aged 60 and over live with a mental disorder, and according to the Global Health Estimates (GHE) 2019, "these conditions account for 10.6% of the total disability (in disability adjusted life years, DALYs) among older adults" (WHO, 2023). This constitutes a critical healthcare problem according to the WHO definition of health as a "state of complete physical, mental and social well-being" (WHO 1946).

Loneliness has a particular influence on the psychosocial well-being of the elderly and is often identified as related to chronic illness and poor self-rated health (Jané-Llopis/Gabilondo 2008). This condition might intensify in a socioeconomic context where increasing mobility often leads to a shortage in the availability of family care (Hoppe et al. 2020), and where the growing elderly population comes into conflict with the non-scaling nature of the human care workforce (OECD, 2020).

Moreover, caregiver burden represents a significant yet often overlooked issue. Informal care within families is typically characterized by silent, wearisome labour, disproportionately carried out by women. As feminist scholars have emphasized, this work is often taken for granted, socially and institutionally invisible, and marked by power asymmetries and a lack of formal recognition or support (Tronto 1993, Kittay 1999, Parks 2010).

In response to the needs of care and companionship in old age, increasing attention is being devoted to the role of socially assistive robots, such as homecare robots, personal assistance robots or companionship robots. These robotic devices hold considerable potential: not only can they enhance care, providing caregivers with practical support such as physical assistance, but also respond to social needs and alleviate loneliness in the users (Ahmed [Buruk/Hamari] 2024, Yen et al. 2024). However, in order to be both effective and ethically acceptable, such technologies must be integrated thoughtfully, following a genuinely human-centered approach. This requires careful consideration of moral values and social impacts, as well as of legal boundaries and practical feasibility. It also demands ongoing dialogue and collaboration among all relevant stakeholders: final users, technical developers, engineers, policy makers, families and caregivers, healthcare personnel and other key actors.

But how is it possible to integrate the potential of social robotics with the needs of vulnerable populations in the most *ethical* way possible? How can we

truly develop *ethical* technologies, where the call for human-centeredness is more than just a formal phrase? Is it really possible to embed ethical and human values into technology?

Is it Possible to Operationalize Ethics?

In fact, demanding that technology be "ethical" is a rather vague formulation. What does it actually mean for a technology to be ethical? If we take the definition of ethics as a systematic discourse and philosophical analysis of moral values — that is, of what is right or wrong, good or bad — we are immediately faced with a situation of pluralism (Neri, 2020). Not only are there different ethical approaches and frameworks, but on the moral level as well, values differ from individual to individual, are context-dependent, and are influenced by culture, society, and other factors. So, when we call for ethicality of robots, *which* values are we referring to? And where is the authority or final decision-making power when it comes to evaluation?

To address this question in the context of Elderly Care Robots (ECRs), a process of specification is necessary. We will therefore break down the overarching question into the following sub-questions:

1. Can technology in general embody or integrate moral values?
2. How can such values be implemented or operationalized in technological design and development?
3. Who is entitled to assess the ethicality of a technology?
4. Which specific values are at stake in the case of ECRs – and how can they be assessed and implemented in practice?

In this section, we will address the first three questions, examining the issue of how ethical values can be integrated into technology in general. In the following section, we will turn to the specific case of social robotics.

The Non-Neutrality of Technology

To answer the question of whether technology can embody or integrate moral values (sub-question A), it is important to revisit the long-standing and influential view of technoscience as a value-neutral enterprise. The idea of science

as an impartial, universal and objective endeavour remains popular today. In school, we learn formulas as if they have always been set in stone, and we hear about scientific discoveries and technological inventions as though their only connection to cultural and social contexts is serendipity (Carrada 2005, Bucchi 2002). The image of science conveyed in public discourse, from the news to science communication, is what Latour would have called *ready-made science*: certain, cold, unproblematic, its process of construction and stabilization sealed within a black box. This stands in contrast to the more accurate and dynamic concept of *science in the making*, which is open to controversy, uncertainty, reversals, and mistakes, where truth and efficiency emerge through compromises among relevant actors (Latour 1987).

Alongside this narrative of science, there persists a theoretical assumption that scientific progress follows a linear, inevitable trajectory, largely uninfluenced by cultural, sociopolitical, or geographical contexts. This view is called scientific "determinism" or "inevitability" and comprises the idea that technology develops independently of society and, rather than being shaped by it, determines the course of social development. (Bijker 2009). This idea originates from the 1922 essay by William Ogburn and Dorothy Thomas, *Are Inventions Inevitable? A note on social evolution*, where they argued that revolutionary inventions are simply the inevitable outcome of cultural and technical components. As they put it: "Given the boat and the steam engine, isn't the steamboat inevitable?" (Ogburn & Thomas 1922, Huyskes 2025). This deterministic view continues to shape how emerging technologies are understood and developed. In the case of robotics, for example, there is a widespread view that imagines a linear and necessary trajectory of progress leading toward a very specific prototype: humanoid robots (with masculine body design), equipped with artificial consciousness or artificial general intelligence, and potentially destined to rebel against humans. This image, popular in the West, is shaped by deep-rooted cultural and literary narratives, and it powerfully influences our collective imaginary (Allen [Wallach/Smit] 2006, Van Grunsven 2022). In parallel, is also common the perception of technology, including Artificial Intelligence, as unbiased, objective, and even superior to fallible human judgment. In the words of sociologist Ruha Benjamin, in fact, "many industries and organizations well beyond health care are incorporating automated tools, from education and banking to policing and housing, with the promise that algorithmic decisions are less biased than their human counterpart." (Benjamin 2019). Indeed, a great deal of technology has been introduced precisely to compensate for human error and limitations: from recruitment tools such as the robot

Tengai, originally marketed as an *unbiased* interviewer[1], to Diella, the Albanian virtual minister created with AI in 2025 to fight corruption[2].

Yet, a more critical look reveals that this deterministic, objective, and neutral narrative of scientific progress has been challenged and largely abandoned within the academic communities of social sciences, philosophy of science, and science and technology studies for nearly 60 years.

STS, SCOT and Feminist Epistemologies

From the 1970s onward, the doctrine of technological determinism and the thesis of scientific neutrality and universalism began to show descriptive fragility, and its assumptions have been critically challenged. Technological determinism was argued to be a poor research strategy because it entails a teleological, linear, and one-dimensional view of technological development,

In its place, research approaches have proliferated in which the social dimension of knowledge plays a central role (Bucchi 2002, Bijker [Hughes/Pinch] 1987; Huyskes 2025). The firsts were the Science–Technology–Society (STS) movement and the Sociology of Scientific Knowledge (SSK), which began to investigate scientists' social responsibilities, demonstrating that technologies are shaped by social practices and decisions and supporting the view of a coevolution of technoscience and society (Keulartz et al. 2004, Bijker 2009). These technology studies replaced the old deterministic view with a more constructivist approach, which sees technological artifacts as "the outcome of negotiations, in which many diverse actors are involved." In this approach, "there is rarely one single path of development but rather a number of potentially viable alternatives […], not completely autonomous at all, […] but rather a fairly random result of social interactions" which "require particular role patterns and lay down a specific 'geography of responsibilities.'" (Akrich 1992, Keulartz et al 2004)To strengthen this critique of technological determinism, it was necessary to show that the workings of technology were socially constructed, with an emphasis on the social (Bijker 2009, Keulartz et al 2004). This contribution came from two main frameworks: SCOT (Social Construction of Technology), in which the contributions of STS and SSK have converged, and Feminist Epistemologies.

1 https://www.tengai-unbiased.com, accessed in Feb 2022.

2 https://www.theguardian.com/world/2025/sep/11/albania-diella-ai-minister-public-procurement.

