The Impossibility of Transparent Social Robots[a]

Giovanna Di Cicco[*]

Abstract

La trasparenza è emersa come uno dei concetti più rilevanti nel dibattito etico che circonda diversi ambiti, tra cui la robotica sociale. Questo articolo esplora il modo in cui la trasparenza si applica ai robot sociali e se possa essere uno strumento efficace per proteggere gli interessi degli utenti da potenziali inganni e dinamiche ambigue implicate nelle interazioni tra esseri umani e robot. L'articolo traccia una distinzione preliminare tra la trasparenza intesa come proprietà della robotica sociale e la trasparenza intesa come attributo dei robot sociali, evidenziandone i diversi significati e implicazioni. La discussione si concentra poi sulla trasparenza dei robot sociali e viene fatta un'ulteriore distinzione tra *trasparenza sui robot sociali* e *trasparenza attraverso i robot sociali*. Partendo dalla descrizione dei tre tipi di inganno proposti da John Danaher, l'inganno di stato interno, messo in atto da robot sociali che mostrano facoltà e stati emotivi che in realtà non hanno, viene identificato come la forma più costitutiva di inganno coinvolta nelle interazioni con i robot sociali. Questo aspetto viene poi considerato alla luce dell'antropomorfismo, per esaminare la progettazione di robot trasparenti, che dovrebbero attenuare le risposte antropomorfiche come possibile rimedio per proteggere gli interessi degli individui ed evitare l'inganno. Tuttavia, poiché l'antropomorfismo sembra essere il fondamento stesso della socialità percepita dai robot, è impossibile rinunciare al loro comportamento ingannevole senza rinunciare anche al loro ruolo sociale. Ciò porta, infine, a sostenere che un robot sociale veramente trasparente non è realizzabile e che la trasparenza non è sufficiente a garantire una robotica sociale responsabile.

Parole chiave: robot sociali, trasparenza, antropomorfismo, pregiudizi cognitivi, inganno dei robot, teoria della mente, interazione umano-robot, etica della tecnologia, roboetica, implicazioni etiche.

[a] Saggio ricevuto in data 30/05/2024 e pubblicato in data 22/01/2025.

[*] Dottoranda, Università degli Studi di Genova – Northwest Italy Philosophy PhD Program (FINO), email: giovanna.d.cicco@gmail.com.

Transparency has emerged as one of the most relevant concepts in the ethical debate surrounding several fields, and social robotics is one of them. This paper explores how transparency relates to social robots and whether it could be an effective tool to protect users' interests from potential deception and misleading dynamics involved in human-robot interactions. The paper outlines a preliminary distinction between transparency understood as a property of social robotics and transparency understood as an attribute of social robots, highlighting their different meanings and implications. The discussion, then, focuses on the transparency of social robots, where a further distinction is drawn between *transparency on social robots* and *transparency through social robots*. Starting from the description of three types of deception proposed by John Danaher, internal state deception, enacted by social robots that display faculties and emotional states they do not really have, is identified as the most constitutive form of deception involved in interactions with social robots. This is then considered in the light of anthropomorphism, to examine the design of transparent robots, which should mitigate the anthropomorphic responses as a possible remedy to protect the interests of individuals and avoid deception. However, since anthropomorphism appears to be the very foundation of robots' perceived sociality, it is impossible to forego their deceptive behaviour without also foregoing their social role. This leads, finally, to argue that a genuinely transparent social robot is not achievable, and that transparency is not enough to ensure a responsible social robotics.

Keywords: social robots, transparency, anthropomorphism, cognitive bias, robot deception, theory of mind, human-robot interaction, ethics of technology, roboethics, ethical implications.

Introduction

Recent advancements in the field of social robotics and artificial intelligence are bound to change the way human beings engage with reality and with each other. This novel context challenges the traditional tools we use to understand others' behaviour and discern between genuine and fake, reality and simulation. While regulations and institutions seem to struggle to cope with evolving technologies and to defend the interests of users, the ethical debate faces unprecedented issues and questions.