The SCOT movement aimed to unveil the role of "relevant social groups", "interpretive flexibility", "stabilization", and "closure" in technological development. To illustrate these concepts, Bijker uses the bicycle as a case study. He demonstrates that the design of the bicycle — far from being the result of a linear, inevitable, or purely technical process — evolved through complex social negotiations among diverse groups, such as bicycle producers, young athletic ordinary users, women cyclists, and anti-cyclists. Each group attributed different meanings to the technology: men saw it as a symbol of power, while women faced practical barriers like clothing constraints. These conflicting interpretations led to design changes, such as the introduction of pneumatic tires to enhance safety and comfort, and to the dominance of one model over others (Bijker 1995). To this critique, feminist epistemologies added the crucial insight that technological development not only integrates society but also reflects and reinforces its underlying power structures. Feminist scholars have shown that the so-called "universal subject" of science — assumed to produce objective and universally valid knowledge — was, in fact, neither universal nor neutral. Rather, it was a highly specific subject, situated in time, space, class, ethnicity, and gender. Typically, this subject was male, white, middle- or upper-class, and Western. This positionality inevitably shaped the production of scientific and technological knowledge.

For example, the case of the contraceptive pill reveals how technology embodies and reproduces gendered hierarchies and power relations. The development and deployment of the pill were influenced not only by scientific and medical agendas but also by social assumptions about female bodies, reproductive control, and gender roles. As discussed by Keulartz et al., if on the one hand the pill could be considered an emancipatory technology, on the other it continued to assign reproductive responsibility primarily to women, naturalizing this asymmetry through its design (Oudshoorn 2003, Tripaldi 2023, Keulartz et al. 2004). Moreover, the pill exposed women to health risks deemed unacceptable for men, using biological framings to obscure social choices and normatively charged evaluations (Oudshoorn 2003). Feminist epistemologies reveal how such technologies are products of androcentric knowledge systems, defining whose bodies are regulated and how. It is also noteworthy that not only the pill reflected existing norms but also actively reshaped norms, transforming the moral understandings of sexuality, altering practices, expectations, and responsibilities (Keulartz et al. 2004).

The same mechanism of integration of social injustices into technoscience does not only affect women, but also other marginalized groups, such as non-

white ethnicities or lower social classes. For example, in healthcare, pulse oximeters have been shown to be less accurate for Black patients because they were primarily designed and adjusted using lighter skin tones: an implicit bias that can affect life or death. (Al-Halawani et al. 2023). Or again, in urban design, the famous and still debated bridges built by architect Robert Moses in Long Island during the 1930s–1940s were so low that public buses could not pass underneath, effectively excluding poorer, often Black communities from certain areas (Winner, 1980; Huyskes, 2025). The examples of exclusions and practical damages are countless: from clinical trials where women are underrepresented — leading to results calibrated on the male physiology and thus often ineffective or harmful for female patients — to car crash tests based on male body, resulting in a higher likelihood of severe outcomes for women, as well as bulletproof vests and even office temperature settings, all originally designed around male standards (Criado Perez 2019).

The partial and exclusionary framing of technoscientific endeavour, which systematically excludes portions of the population from participating in the construction of knowledge, is therefore not only morally unjust, but has practical consequences and is epistemically limiting. When scientific and technological development is shaped by a limited set of social experiences and values, the resulting outputs are ineffective or even harmful for large segments of the population (Harding 1986; Grasswick 2018).

This dynamic becomes even more evident in the case of Artificial Intelligence and digital technologies, where the societal implications of biased data and exclusionary design perspectives are integrated and amplified (Bartoletti 2020; Huyskes 2025).

Digital Technologies as a Lens to Read Embedded Values

Around the late 2010s, particularly in the United States, a series of troubling cases involving algorithms, software, and digital technologies began to draw public and academic attention. While these technologies were introduced with the stated aim of overcoming human limitations — such as bias in hiring, risk assessment in the criminal justice system, or public housing allocation — they often produced outcomes that were deeply problematic.

In 2018, Joy Buolamwini, an African American researcher at MIT, discovered that the facial recognition software she was working on failed to accurately recognize her face. Further investigation revealed a significant disparity: while the maximum error rate for light-skinned males was just

0.8%, this increased dramatically for other demographic groups. The most misclassified group were dark-skinned females, that faced error rates of up to 34.7% (Buolamwini and Gebru 2018). Disparities like these can carry serious consequences, particularly when facial recognition technologies are used in sensitive domains like law enforcement, where misidentification can lead to wrongful arrests or deportation. Similarly, in 2016, an investigative report by *ProPublica* revealed that U.S. courts were using an algorithm known as COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) to estimate the risk of criminal recidivism. Based on a dataset of over 10,000 defendants, the investigation found that the software falsely predicted a high risk of recidivism in 45% of Black defendants — nearly double the rate for white defendants (23%), who were instead more likely to be incorrectly classified as low risk (Angwin et al., 2016). The algorithm was not only ethically unjust but also epistemically flawed and technically ineffective, exhibiting low predictive accuracy. In another example, Amazon ran an AI-based recruitment trial in 2014 to streamline hiring processes. By 2017, it was discontinued after the algorithm was found to systematically disadvantage women (Dastin, 2018).

In these cases, it becomes particularly evident how Artificial Intelligence both integrates and reproduces existing social dynamics. In all three of the previously mentioned cases, the discriminatory outcomes were not simply incidental but rather emerged directly from the type of AI technology employed: supervised machine learning models trained on real-world social data. In supervised machine learning, neural networks are trained on large datasets and learn to recognize patterns by adjusting internal parameters to minimize prediction errors across labelled examples. Once these statistical regularities are identified, the model uses them to make new decisions and perform future tasks. However, in this workflow, the model can encode and amplify existing patterns of bias in the training data, also because it identifies correlations without understanding the underlying causes or broader social contexts (Barocas [Hardt/Narayanan] 2023). Moreover, this process is often opaque, even to the developers themselves, making it difficult to identify or mitigate discriminatory outcomes.

In the case of facial recognition, for instance, training datasets were predominantly composed of lighter-skinned individuals (79.6% in IJB-A and 86.2% in Adience), leading to biased performance (Buolamwini and Gebru, 2018). In the COMPAS and Amazon cases, the training data itself was historically biased, rooted respectively in a long legacy of discrimination against

African American communities in the justice system and gender disparities in the tech labor market.

A similar dynamic can also emerge from the choice and selection of proxy variables. For instance, Obermeyer et al. (2019) found that a widely used algorithm in the U.S. healthcare system predicted patients' health needs based on past health expenditure. Because less money is typically spent on Black patients with the same medical needs as white patients, the software consistently underestimated the needs of Black individuals. As a result, the number of Black patients eligible for extra care was reduced by over half compared to those who would have actually needed it (Obermeyer et al. 2019; Benjamin 2019).

Some biases may emerge from the very structure of the algorithm, independent of the data it processes: for instance, in collaborative filtering systems like those used by Netflix or Amazon, specific algorithmic design choices can lead to biased recommendation patterns. Statistical biases — such as the *cold-start* problem, *popularity bias*, and *homogenization bias* — can directly produce discriminatory outcomes by reinforcing existing stereotypes or systematically favoring content associated with majority groups. This is also evident in the case of search engines, where such biases can result in the preferential ranking of commercial entities or businesses from dominant groups, with tangible economic consequences. (Stinson 2022). Importantly, this mechanism operates even at the foundational level of algorithmic design. The development and architecture of any algorithm — like that of any technology — inevitably reflect human decisions, which are inherently situated and partial (Thaler & Sunstein, 2009). A striking example emerges when AI is applied to the legal domain. Here, attempts to formalize law into machine-readable code require interpretative decisions that become hardcoded into the system, effectively excluding alternative legal interpretations. As Surden (2017) argues, law is inherently interpretive, and the process of translating legal norms into code is never neutral. On the contrary, it necessitates selecting specific legal approaches, assumptions, and values. Even seemingly technical design choices — such as making court documents publicly searchable — reflect value-laden trade-offs, often balancing transparency against privacy. The integration of worldviews, goals, and power relations into technological design is also evident in the field of robotics. When we design a robot, we inevitably make decisions about aesthetics, form, functionality, and even moral behavior. These decisions are never entirely neutral or purely technical. Even when we consider only the aesthetic dimension, the dominant prototype in robotics tends to reflect a male-coded body, while alternative designs are often hyper-sexualized representations of

female forms. This becomes even more evident in the case of sex robots, where the function, features, and appearance are simultaneously technical and ethical, shaped by deeply rooted sociocultural influences (van Grunsven/van Wynsberghe 2019, van Grunsven [Stone/Marin] 2024).

Algorithms and robotic systems clearly show how technology incorporates existing social patterns, inequalities, and normative choices at multiple levels, in various forms, and across a wide range of domains: from the tech industry to the legal system, from healthcare to entertainment. The problem is even greater considering that these socially shaped systems are developed under a narrative of neutrality, which conceals their value-laden nature.

The Implementation of Moral Values into Technological Design

Returning, then, to the question of whether technology can embody or integrate moral values, an affirmative answer is logically entailed. As we have seen in the previous sections, technology not only can integrate moral values, it also cannot not do so. In other words, it is inherently imbued with values by its very nature. If the technoscientific process is intrinsically value-laden, and if technological outcomes are not predetermined — as extensively argued by anti-deterministic and social constructivist approaches — then this opens up a significant space for human agency in shaping technological development. It becomes not only possible, but also unavoidable and necessary to make deliberate decisions about which values should be embedded within technological systems.

We thus arrive at the sub-question B: how can such values be consciously implemented or operationalized in technological design and development? To this end, philosophical and ethical reflection assumes a central and guiding role.

Value-Sensitive Design, Care Ethics and User-Centered Design Approach

One of the earliest approaches to advocate for the possibility of actively integrating specific values into technological design and engineering is Value-Sensitive Design (VSD), defined as "a theoretically grounded approach to the design of technology that accounts for human values in a principled and comprehensive manner throughout the design process" (Friedman and Kahn 2003, van Wynsberghe 2012). Friedman and Kahn criticize approaches that treat ethics as a marginal or after-the-fact concern in technological develop-

ment as insufficient. Instead, they advocate for the systematic integration of fundamental human values — such as human welfare, ownership and property, privacy, freedom from bias, universal usability, trust, autonomy, informed consent, accountability, identity, and environmental sustainability — from the early stages of the design process. To this end, their VSD methodology proposes an iterative and integrative process that integrates conceptual, empirical and technical investigations. The *conceptual* component uses philosophical analysis to clarify how values are supported or diminished by technological designs, who is affected, and how trade-offs among competing values should be navigated in the design. The *empirical* investigation draws on social science to understand how people experience and interpret those values. The *technical* part assesses existing or potential technological solutions in terms of their capacity to uphold or undermine those values (Friedman and Kahn, 2003).

A particularly relevant variant of Value-Sensitive Design (VSD) can be found in the field of robotics, especially in the context of robots designed for the care of vulnerable populations: the Care-Centered Value-Sensitive Design (CCVSD) (Van Wynsberghe, 2012). Drawing on the philosophical framework of care ethics, van Wynsberghe applies the values identified by care ethicist Joan Tronto — attentiveness, responsibility, competence, and reciprocity (Tronto, 1993) — to the domain of elderly care robotics. This approach explores how these values can be operationalized within different robotic designs. In particular, the focus is on making each issue more specific and less abstract, taking into consideration various factors such as context (place), the actors involved, the care practices (e.g., lifting, bathing, feeding, delivery of goods, social interaction), the type of robot (assistive, enabling, replacement), and the manifestation of moral elements. A paradigmatic example of CCVSD is the practice of lifting elderly patient with low mobility, traditionally treated as a purely technical or engineering task. Yet, as van Wynsberghe illustrates, lifting is ethically charged: it occurs at a moment of maximal patient vulnerability and can result in feelings of objectification. This practice comprises value-laden and sensitive actions like eye contact that are integral to establishing and maintaining a bond of trust and preserving the patient's dignity (Van Wynsberghe, 2012). It is therefore possible to consider this ethically relevant practice right from the initial stages of design and functionality choices. When lifting is entirely delegated to autonomous robots (e.g., RIBA), these relational cues are often lost. In contrast, one can choose to favor the creation and use of enabling technologies, such as exoskeletons, which allow the caregiver to pre-

serve the ethical sensitivity of the practice (ibd.). However, if moral evaluation is context-specific, dependent on the individual characteristics of a patient and shaped by cultural variation, pluralistic values, and situational nuance, then a central question emerges: who decides which values are embedded in the technology? We thus arrive at the sub-question C, namely who is entitled to assess the ethicality of a technology.

A deeply rooted view in Western philosophy holds that the final say in ethical decisions belongs to the citizens directly affected (Dewey 1927, Arendt 1958, Rawls 1993). Only those who live the problems have the experiential standpoint needed to judge them, and only their participation can make science and technology genuinely democratic.

This is where methodological frameworks such as the User-Centered Design (UCD) become crucial. UCD encompasses approaches such as *participatory design* and *co-design*, which aim to include stakeholders, and especially end-users, throughout the design process (Sanders & Stappers, 2008). In this view, the active engagement of users is not merely a desirable feature but a necessary condition for ensuring that the technology remains truly human-centered (Gould/Lewis, 1985). By involving users through iterative cycles of co-construction, these approaches help to align technical functionalities with the lived values, concerns, and actual needs of end users. Many studies in healthcare show that integrating patient feedback is key to improving services (De Rosis [Ferrè/Pennucci] 2022).

In this way, the dimension of public trust and consumer adoption can be directly connected to that of moral reasoning. On one hand, the social acceptance of a technology is a pivotal for its success and long-term adoption. On the other, listening to end-users' voices serves to operationalize a range of ethical frameworks: from consequentialist ethics, where moral evaluation is based on promoting the greatest well-being for the greatest number (Neri 2020), to the relational and non-hierarchical approach of care ethics (Gilligan 1977, Botti 2015), to the mutual and dynamic processes addressed in pragmatist ethics (Keulartz et al., 2004), and even to non-paternalistic, self-defined forms of deontological reasoning. Importantly, this approach can also inform legal and regulatory frameworks. By grounding regulation in bottom-up evaluations and user participation, it can offer concrete guidance for policies that reflect real-world contexts and ethical concerns (WHO, 2007).

Operationalizing Ethics in the Case of Elderly Care Robots

After examining how technology in general can embody or integrate moral values, with a focus on various cases of AI and robotics, we have outlined several approaches for implementing such values in technological design and development and identified who should have the final say in the ethical assessment. With this foundation, we can now return to the specific case of robots for elderly care. There remains, in fact, the final and central sub-question D: which specific values are at stake in the case of Elder Care Robots (ECRs), and how can they be assessed and implemented in practice? To this end, the discussion will be divided into three parts. First, we will define what social robots and elderly companionship robots are, providing a general overview along with a review of their documented effects. Second, we will address the question: what does it mean for ECRs to be "ethical"? Finally, we will explore how such ethical considerations can be implemented in legal and regulatory frameworks.