One main ethical concern raised about social robotics is that human-social robot interactions might be inherently deceptive and inauthentic, as they provide the illusion of robots being something they are not and having attributes they do not actually have. Although social robots are not necessarily humanoid or human-like, they usually display evocative features, such as certain facial expressions, proxemic and postural attitudes or vocal tones. They are designed to perform tasks focused on interacting with human beings, behaving as credible social actors, and eliciting empathic and emotional reactions. Therefore, they behave *as if* they had emotions, intentions, preferences, or goals, where the words "*as if*" precisely reflect the dimension of simulation involved. Some of the risks of such deception are those related to privacy and information sharing, the building of unidirectional bonds or trust, the misinterpretation of robots' behaviour, and the information and power asymmetry between users and companies[1].

Understanding the implications of deceptive practices involved in social robotics and developing strategies to defend the interests of users has become a key issue in the technology ethics debate. *Transparency* has then emerged as an increasingly relevant concept and one of the most advocated strategies[2]. Indeed, the AI HLEG, established by the European Union, identifies it as one of the principles for a sustainable and trustworthy artificial intelligence[3].

However, the great success of this concept comes with an equal amount of uncertainty regarding its understanding and applications. This results in what Emmanuel Alloa describes as a *magic concept*, characterised by a great normative attractiveness and an exceedingly positive connotation, yet presenting multiple and overlapping definitions[4]. As scholars point out, despite the relevance assigned to transparency in the ethical debate on social robotics, there is currently a lack of an extensive literature or agreed definitions[5].

Therefore, in order to understand whether transparency is a suitable and sufficient strategy to avoid deception and ensure a sustainable development for social robots, this paper will try to clearly define what transparency means for social robotics, how it interacts with different forms of deception perpetrated by social robots, and what are the limits of its application.

1. *Why transparency and which transparency*

Firstly, it is worth noting that the relevance of transparency in the technology ethics debate is part of a broader flourishing of this concept in the 21st century. Such notion has been regarded as a socio-political tool to serve democracy, playing a major role in the fight against corruption and stimulating responsible and informed decision-making.[6]. However, Alloa highlighted that transparency can be associated with

[1] R. Wullenkord & F. Eyssel, *Societal and Ethical Issues in HRI*, in «Current Robotics Reports», 1, 2020, pp. 85-96.

[2] See S. Turkle, *Authenticity in the age of digital companions*, in «Interaction Studies: Social Behaviour and Communication in Biological and Artificial Systems», 8, n. 3, pp. 501-517; P.G.R. de Almeida, C.D. dos Santos & J.S. Farias, *Artificial Intelligence Regulation: a framework for governance*, in «Ethics and Information Technology», 23, 2021, pp. 505-525; and A. Jobin, M. Ienca & E. Vayena, *The global landscape of AI ethics guidelines*, in «Nature Machine Intelligence», 1, 2019, pp. 389-399.

[3] AI HLEG - High-Level Expert Group on Artificial Intelligence, *The Assessment List for Trustworthy Artificial Intelligence* (ALTAI), 17 July 2020.

[4] E. Alloa, *Transparency: A magic concept of modernity*, in E. Alloa & D. Thomä (eds.), *Transparency, Society, Subjecticity. Critical Perspectives*, Palgrave Macmillan, London, 2018, pp. 21-55: 29.

[5] A. Theodorou, R.H. Wortham & J.J. Bryson, *Designing and implementing transparency for real time inspection of autonomous robots*, in «Connection Science», 29, n. 3, pp. 230-241.

[6] J.C. Bertot, P.T. Jaeger, J.M. Grimes, *Using ICTs to create a culture of transparency: E-government and social media as openness and anti-corruption tools for societies*, in «Government Information Quarterly», 27, n. 3, 2010, pp. 264-271: 264.

different aspirations and contexts, such as gaining access to information, safeguarding justice, providing accountability, and encouraging virtuous conduct.[7]. Therefore, when we talk about transparency in social robotics, we can group the various nuances of the concept into two main original understandings: *transparency of social robotics*, as an attribute of such field of research and production, and *transparency of social robots*, as an attribute of social actors.

Transparency of social robotics refers more precisely to information transparency, achieved through disclosure procedures that allow information, data, or behaviour to be visible and understandable. It thus relates to the possibility for governments and general public to see practices and activities underlying the research and production of social robots. Some authors have emphasised, for instance, the importance of knowing the working algorithms and rules of AI[8] and the methods and data that have been used in their training[9]. Other significant information concern how the data collected by social robots is processed, the business operations of the producing companies, the power dynamics in which they are involved and the operational and procedural processes of their activities.