Social Robotics and Robots to Combat Loneliness

Robots are currently among the most widely discussed and anticipated technological innovations, both in Western and Eastern contexts. In public discourse and the collective imagination, they are often seen as symbols of progress and future potential (Alnajjar et al. 2021). Originally designed to carry out physical tasks, robots are now increasingly being developed to engage in social interaction. One of the most sensitive and pressing social needs today is companionship, particularly for vulnerable groups such as the elderly. In response, growing attention is being directed toward the development of "companion robots", leading to the emergence of a dedicated and rapidly evolving field: social robotics. Social robots are "a new class of machines designed to function as 'social partners' for humans" (Damiano/Dumouchel, 2020). They are used in a wide variety of contexts: wellbeing, education, socialization, healthcare, disability assistance, elderly care, motivation and behavioral influence, rehabilitation, entertainment, navigation and guidance, shopping, and general assistance (Ahmed [Buruk/Hamari], 2024). They have proven particularly effective in contexts where human companionship is unavailable or limited (e.g. during pandemics) or in contexts where individuals are socially isolated because of physical or mental conditions (ibd.). The central strategy of social robotics is to create "socially interactive robots" (Fong [Nourbakhsh/Dautenhahn] 2003), meaning robots capable of generating a

believable social presence. This is defined as the robot's ability to give users the "sense of being with another" (Biocca et al., 2003) or the "feeling of being in the company of another" (Heerink et al., 2008). Achieving this involves transferring and adapting features of human face-to-face social interaction to human-robot interaction. One of the most crucial features is the ability to communicate through emotion (Damiano/Dumouchel, 2020). To this end, social robots are often equipped with advanced capabilities that enable them to foster deep emotional connection and facilitate social engagement, while also assisting with practical tasks (Ahmed [Buruk/Hamari], 2024). These characteristics do not necessarily need to be human-like. According to Fong et al. (2003), social robots can be categorized based on their appearance and features into four main types: anthropomorphic (human-like), zoomorphic (animal-like), functional (machine-like), and caricatured (object-like) (Fong et al. 2003, Ahmed [Buruk/Hamari] 2024). What makes a robot socially effective is not its resemblance to a human, but rather its capacity to trigger certain innate human responses. As Sherry Turkle explains, robots can activate our "Darwinian buttons", for example, by making eye contact, which causes people to respond as if they were engaging in a real social relationship (Turkle 2007). Designers often include features like large eyes and rounded heads to resemble infants. According to Konrad Lorenz, founder of ethology, such "baby-like" traits instinctively elicit tenderness and care from humans (Lorenz, 1970. Kaplan 2001). Some well-known social robots include Nao, Pepper, Buddy, Jibo, and Amazon's Astro: Nao is a small humanoid robot used mainly in research, education, and therapy, capable of speech, movement, and emotion recognition; Pepper, also from SoftBank, is designed to detect emotions and interact socially, and is widely adopted in research and pilot programs. Buddy is a friendly home companion with an expressive screen face, helping with reminders, entertainment, and social connection; Jibo and Astro, both aimed at private home use, combine features like voice interaction, mobility, and social presence — though with mixed commercial success. These robots can foster social presence, empathy, and contextual awareness. They often respond to voice, recognize faces, interpret emotions, and support users in both practical and emotional ways (Ahmed [Buruk/Hamari] 2024).

The field of social robotics for the elderly seeks to address the challenges posed by the aging population, shortages in care services, and the increasing burden on caregivers, as ensuring a high quality of life for older adults has become an increasingly important priority (WHO 2024). In this context, robots have been developed to assist with a wide range of tasks: physical support

(e.g., lifting, mobility aid), housework, guided physical exercise, cognitive assistance (e.g., reminders), and telemedicine (Hoppe et al 2020, Ahmed et al 2024). However, one of the most widespread and often overlooked issues in elderly care is social isolation, which is strongly linked to poor mental health outcomes such as depression, anxiety, and cognitive decline (Jané-Llopis/Gabilondo 2008, Yen et al. 2024). This is where Elderly Companionship Robots (ECRs) come into play. These social robots aim to alleviate loneliness by providing interactive, emotionally supportive presence. A particularly researched subgroup within elderly care is individuals living with dementia, for whom a wide variety of robotic devices have been developed. Examples include Paro, a baby seal-like robot covered in soft fur that reacts to touch and sound; Hyodol, a doll-like companion robot widely used in South Korea, even with cognitively healthy older adults; and animal-like robots such as the robotic dog Aibo and the Joy-for-All cat. Studies have shown positive outcomes, including reduced loneliness, improved communication, and emotional engagement (Tamura et al. 2004, Sharkey/Sharkey 2012, Robinson et al. 2013, Lee et al. 2023, Yen et al 2024). Similarly, studies on Aibo report comparable emotional benefits to those seen with real animals (Collins [Millings/Prescott] 2013). The same robots are also used in therapy with cognitively healthy older adults, where they serve as tools for stimulation, interaction, and emotional support. These include humanoid robots such as Nao and its therapeutic version Zora, Pepper, Aibo, and many others. Research suggests that these robots can support wellbeing, stimulate conversation, and help reduce feelings of isolation and anxiety, at least in the short term (Skarkey/Sharkey 2012, Anderson/Felzmann, 2024). For instance, a randomised controlled trial in a New Zealand care facility found that interactions with Paro significantly reduced loneliness compared to usual activities or even a resident dog. Residents engaged in more physical interactions and discussion, and Paro also appeared to act as a social catalyst within the group (Robinson et al. 2013).

However, evidence is mixed. While several narrative and qualitative reviews highlight clear psychosocial benefits, especially in terms of engagement, emotional expression, and reduced use of medication, some meta-analyses have found no statistically significant effects on outcomes such as quality of life or depression (Pu et al. 2019, Lu et al. 2021, Anderson/Felzmann, 2024). This may also be due to the inherent difficulty in detecting such effects through standardised measures, as highlighted by several narrative and qualitative analyses that instead report perceived benefits (Pu et al., 2019). Moreover, the biggest challenge regards long-term outcomes: for example, the study by Lee

et al. (2023) on Hyodol showed reduced depressive symptoms at three months, but no lasting effect at six, and some Japanese reports describe a rapid decline in users' interest over time (Sharkey/Sharkey, 2010).

How Can Companion Robots be Ethical?

Despite potential benefits, social robots have raised significant concerns among scholars, especially for what concerns the formation of affective bonds in human-robot relationship. What happens when our parents or grandparents become emotionally attached to robots that, at least *prima facie*, cannot genuinely reciprocate their feelings? Do the abovementioned features of social robotics (such as human-like appearance and behavior, the ability to evoke tenderness, personalized memory, and kind, believable speech) not only enhance user engagement but also involve a form of deception? Is it ethical to design companion robots with the specific aim of eliciting affective attachment, especially in vulnerable users like older adults?

The Accusation of Deception

When discussing ethics in the context of social robots, particularly companion robots, the specific moral values at stake are typically those related to human dignity, affective reciprocity, and the risk of emotional deception or manipulation. These moral values lie at the heart of a complex and ongoing ethical debate, with many scholars arguing that the use of social robots in vulnerable contexts, such as elderly care, is problematic. Among the first and most influential critiques is that of Sparrow and Sparrow, who argue:

> "Any beneficial effects of robot pets or companions are a consequence of deceiving the elderly person into thinking that the robot is something with which they could have a relationship. […] We believe that it is not only misguided, but actually unethical, to attempt to substitute robot simulacra for genuine social interaction." (Sparrow/Sparrow 2006)

Their position reflects a broader deontological perspective, concerned with authenticity, truthfulness, and respect for human dignity. Similar arguments have been developed by many others. For instance, Sherry Turkle, based on extensive empirical and theoretical work, including case studies of older adults interacting with social robots, has raised critical concerns. She asks:

> "The fact that our parents, grandparents and our children might say 'I love you' to a robot who will say 'I love you' in return, does not feel completely comfortable; it raises questions about the kind of authenticity we require of our technology." (Turkle 2006)

Van Wynsberghe also takes a critical stance toward this technology, focusing on issues of relational authenticity and vulnerability in care contexts:

> "Through an analysis of the value of reciprocity from care ethics, […] it becomes clear that HRI cannot achieve this bidirectional value of reciprocity; a robot must deceive users into believing it is capable of reciprocating to humans or is deserving of reciprocation from humans. […] social robots designed for reciprocity use reciprocity as an instrumental value to enhance acceptability of the robot." (van Wynsberghe 2022)

From these perspectives, emotional authenticity is considered central. The idea is that genuine emotions are understood as first-person, internal experiences; since social robots do not possess such inner emotional states and cannot genuinely feel or reciprocate emotions, their simulations of social or emotional behavior are inherently deceptive. Therefore, designing robots to evoke emotional attachment is, from this standpoint, ethically wrong. But is this truly the only, or the best, way to frame the issue? According to some scholars, the answer may be more nuanced.

Another Perspective: Emotions as Coordination and the Relational View of the Self

In recent years, new perspectives in the philosophy of mind and emotions, and more generally in understandings of the human mind and self, have begun to emerge as alternatives to the classical, standard views. According to Dumouchel and Damiano (2017), for example, who have developed a new theory of emotions applied to Human-Robot Interaction called the "Theory of Affective Coordination", the common argument of deception is based on a fundamental misunderstanding. They argue that, while modern philosophy of mind formally rejects Cartesian dualism and acknowledges the mind as a cognitive system like any other, it nevertheless tends to preserve the view of the human mind as a paradigmatic epistemic agent—almost ontologically distinct. The deception argument relies precisely on this residual dualism: it assumes that external expressions are only genuine if they mirror internal, private states.

Thus, when robots simulate emotions, they are accused of misleading users into believing that those emotions are authentic (Dumouchel/Damiano 2017). This view, however, is regarded as philosophically untenable. Even Descartes had to postulate an external agent — the evil genius — to demonstrate solipsism, since without an external point of reference, the very notion of being wrong loses coherence: one cannot recognize an error without some form of comparison or feedback. This suggests that truth and meaning are not purely internal, but necessarily arise within relational and intersubjective contexts (ibid.). Affectivity, too, is embedded in intersubjective interaction: agents adjust their behavior in response to others, whose reactions shape our strategies, expectations, and relationships. Dumouchel and Damiano thus propose that the mind, as well as emotions and affective processes, exists within a relational space. Emotions, in Dumouchel's view, are not private mental states or internal attributes of the individual, but relational properties emerging from, and functional to, a mechanism of strategic interindividual coordination (Dumouchel 1995, Dumouchel/Damiano 2017).

Interestingly, this reconceptualization of emotions aligns with emerging views of the self, mind, and affectivity that challenge traditional notions of the mind as internal, private, rational, and representational. Rooted in American pragmatism, such perspectives go back to philosophers like William James, who famously inverted the intuitive model of emotion. According to James, emotions are not internal states that precede bodily expression, but rather the bodily expression constitutes and influences the emotional experience itself (Murphy, 1997). Numerous psychological experiments have attempted to demonstrate how bodily or environmental cues influence emotional states (Laird/Lacasse 2014): eye contact can increase feelings of affection (ibd.); the classic Strack, Martin, and Stepper (1988) study showed that holding a pen in one's mouth to simulate smiling could influence emotional evaluation; the "suspension bridge" experiment on misattributed arousal showed that participants were more likely to feel attracted under conditions of physiological stress (Dutton/Aron 1989). Social influences also play a role: for example, simply being told that your heart is beating differently may change your emotional perception (Valins 1966, Taylor 1975). This is consistent with research in behavioral sciences that challenge the idea of the mind as a rational, introspective unit. Asch's conformity experiment (1955) and moral psychology studies suggest that moral judgments are often intuitive rather than deliberative. According to Bem (1972), introspection is often a post-hoc rationalization of behavior, not an access to internal causes. From pragmatist foundations,

radical embodied and enactive approaches to cognition have emerged. Radical embodied cognitive science argues that cognition, perception, and emotion are not internal processes but are grounded in bodily and environmental interaction (Chemero 2013). Similarly, enactive accounts claim that cognition is not individual but arises from dynamic interaction between agents (De Jaegher & Di Paolo 2007). This shift gives new importance to the body and environment in shaping mental life. In psychology, sensorimotor psychotherapy emphasizes how working on the body rather than rational cognition can be effective in trauma treatment (Ogden & Minton, 2000). In robotics, an embodied approach has shown promise: morphology can reduce computational demands and enhance task efficiency (Hoffmann & Pfeifer, 2012); neuromorphic and active perception research suggests that vision and understanding arise through movement and sensorimotor interaction rather than internal mental representation (D'Angelo et al., 2025). At a deeper level, these shifts invite us to reconceptualize the very notion of the self. As seen above, our scientific models are deeply shaped by cultural, social, and moral values. The individualist, rationalist image of the self reflects Western socio-economic and cultural paradigms, particularly those of capitalism and liberalism. By contrast, African approaches to the selfhood emphasize its shared and emergent nature through communal relationships (Jecker/Atuire, 2025). Feminist and care ethics similarly criticize the notion of the atomized individual and promote ideas such as relational autonomy and interdependence (Barclay, 2000).

If we then connect these views to proposals such as Bruno Latour's Actor-Network Theory (ANT), which holds that both human agents and technological artefacts contribute symmetrically to hybrid networks of relations, we arrive at a radically relational and distributed perspective (Latour 2005). Within this framework, robots may be seen as active participants in affective and social interactions, not because they possess personal or inner consciousness, but because subjectivity itself is co-constructed within a relational network. If there is no longer a Cartesian dichotomy, and affectivity is instead distributed across a relational network of various actants, then the affective dimension is no longer confined to human beings. This means that the assumptions underlying the deception argument lose their force: if, in order to avoid deception, it was once required that external emotional expressions match internal emotional states, and that requirement no longer holds, then human–robot interaction is no longer inherently deceptive. As a consequence, social robots can be understood as social partners, and the human–robot relationship can be

framed as a new form of interaction within the human affective ecosystem, rather than as the more controversial replacement for human relationships.

New Ways to Determine Deception

Social robots, therefore, are not deceptive *per se*. But this does not mean they are necessarily non-deceptive either: they can be. Hence, caution is needed, especially because we are dealing with vulnerable users, whose protection is of primary importance. Therefore, we must carefully assess under what specific conditions deception might occur. Looking more closely at the Theory of Affective Coordination, another consequence emerges: what is morally relevant is not the "truth" of another's emotion, understood as correspondence with an internal, private mental state, but the mutual commitment to act within a reciprocal emotional coordination. This standpoint aligns with the pragmatist account of truth, according to which truth is not about an unattainable dimension of objectivity, but about its function in action. (Murphy 1987) As Charles Sanders Peirce put it: "Reality […] consists in the particular sensible effects that things participating in it produce"; or in William James' words: "Ideas […] become true only insofar as they help us to establish a satisfactory relationship with all the other parts of our experience." (ibd.) Bringing these perspectives together, we can redefine deception: an agent deceives us not when its actions fail to correspond to some internal "true" emotional state, but when it betrays our expectations of mutual behavior. Consequently, determining when social robots may be deceptive cannot rely solely on abstract theorizing; it requires bottom-up, context-specific, culturally sensitive, and user-centered research. To this end, future empirical work is needed to explore older adults' emotional expectations toward social robots. While some studies have addressed this question in general terms (Hoppe et al. 2020), more targeted research focusing specifically on affective expectations is still lacking.

This reflects a first, foundational layer of the concept of deception. But several scholars have proposed more nuanced distinctions to better guide both ethical reflection and user protection, particularly in the case of vulnerable populations (Danaher 2020, Umbrello/Natale 2024). Most notably, two categories often emerge: the first is what we have discussed so far – a form of structural or design-based deception, resulting from the robot's social cues or the user's psychological tendencies (such as agency attribution or personification). The second, more serious, form is what Danaher (2020) calls "hidden state deception" and Umbrello and Natale (2024) refer to as "strong deception". This stronger form of deception refers to situations in which a robot deliber-

ately conceals or misdirects attention away from certain capacities or functions, e.g. covertly recording users through a hidden camera (Danaher, 2020), or when users are led to misunderstand the artificial nature of the system (Umbrello/Natale 2024). In these cases, "the deception serves some purpose other than truth, one that is typically to the advantage of the deceiver and the disadvantage of the deceived" (Danaher, 2020). According to Umbrello and Natale, strong deception involves three features: (1) lack of transparency, (2) an intent to mislead, and (3) the potential to undermine user autonomy or control. While the first type of deception (structural or affective) can be meaningfully examined within a philosophical framework, as discussed above, the second type (intentional and harmful deception) may be more appropriately addressed through legal and regulatory instruments.