On the one hand, this openness provides a greater understanding of robotic technologies and related risks, allowing individuals to be more conscious in their use. On the other hand, it exposes the actions of companies to the judgement of public and laws, allowing for a greater scrutiny of their legitimacy and to hold them accountable for their decisions. At the same time, it seems to be a potential instrument of moralisation and self-regulation that could induce restraint and best practices[10]. If companies are forced to provide details and reasons for their actions, then what they do is there for all to see and they are much more likely to act in a virtuous manner.

Understanding transparency in this way does not present unique peculiarities related to social robotics and can be traced back to the debate on transparency in the socio-political perspective. Moreover, it is worth noting that Transparency as an alternative to stricter regulation or as a regulation in itself has been questioned and shows limits in its application and outcomes[11].

Transparency of social robots, instead, refers to the user's ability to clearly understand the artificial social partner, so as to accurately grasp the functioning of the robot he or she interacts with. On a pragmatic level, this could foster human-robot cooperation by ensuring a safer and more effective use of the robot, such as knowing when to consider it reliable or is acting unexpectedly, how it makes decisions, or how

[7] E. Alloa, *Transparency: A magic concept of modernity*, cit., pp. 31-32.

[8] M.C. Buiten, *Towards Intelligent Regulation of Artificial Intelligence*, in «European Journal of Risk Regulation», 10, n. 1, 2019, pp. 41-59.

[9] M. Butterworth, *The ICO and artificial intelligence: The role of fairness in the GDPR framework*, in «Computer Law & Security Review», 34, n. 2, 2018, pp. 257-268.

[10] E. Alloa, *Seeing Through a Glass Darkly. The Transparency Paradox*, in E. Alloa (Ed), *This Obscure Thing Called Transparency. Politics and Aesthetics of a Contemporary Metaphor*, Leuven University Press, Leuven, 2022, pp. 9-25: 12.

[11] A. Etzioni, *The Limits of Transparency*, in E. Alloa & D. Thomä (eds.), *Transparency, Society, Subjecticity. Critical Perspectives*, Palgrave Macmillan, London, 2018, pp. 179-201.

to interpret its behaviour.[12]. On a more ethically relevant level, the main benefit of transparent robots is considered to be the elimination, or reduction, of users' deception, making him less vulnerable to the possible exploitation of his trust and emotional response.

Schött, Amin and Butz points out that transparency of social robots can be understood either as *transparency on the robot*, where information is provided from outside, or as transparency *through the robot*, where information is integrated into the design itself[13]. Transparency on the robot can be achieved, for instance, through what is conveyed by marketing, advertising, instruction manuals or websites. Transparency through the robot refers to constitutive elements embedded in the robot design itself, which thus becomes the source of transparent information. This can be done in an explicit manner, such as by having the robot remind the user that it is an artificial entity, that it has no feelings, that it does not belong to any gender or that it cannot answer questions about its pretended past or its emotional states[14]. But it can also occur implicitly, when the robot is constructed in such a way that it does not resemble a human being, when it appears obviously mechanical, or has a voice clearly recognisable as artificially synthesised[15].

In both transparency on the robot and transparency through the robot, then, the goal is to provide a look at the reality that lies beyond the social appearance of the artificial agent, beyond the *as if* it performs. To aim for a transparent robot means to aim for a robot that is accurately perceived by the user, who can clearly recognise its artificial nature and real properties. In this way, transparency would become the way to avoid manipulation of users and preserve them from developing inappropriate and one-sided emotional responses or bonds. However, to understand whether this is the case, it is worth investigating what we mean when we talk about deception involved in interactions with social robots and how this responds to attempts at transparency.

2. *Social robots, deception and anthropomorphism*

There are several ways for a robot to deceive users and John Danaher has specifically outlined three[16].

1) *External state deception* occurs when the robot deceives the user on something that does not concern the robot itself by providing false information. As Danaher points out, external state deception is similar to cases where humans lie, so it follows

[12] A. Theodorou, R.H. Wortham & J.J. Bryson, *Designing and implementing transparency for real time inspection of autonomous robots*, cit., pp. 232-234.