Ethics by Law in Social Robotics: Preliminary Reflections on Article 5 of the AI Act

Digital technologies evolve at a pace that often exceeds the capacity of legal and regulatory systems to respond effectively. Regulatory frameworks tend to operate on significantly slower structural timelines, leaving many emerging technologies unregulated or only partially addressed. The European Union is one of the few institutional bodies that has formalized ethical principles and translated them into regulatory frameworks for AI technologies, albeit with significant challenges such as the abstract nature of many guidelines and their limited binding power.

The most comprehensive effort in this direction is the AI Act, which was proposed in 2021 and officially entered into force on August 1st 2024 (European Commission, 2024). Some legal scholars working on social robotics have pointed out that the issue of potentially deceptive social robots may fall under Article 5 of the AI Act, which concerns prohibited AI practices (Bertolini, 2024). The article reads as follows:

1. The following AI practices shall be prohibited:
    a. the placing on the market, the putting into service or the use of an AI system that deploys subliminal techniques beyond a person's consciousness or purposefully manipulative or deceptive techniques, with the objective, or the effect of materially distorting the behaviour of a person or a group of persons by appreciably impairing their ability to

make an informed decision, thereby causing them to take a decision that they would not have otherwise taken in a manner that causes or is reasonably likely to cause that person, another person or group of persons significant harm; Related: Recital 29

b. the placing on the market, the putting into service or the use of an AI system that exploits any of the vulnerabilities of a natural person or a specific group of persons due to their age, disability or a specific social or economic situation, with the objective, or the effect, of materially distorting the behaviour of that person or a person belonging to that group in a manner that causes or is reasonably likely to cause that person or another person significant harm; Related: Recital 29. (European Commission, 2024)

The most relevant part of this provision for the present analysis is the explicit prohibition of deceptive technologies, under which we can include including robotic systems, that employ "manipulative or deceptive techniques, with the objective or the effect of materially distorting the behavior of a person". This provision may represent a striking legal translation of the ethical concerns previously discussed, especially those addressed in the philosophical literature on deception, emotion simulation, and relational ethics, since a misplaced emotional attachment could lead to unintended or harmful behavioral changes. As such, philosophical and ethical conclusions remain highly relevant to legal interpretation and implementation. If we accept the ethical arguments explored earlier — particularly those informed by relational theories and the Theory of Affective Coordination — then social robots created to generate feelings of affection in the user should not be considered inherently deceptive, and therefore not inherently subject to the bans outlined in Article 5.Companion robots for the elderly and other vulnerable population, therefore, can be considered generally permissible and not automatically excluded by the regulation.

However, the fact that not all social robots fall under the scope of Article 5 does not imply that none of them do. Following the ethical analysis previously outlined, the lighter and structural form of "design-based deception" must be distinguished from more problematic forms of deception that involve manipulation or exploitation. Given the vulnerability and delicate moral status of the end-users, it is crucial to safeguard their human dignity, ensuring that both no unintended harmful effects occur and no explicit intention to deceive is present. Within the regulatory framework of the AI Act, this highlights the need for a more nuanced evaluation and the development of more specific

guidelines, aimed at more narrowly and precisely identifying which types of companion robots are ethically and legally permissible, and which may raise deeper concerns.

To this end, much work remains to be done. However, even from the embryonic analysis provided here, it is possible to outline some very preliminary guidelines. On the one hand, this concerns the methodological approach. As shown by the value-sensitive design perspective, the integration of moral values can—and should—occur at the very beginning of the design process. The form, appearance, and functionality of a robot can take many different shapes and not necessarily follow the depictions found in dominant cultural narratives. Furthermore, as highlighted by the need for user-centered design approaches, the final say in this technological construction must lie with the end-users themselves; in the case of ECRs, with elderly people. As Sharkey & Sharkey emphasize, "it is important to ensure that robots introduced into elder care do actually benefit the elderly themselves, and are not just designed to reduce the care burden on the rest of society." (Sharkey/Sharkey 2012). It is therefore essential that any design and development processes be informed by, and co-constructed with, elderly users. On the other hand, it is necessary to employ theoretical tools from ethics and philosophy alongside those of the law, using legal constraints to operationalize ethical considerations in ways that maximize the protection of vulnerable populations. For instance, legal constraints can be applied to manufacturers regarding product features, or to distributors and retailers concerning marketing practices—thus minimizing risks related to profit-driven exploitation. This could include, for example, applying specific regulations used for medical devices. By combining the top-down theoretical insights of ethics and philosophy with the bottom-up, user-centered and value-sensitive design approaches, it becomes possible to begin outlining context-sensitive guidelines tailored to different types of deception and specific use cases.

Preliminary Guidelines

Building on the previous analysis, I propose some preliminary regulatory and design guidelines, which need to be further developed and specified in future work. These recommendations follow the above-mentioned distinction between banal and strong deception, to which I suggest adding a third area of concern: commercial exploitation, a particular form of strong deception. Indeed, the main risk posed by social robots is that they may be designed to

exploit users' psychological vulnerabilities for commercial gain, much like social media platforms have proven to do — enabled by privatization, deregulation, monopolies, and disproportionate power distribution. For example, they could manipulate users' attention and emotions to increase engagement or use personalized interactions to influence behavior in ways that primarily serve profit-driven goals rather than users' own interests. This includes tactics such as reinforcing addictive behaviors, subtle persuasion without clear consent, and controversial personalization up to undisclosed advertising.

In light of this, and following the threefold definition of deception, the following guidelines can be outlined, addressed to different stakeholders:

- Avoiding Banal Deception
  - Ensure emotional safety: conduct field studies using user-centered and value-sensitive design focused on users' affective expectations and inform regulation on these findings. (researchers, companies, national and supranational institutions…)
  - Promote understanding of human-robot interaction as a novel relational form and not as a replacement for human-human relations, favoring non-human-like designs in elderly care and discouraging the employment of hyper-realistic anthropomorphic robots. (companies, manufacturers, developers…)
  - In this perspective, design and usage guidelines should ensure that robots are not used exclusively nor perceived as replacements. (lawyers, manufacturers, final users…)
- Avoiding Strong Deception
  - Ensure transparency and privacy protection: all features must be clearly declared, with clear communication to guarantee informed consent as much as possible. (lawyers, companies, retailers and distributors, final users…)
  - Prohibit intentional manipulation (e.g., hidden cameras, covert recording, advertising purposes, and other deceptive practices). (lawyers, institutions, manifacturers…)
- Avoiding Commercial Exploitation
  - Explore reclassification of certain social robots under stricter regulatory categories, such as medical devices, to provide stronger control and protection for vulnerable populations. (lawyers, researchers, institutions…)

- Promote transparency about commercial intents and practices behind social robots. (lawyers, companies, retailers and distributors…)
- Encourage business models through legal boundaries that prioritize user well-being over profit maximization. (lawyers, researchers, institutions…)

Conclusion

Despite the complexity and abstract nature of regulations such as the AI Act, it is nonetheless possible and necessary to specify clear frameworks to guide the ethical governance of technologies. This ensures the protection of end users without resorting to blanket bans on potentially beneficial innovations, such as elderly care and companionship robots. Indeed, legal instruments are crucial to operationalize ethics and embed values within technology, acknowledging that technology is never neutral but always reflects social biases, values, and worldviews. At the same time, technology is neither inevitable nor predetermined. On the contrary, it can be shaped towards multiple possible futures and diverse forms by integrating specific values and purposes into its design and deployment.