[13] S.Y. Schött, R.M. Amin, R.M. Butz, *A literature survey of how to convey transparency in co-located human-robot interaction*, in «Multimodal Technol. Interact.», 7, n. 25, 2023, p. 9.

[14] B. Leong & E. Selinger, *Robot Eyes Wide Shut: Understanding Dishonest Anthropomorphism*, in «Proceedings of the Association for Computing Machinery's Conference on Fairness, Accountability, and Transparency», 2009, pp. 299-308: 307.

[15] C. Balkenius & B. Johansson, *Almost Alive: Robots and Androids*, in «Frontiers in Human Dynamics», 4, 2022, pp. 1-7: 5-6.

[16] J. Danaher, *Robot Betrayal: a guide to the ethics of robotic deception*, «Ethics Inf. Technol.», 22, 2020, pp. 117-128: 121.

the same moral principles[17]. Therefore, in this context transparency, in terms of information about the design process, can be crucial in ensuring that artificial agents are constructed in such a way that they never lie to the user.

2) *Hidden state deception*, instead, occurs when the robot possesses certain capabilities, but it keeps them hidden from users by omitting or denying them. They might include hidden recording devices or undeclared personal data retention. Again, transparency on the robot plays a key role here, as it allows users to have full knowledge about all the functionalities of the robot they interact with. Users should be aware of the possibility of audio or video recording, of how the robot handles personal information it collects, and what kind of physical force it may exert. As with the previous case, we do not consider it morally acceptable to take advantage of trust or naivety of individuals in order to covertly act against their interest, and neither should it be acceptable for social robots.

In both cases, if users have access to comprehensive and meaningful information about the robot, they are given the tools to rationally choose the best way to deal with it.

3) *Superficial state deception*, finally, includes cases where the robot pretends to have abilities or internal states that it does not actually possess. This form of deception is particularly relevant for social robots, since their very ability to pose as social actors, and create relationships with humans, relies on the simulation of feelings and emotional states capable of evoking an empathic response. Such simulation is not always necessarily a malicious tactic against users, but represent a fundamental design element, which is ultimately useful for the legitimate tasks the robot is designed to perform. In recent years, studies of human-robot interaction (HRI) have played a vital role in understanding how individuals respond to artificial agents and how to improve their interactions so that they become as friendly and natural as possible[18]. To understand how and whether transparency can be useful in defending individuals against superficial state deception, it is then imperative to look at the cognitive and behavioural dynamics it involves.

3. Superficial state deception is not a choice

HRI pragmatic experiments have shown that humans tend to apply to interactions with robots the same social norms and inferences they apply to interactions with living beings[19]. Subjects were shown to attribute meaning and intention to the behaviour of

[17] *Ibid.*

[18] L.T. Cordeiro Ottoni & J. de Jesus Fiais Cerqueira, *A Review of Emotions in Human-Robot Interaction*, 2021 Latin American Robotics Symposium (LARS), 2021 Brazilian Symposium on Robotics (SBR), and 2021 Workshop on Robotics in Education (WRE), Natal, Brazil, 2021, pp. 7-12.

[19] C. Nass, J. Steuer & E.R. Tauber, *Computers are Social Actors*, in «Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 1994)», Association for Computing Machinery, New York, NY, USA, pp. 72–78.

social robots[20] and to place them within social categories, to the point of transferring prejudices or notions, such as gender or identity, onto them[21]. By encouraging such tendencies, social robotics hopes to produce artifacts capable of performing roles traditionally reserved to conscious beings, such as those of care robots, hospitality robots or sex robots. At the same time, it seems to somehow encourage subjects' inaccurate representation of reality, potentially leading them to experience inauthentic relationships.

This has led some authors to believe that social robotics engages in outright deception to the detriment of the user, leading them to indulge in empathic and emotional attachments that are not justified. Robert Sparrow describes it as an excessive sentimentalism of users, who are induced to violate the *prima facie* duty to pursue an accurate representation of reality[22]. The perspective held by Sparrow, however, is challenged by Mark Coeckelbergh, who suggests considering the illusion carried on by social robots not as a deception but as a performance, similar to the one of magic shows[23]. In this view, thus, designers and users are in a relationship resembling the one between magicians and their audience, where they cooperate in maintaining the illusion they voluntarily take part in.

Both sides of the argument, anyway, consider the successful illusion operated by social robots to some voluntary disposition of the subject and thus fail to grasp the problematic core of the issue. HRI studies point out that the creation of empathic and emotional bridges with social robots largely rests on human cognitive mechanism of anthropomorphism: the tendency to attribute human properties and mental states, such as emotions, motivations, or intentions, to nonhuman entities. This emerges both through the analysis of behavioural results and by looking at neurophysiological findings and brain activity reports[24]. Understanding the dynamics of anthropomorphism, then, highlights that the illusion underlying human-robot interactions is not resulting from a choice, but from a conditioned response.

[20] E. Schellen, F. Bossi & A. Wykowska, *Robot Gaze Behavior Affects Honesty in Human-Robot Interaction*, in «Front. Artif. Intell.», 4, 2021; M. Salem, F. Eyssel, K. Rohlfing, S. Kopp & F. Joublin, *To Err is Human(-like): Effects of Robot Gesture on Perceived Anthropomorphism and Likability*, in «Int. J. Soc. Robot», 5, 2013, pp. 313- 323.

[21] See S.J. Stroessner & J. Benitez, *The Social Perception of Humanoid and Non-Humanoid Robots: Effects of Gendered and Machinelike Features*, in «International Journal of Social Robotics», 11, 2019, pp. 305-315; J. Bernotat, F. Eyssel & J. Sachse, *The (Fe)male Robot: How Robot Body Shape Impacts First Impressions and Trust Towards Robots*, in «International Journal of Social Robotics», 13, 2021, pp. 477-489.

[22] R. Sparrow, *The March of the robot dogs*, in «Ethics and Information Technology», 4, n. 4, 2002, pp. 305-318.

[23] M. Coeckelbergh, *How to describe and evaluate "deception" phenomena: recasting the metaphysics, ethics, and politics of ICTs in terms of magic and performance and taking a relational and narrative turn,* in «Ethics Inf. Technol.», 20, 2018, pp. 71-85.

[24] G. Di Cesare, F. Vannucci, F. Rea, A. Sciutti & G. Sandini, *How attitudes generated by humanoid robots shape human brain activity,* in «Scientific Reports», 10, n. 16928, 2020.

4. *Anthropomorphism and theory of mind: how social robots trick us*

Anthropomorphism is usually regarded as a cognitive *bias* that leads us to misinterpret the behaviour of nonhumans, inferring inaccurate causes for it. As human beings, we tend to detect something human in every thing, and this constitutes a structural element of how our minds work, which can be even found in ancestral forms, such as pareidolia[25].

Anthropomorphism is understood as a part, typically considered improper, of the more general human faculty of adopting others' point of view and to imagine what they might be feeling or thinking. This ability, often called "theory of mind" (ToM), refers to the possibility of creating explicit meta-representations of others' mental states, inferring beliefs, motivations, or goals, so that their condition can be evaluated according to their own parameters[26]. To imagine how others might experience a certain situation is different from imagining how we will experience that same situation[27]. It is this very act of perspective taking that underlies the peculiarity of empathic experiences in human beings.

Empathy can be understood as a neurobehavioural process with evolutionary underpinnings, emerging in a variety of human and non-human animals, and consisting of a spontaneous response to specific external stimuli[28]. However, human empathy seems to extend to more situations and individuals. Sometimes we experience empathy for strangers, individuals who are distant in time and space, or even fictional characters and animals of other species. When we adopt the other's point of view, it is no longer relevant who we are or what relationship we have with our social counterparts. ToM allows us to access what lies behind their external behaviours through a spontaneous inferential mechanism, which associates those behaviours with the inner states generating them[29]. However, since we can never have immediate access to internal states of others, this inference is inevitably grounded in our own personal existence, in how we experience those internal states as individuals and as human beings.