In the specific case of Elderly Care Robots (ECRs), particular attention must be paid to the risks of deception and lack of reciprocity, with a focus on human dignity and well-being. Several paths can be pursued to achieve this goal: on the one hand, developing ethical evaluations grounded in rigorous philosophical and theoretical work that incorporate emerging relational perspectives; and on the other, undertaking bottom-up, field-based analyses centered on the interests, care, and lived experiences of end users. The overarching purpose is to translate ethical and empirical insights into robust regulatory measures that effectively guide innovation, while acknowledging the complexity of this task and that substantial work remains to be done.

References

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.

Akrich, M. (1992). The de-scription of technical objects. In: W.E. Bijker and J. Law, eds. Shaping technology/building society. Cambridge, MA: MIT Press, pp.205–224.

Al-Halawani, R., Charlton, P. H., Qassem, M., & Kyriacou, P. A. (2023). A review of the effect of skin pigmentation on pulse oximeter accuracy. *Physiological measurement*, *44*(5), 05TR01.

Allen, C., Wallach, W., Smit, I. (2006). Why machine ethics? IEEE Intelligent Systems, 21(4), pp.12–17.

Alnajjar, F., Bartneck, C., Baxter, P., Belpaeme, T., Cappuccio, M., Di Dio, C., Eyssel, F., Handke, J., Mubin, O., Obaid, M. and Reich-Stiebert, N. (2021). *Robots in education: An introduction to high-tech social agents, intelligent tutors, and curricular tools*. Routledge.

Anderson, M. L., & Felzmann, H. (2024). 3. Ethical complexities within the appearance and usage of social robots: A scoping review. Irish Journal of Applied Social Studies, 24(1), 3.

Angwin, J., Larson, J., Mattu, S., Kirchner, L. (2016). Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. ProPublica, 23 May.

Arendt, H. (2022). The human condition. University of Chicago press.

Asch, S. E. (1955). Opinions and social pressure. Scientific American, 193(5), 31–35.

Barclay, L. (2000). Autonomy and the social self. Relational autonomy, 52–71.

Barocas, S., Hardt, M., Narayanan, A. (2023). *Fairness and machine learning: Limitations and opportunities*. MIT press.

Bartoletti, I. (2020). An artificial revolution: on power, politics and AI. Black Spot Books.

Benjamin, R. (2019). Assessing risk, automating racism. Science, 366(6464), pp.421–422.

Bem, D. J. (1972). Self-perception theory. In: L. Berkowitz, ed. Advances in experimental social psychology, Vol. 6. New York: Academic Press.

Bertolini, A. (2024). The subtle line between personalization and user manipulation in a European regulatory perspective. A proposal for a technology-assessment methodology for Artificial Intelligence Systems. In *2024*

IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 4777–4784). IEEE.

Bijker, W.E. (1995). Of bicycles, bakelites, and bulbs: Toward a theory of sociotechnical change. Cambridge, MA: MIT Press.

Bijker, W.E. (2009). Social construction of technology. In: J.K.B. Olsen, S.A. Pedersen and V.F. Hendricks, eds. A companion to the philosophy of technology. Oxford: Wiley-Blackwell, pp.88–94.

Bijker, W.E., Hughes, T.P., Pinch, T., eds. (1987). The social construction of technological systems. Cambridge, MA: MIT Press.

Botti, C. (2015). Prospettive femministe nel dibattito bioetico contemporaneo. In *Donne, diritto, diritti. Prospettive del giusfemminismo* (pp. 97–115). Giappichelli.

Bucchi, M. (2002). Scienza e società: introduzione alla sociologia della scienza. Bologna: Il Mulino.

Buolamwini, J., Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Conference on fairness, accountability and transparency. PMLR, pp.77–91.

Carrada, G. (2005). Comunicare la scienza: kit di sopravvivenza per ricercatori. Vol. 12. Milano: Alpha Test.

Chemero, A. (2013). Radical embodied cognitive science. Review of General Psychology, 17(2), 145–150.

Coalition for Future Mobility (n.d.) *Benefits of self-driving vehicles*. Available at: https://coalitionforfuturemobility.com/benefits-of-self-driving-vehicles/ (Accessed: 23 June 2025).

Collins, E.C., Millings, A., Prescott, T.J. (2013). Attachment to assistive technology: A new conceptualisation. In: Assistive technology: From research to practice. IOS Press, pp.823–828.

Criado-Perez, C., 2019. Invisible women. New York: Abrams Press.

D'Angelo, G. V., Clerico, V., Bartolozzi, C., Hoffmann, M., Furlong, P. M., Hadjiivanov, A. (2025). Wandering around: A bioinspired approach to visual attention through object motion sensitivity. Neuromorphic Computing and Engineering.

Damiano, L., Dumouchel, P.G. (2020). Emotions in relation. Epistemological and ethical scaffolding for mixed human-robot social ecologies. HUMANA.MENTE Journal of Philosophical Studies, 13(37), pp.181–206.

Danaher, J. (2020). Robot betrayal: A guide to the ethics of robotic deception. Ethics and Information Technology, 22(2), pp.117–128.

Dastin, J. (2018). 'Amazon scraps secret AI recruiting tool that showed bias against women', *Reuters*, 10 October. Available at: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G (Accessed: 23 June 2025).

De Jaegher, H., Di Paolo, E. (2007). Participatory sense-making: An enactive approach to social cognition. Phenomenology and the cognitive sciences, 6, 485–507.

De Rosis, S., Ferrè, F., Pennucci, F. (2022). Including patient-reported measures in performance evaluation systems: patient contribution in assessing and improving the healthcare systems. *The International Journal of Health Planning and Management*, *37*, 144–165.

Dewey, J. (1927). The Public and Its Problems, Henry Holt.

Dumouchel, P. (1995). *Émotions: essai sur le corps et le social*. Synthélabo.

Dumouchel, P., Damiano, L. (2017). Living with robots. Cambridge, MA: Harvard University Press.

Dutton, D. G., Aron, A. (1989). Romantic attraction and generalized liking for others who are sources of conflict-based arousal. Canadian Journal of Behavioural Science, 21(3), pp.246–255.

European Commission (2024). Artificial Intelligence Act (Regulation (EU) 2024/1689).

Fong, T., Nourbakhsh, I., Dautenhahn, K. (2003). A survey of socially interactive robots. Robot Auton Syst 42:143–166. https://doi.org/ 10.1016/S0921-8890(02)00372-X.

Friedman, B., Kahn, P.H. Jr. (2003). Human values, ethics, and design. In: J.A. Jacko and A. Sears, eds. The human–computer interaction handbook. Mahwah, NJ: Lawrence Erlbaum Associates, pp.1177–1201.

Gilligan, C. (1977). In a different voice: Women's conceptions of self and of morality. *Harvard educational review*, *47*(4), 481–517.

Gould, J. D., Lewis, C. (1985). Designing for usability: key principles and what designers think. *Communications of the ACM*, *28*(3), 300–311.

Grasswick, H. (2018). *Feminist Social Epistemology*. In: Zalta, E.N. (ed.) *The Stanford Encyclopedia of Philosophy* (Fall 2018 Edition). Available at: https://plato.stanford.edu/archives/fall2018/entries/feminist-social-epistemology/ (Accessed: 24 June 2025).

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. Psychological Review, 108(4), pp.814–834.

Harding, S. G. (1986). The science question in feminism. Ithaca: Cornell University Press.

Hoffmann, M., Pfeifer, R. (2012). The implications of embodiment for behavior and cognition: animal and robotic case studies. arXiv preprint. arXiv:1202.0440.

Hoppe, J. A., Johansson-Pajala, R. M., Gustafsson, C., Melkas, H., Tuisku, O., Pekkarinen, S., Hennala, L., Thommes, K. (2020). Assistive robots in care: Expectations and perceptions of older people. In: Aging between participation and simulation: Ethical dimensions of socially assistive technologies in elderly care, pp.139–156.