On the one hand, this implies that, although excessive human-likeness is shown to produce a feeling of uncanny and discomfort[30], a general resemblance to human beings is more likely to trigger anthropomorphism. Indeed, HRI research

[25] L.F. Zhou & M. Meng, *Do you see the 'face'? Individual differences in face pareidolia*, in «Journal of Pacific Rim Psychology», 14, 2020.

[26] V. Stone, *The moral dimensions of human social intelligence*, in «Philosophical Explorations: An International Journal for the Philosophy of Mind and Action», 9, n. 1, pp. 55-68.

[27] C.D. Batson, S. Early & G. Salvarani, *Perspective taking: Imagining how another feels versus imagining how you would feel*, in «Personality and Social Psychology Bulletin», 23, n. 7, 1997, pp. 751-758.

[28] J. Decety, G.J. Norman, G.G. Berntson & J.T. Cacioppo, *A neurobehavioral evolutionary perspective on the mechanisms of underlying empathy*, in «Prog. Neurobiol.», 98, 2012, pp. 38-48.

[29] V. Stone, *The moral dimensions of human social intelligence*, op. cit.

[30] M. Mori, K.F. Macdorman & N. Kageki, *The Uncanny Valley*, in «IEE Robotics & Automation Magazine», 19, n. 2, 2012, pp. 98-100.

reports that we relate, empathise, and trust artificial agents more easily when they have humanoid attributes, both on an aesthetic and behavioural level[31].

On the other hand, the more distant the entity is from us, either socio-culturally or evolutionarily, the more likely it is that the inference is inaccurate and that he or she experiences or externalizes those inner states differently from us. But whereas in the case of other living beings we are faced with the unrealizable attempt to understand the nature of their inner life, social robots are produced by us, so we may know how they operate.

When it comes to social robots, anthropomorphism leads us to an incorrect inference, since, as Paul Dumouchel points out, their behaviours cannot really be interpreted as read-outs of any internal state but are «signs without referents»[32]. The empathic and emotional response of individuals towards social robots does not depend on a defect of reason or will, nor on false beliefs or intentional participation in an illusory reality[33]. Subjects involved in empirical experiments and users of social robots are aware that they are dealing with machines without internal states but tend to respond empathically to them regardless. Such responses, therefore, cannot be seen as a choice. Instead, they are to be understood as spontaneous and pre-reflexive cognitive mechanisms, which are deliberately elicited through specific design and marketing strategies.

5. The impossible transparent robot: inherent limits of transparency in social robotics

Having clarified the effects of anthropomorphically inspired design on our cognitive biases, we can evaluate how effective transparency of social robots is in preventing individuals from being manipulated.

Transparency on the robot can certainly be a regulatory requirement to ensure that companies and research do not convey misleading information and provide a truthful representation of social robots, at least on a theoretical level. Authors have often expressed this need and highlighted problematic human-washing (or machine-washing) practices carried out by companies[34]. In analogy to the concept of greenwashing, human-washing describes the strategy of companies to deliberately manipulate their communications by creating a symbolic veil; a misleading façade that generates information asymmetry and portrays social robots as more competent, harmless, or similar to us. Demanding transparency from companies about the real properties of social robots, therefore, means removing the opacity of this façade,

[31] See M. Li & A. Suh*, Machinelike or Humanlike? A literature Review of Anthropomorphism in AI-Enabled Technology*, in «Hawaii International Conference on System Sciences», 2021; A. Sacino et al., *Human- or object-like? Cognitive anthropomorphism of humanoid robots*, in «PLoS ONE», 17, n. 7, 2022.

[32] P. Dumouchel, *Making Faces*, in «Topoi», 41, 2022, pp. 631-639: 637.

[33] L. Damiano, P. Dumouchel, *Anthropomorphism in Human-Robot Coevolution*, in «Frontiers in Psychology», 9, n. 468, 2018.

[34] G. Scorici, M.D. Schultz & P. Seele, *Anthropomorphization and beyond: conceptualizing humanwashing of AI-enabled machines*, in «AI & Society», 39, pp. 789-795, 2024.

making it transparent, so that we can access the reality it conceals. However, two considerations should be taken into account.

First, as Alloa points out, just because a medium is transparent does not mean that there is no mediation[35]. Transparency of information is not a given property but something that is made, the result of a process that is never ethically neutral[36]. When information is disclosed, someone has decided on which information, as well as how to interpret and elaborate it to make it understandable and accessible. This process needs to be understood and regulated, so that it does not become a new façade for the sake of marketing.

Second, we have observed that accurate theoretical knowledge is not sufficient to empirically avoid the emergence of misleading dynamics between individuals and social robots. In a sense, we are evolutionarily programmed to interpret robot behaviour through the lens of anthropomorphism.

Transparency through the robot is thus proposed as a strategy to mitigate the effects of anthropomorphism throughout the interaction itself. One example of this strategy has been highlighted by van Straten and Kühne in a study on the interaction between children and social robots, where children's tendency to anthropomorphise and trust robots was found to decrease when they interacted with robots consistently communicating the absence of human psychological capacities[37].

Illusion of transparency, in social psychology, refers to the biased perception that our internal experiences are more visible to others than they really are and that others can perceive our actual personal thoughts, emotions, or mental states. Human beings are never transparent, but we have access to their internal experience through the correct interpretation of their behaviour. Therefore, applying the same notion to social robots, we can conclude that the more the robot is transparent, the more our interpretation of its behaviours should allow us to perceive its lack of inner experience. The design of an entirely transparent robot, then, should convey by the interaction itself that those behaviours are mere simulacra.

However, as seen above, the possibility for the robot to be perceived as a social actor and to create an empathic and engaging interaction is based on the very triggering of anthropomorphism. This means that transparency and perceived sociality of the robot are inversely proportional and to forego the design of social robots with misleading features triggering anthropomorphism is to forego the design of social robots altogether. A genuine transparent social robot is hence not really possible.

[35] E. Alloa, *Transparency: A magic concept of modernity*, op. cit., p. 36.

[36] M. Turilli & L. Floridi, *The ethics of information transparency*, in «Ethics. Inf. Technol.», 11, pp. 105-112: 109.

[37] C.L. van Straten, J. Peter, R. Kühne, *Transparent robots: How children perceive and relate to a social robot that acknowledges its lack of human psychological capacities and machine status*, in «Int. J. Human-Computer Studies», 177, 2023.

Conclusion

In the end, transparency alone does not appear to be able to protect individuals from one of the major sources of exploitation risk: the one that comes from directly interacting with social robots and misinterpreting their behaviours. In fact, since such interactions are mediated by the pre-reflective cognitive mechanism of anthropomorphism, transparency of information about how the robot is constituted does not prevent us from empathising with it and ascribing it mental states it does not possess.

I argue that we should embrace this impossibility and use such awareness to shape adequate strategies for regulating social robots in the world. The use of social robots has been shown to be a potential resource in controlled settings, such as in investigating the functioning of human relationships or in treating social disorders[38]. And there might be other cases where social robots are beneficial, so much so that we agree «to conscientiously harness our weird sensibilities so that our instinctual responses work for us and not against our best interests»[39]. If we accept that some degree of deception is always involved in the relationships between humans and social robots, we can begin to engage in discussions about whether, when and how such deception is something we are willing to allow as a society.

This means further investigating the limits of transparency and identifying the empirical consequences of human-robot interactions that are not free of deception. Additional studies and cognitive experiments are needed to determine the potentially disruptive effects of manipulating cognitive biases on how we interact with each other and the possible benefits that might emerge from deception in specific settings. Furthermore, an interdisciplinary dialogue between HRI, engineering, cognitive science and ethics needs to be developed in order to reach a coherent definition of transparency and a viable implementation strategy. Finally, governments and institutions need to produce strong regulations where transparency is not the goal, but a tool to ensure that social robotics meets the standards of such regulations. To do so, transparency is necessary, but is not enough. Instead, we need to regulate how social robot should be designed, for what purpose, and what the alternatives are.

[38] A. Kouroupa, et al., *The use of social robots with children and young people on the autism spectrum: A systematic review and meta-analysis*, in «PLoS One», 17, n. 6, 2022.

[39] B. Leong & E. Selinger, *Robot Eyes Wide Shut: Understanding Dishonest Anthropomorphism*, in «Proceedings of the Association for Computing Machinery's Conference on Fairness, Accountability, and Transparency», 2009, pp. 299-308: 308.