Huyskes, D. (2025). Constructing automated societies: Socio-cultural determinants and impacts of automated decision-making in public services, Università degli Studi di Milano.

Institute for Health Metrics and Evaluation (IHME) (2019). Global Health Data Exchange (GHDx). [online] Available at: https://vizhub.healthdata.org/gbd-results/.

Jecker, N. S., Atuire, C. A. (2025). Authors meet critics: What is a person? Untapped insights from Africa. Journal of medical ethics, 51(4), 249–250.

Jané-Llopis, E., Gabilondo, A., eds. (2008). Mental health in older people: Consensus paper. Luxembourg: European Communities.

Kaplan, F. (2001), January. Artificial attachment: Will a robot ever pass Ainsworth's strange situation test. In *Proceedings of humanoids* (pp. 125–132).

Keulartz, J., Schermer, M., Korthals, M., Swierstra, T. (2004). Ethics in technological culture: A programmatic proposal for a pragmatist approach. Science, Technology & Human Values, 29(1), pp.3–29.

Kittay, E. (1999). Love's labor: Essays on women, equality, and dependency. New York: Routledge.

Laird, J. D., Lacasse, K. (2014). Bodily influences on emotional feelings: Accumulating evidence and extensions of William James's theory of emotion. Emotion Review, 6(1), pp.27–34.

Latour, B. (1987). Science in action: How to follow scientists and engineers through society. Cambridge, MA: Harvard University Press.

Latour, B. (2005). Reassembling the social: An introduction to actor-network-theory. Oxford University press.

Lee, O. E., Nam, I., Chon, Y., Park, A., Choi, N. (2023). Socially assistive humanoid robots: Effects on depression and health-related quality of life among low-income, socially isolated older adults in South Korea. *Journal of Applied Gerontology*, 42(3), 367–375.

Lorenz, K (1970). Studies in animal and human be haviour. Harvard University Press.

Lu, L. C., Lan, S. H., Hsieh, Y. P., Lin, L. Y., Lan, S. J., Chen, J. C. (2021). Effectiveness of companion robot care for dementia: a systematic review and meta-analysis. *Innovation in aging*, *5*(2), igab013.

Murphy, J. P. (1997). Il pragmatismo. Bologna: Il Mulino.

Neri, D. (2020). *Filosofia morale. Manuale introduttivo*. goWare & Guerini editore.

Obermeyer, Z. et al., 2019. Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), pp.447–453.

OECD (2020), *Who Cares? Attracting and Retaining Care Workers for the Elderly*, OECD Health Policy Studies, OECD Publishing, Paris, https://doi.org/10.1787/92c0ef68-en.

Ogden, P., Minton, K. (2000). Sensorimotor psychotherapy: One method for processing traumatic memory. Traumatology, 6(3), 149–173.

Ogburn, W.F., Thomas, D. (1922). Are inventions inevitable? A note on social evolution. Political Science Quarterly, 37(1), pp.83–98.

Oudshoorn, N. (2003). *The male pill: A biography of a technology in the making*. Duke University Press.

Parks, J. A. (2010). Lifting the burden of women's care work: Should robots replace the "human touch"? Hypatia, 25(1), pp.100–120.

Pu, L., Moyle, W., Jones, C., Todorovic, M. (2019). The effectiveness of social robots for older adults: a systematic review and meta-analysis of randomized controlled studies. *The gerontologist*, *59*(1), e37-e51.

Rawls. J. (1993), Political Liberalism, Columbia University Press.

Robinson, H., Macdonald, B., Kerse, N., Broadbent, E. (2013). [Various contributions]. Journal of the American Medical Directors Association, 14. DOI: 10.1016/j.jamda.2013.02.007.

Sanders, E. B. N., Stappers, P. J. (2008). Co-creation and the new landscapes of design. *Co-design*, *4*(1), 5–18.

Sharkey, A., Sharkey, N. (2012). Granny and the robots: Ethical issues in robot care for the elderly. Ethics and Information Technology, 14, pp.27–40.

Stinson, C. (2022). Algorithms are not neutral: Bias in collaborative filtering. AI and Ethics, 2(4), pp.763–770.

Sparrow, R., Sparrow, L. (2006). In the hands of machines? The future of aged care. Minds and Machines, 16, pp.141–161

Strack, F., Martin, L. L., Stepper, S. (1988). Inhibiting and facilitating conditions of the human smile: A nonobtrusive test of the facial feedback hypothesis. Journal of Personality and Social Psychology, 54(5), pp.768–777.

Surden, H. (2017). Values embedded in legal artificial intelligence. University of Colorado Law Legal Studies Research Paper, No. 17–17. SSRN.

Tamura, T., Yonemitsu, S., Itoh, A., Oikawa, D., Kawakami, A., Higashi, Y., Nakajima, K. (2004). Is an entertainment robot useful in the care of elderly people with severe dementia? *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, *59*(1), M83-M85.

Taylor, S. E. (1975). Relationship between physiological feedback and the attribution of emotion. Journal of Personality and Social Psychology, 31(2), pp.306–314.

Tengai (n.d.) Our story. Available at: https://tengai.io/our-story/ (Accessed: 23 June 2025).

Thaler, R. H., Sunstein, C.R. (2009). Nudge. London: Penguin.

Tripaldi, L. (2023). Gender tech: Come la tecnologia controlla il corpo delle donne. Roma: Gius. Laterza & Figli.

Tronto, J. (1993). Moral boundaries: A political argument for an ethic of care. New York: Routledge.

Turkle, S., Taggart, W., Kidd, C. D., Dasté, O. (2006). Relational artifacts with children and elders: The complexities of cybercompanionship. Connection Science, 18(4), pp.347–361.

Turkle, S. (2007). Authenticity in the age of digital companions. *Interaction studies*, *8*(3), 501–517.

Umbrello, S., Natale, S., 2024. Reframing deception for human-centered AI. *International Journal of Social Robotics*, *16*(11), 2223–2241.

United Nations (2015). Transforming our world: The 2030 Agenda for Sustainable Development. Resolution adopted by the General Assembly on 25 September &2015, A/RES/70/1, pp.1–13.

Valins, S. (1966). Cognitive effects of false heart-rate feedback. Journal of Personality and Social Psychology, 4(4), pp.400–408.

van Grunsven, J., van Wynsberghe, A. (2019). A semblance of aliveness: How the peculiar embodiment of sex robots will matter. *Techné: Research in Philosophy and Technology*.

van Grunsven, J. (2022). Anticipating sex robots: A critique of the sociotechnical vanguard vision of sex robots as 'good companions'. In: I. van de Poel and S. Royakkers, eds. Being and value in technology. Cham: Springer, pp.63–91.

van Grunsven, J., Stone, T., Marin, L. (2024). Fostering responsible anticipation in engineering ethics education: how a multi-disciplinary enrichment of the responsible innovation framework can help. *European journal of engineering education*, *49*(2), 283–298.

van Wynsberghe, A. (2012). Designing robots for care: Care centered value-sensitive design. Science and Engineering Ethics, 19(2), pp.407–433.

van Wynsberghe, A. (2022). Social robots and the risks to reciprocity. AI & Society, 37(2), pp.479–485.

Winner, L., (1980). 'Do Artifacts Have Politics?', *Daedalus*, 109(1), pp. 121–136.

World Health Organization (WHO) (1946). Constitution of the World Health Organization. [online] Available at: https://www.who.int/about/governance/constitution

World Health Organization (WHO) (2007). People centred health: A framework for policy. [online] Available at: http://www.wpro.who.int/NR/rdonlyres/

World Health Organization (WHO) (2023). Mental health of older adults. [online] Available at: https://www.who.int/news-room/fact-sheets/detail/mental-health-of-older-adults

World Health Organization (WHO) (2024). Ageing and health. [online] Available at: https://www.who.int/news-room/fact-sheets/detail/ageing-and-health

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).