ARTICLE

Check for updates

Mitigating emotional risks in human-social robot interactions through virtual interactive environment indication

Aorigele Bao[1,2], Yi Zeng[2,3,4 ✉] & Enmeng Lu[3,4]

Humans often unconsciously perceive social robots involved in their lives as partners rather than mere tools, imbuing them with qualities of companionship. This anthropomorphization can lead to a spectrum of emotional risks, such as deception, disappointment, and reverse manipulation, that existing approaches struggle to address effectively. In this paper, we argue that a Virtual Interactive Environment (VIE) exists between humans and social robots, which plays a crucial role and demands necessary consideration and clarification in order to mitigate potential emotional risks. By analyzing the relational nature of human-social robot interaction, we discuss the connotation of such a virtual interactive environment that is similar to the emotional states aroused when reading novels. Building on this comprehension, we further demonstrate that manufacturers should carry out comprehensive Virtual Interactive Environment Indication (VIEI) measures during human-social robot interaction with a stricter sense of responsibility when applying social robots. Finally, we contemplate the potential contributions of virtual interactive environment indication to existing robot ethics guidelines.

[1] Department of Philosophy, School of Humanities, University of Chinese Academy of Sciences, Beijing, China. [2] Institute of Philosophy, Chinese Academy of Sciences, Beijing, China. [3] Institute of Automation, Chinese Academy of Sciences, Beijing, China. [4] Center for Long-term Artificial Intelligence, Beijing, China. ✉email: yi.zeng@ia.ac.cn

Introduction

Social robots are considered as machines that can interact with humans in a social manner (Fong et al. 2003). By interacting with humans, social robots have been able to evoke human emotions to some extent (Breazeal 2002, 2003). Social robots are widely used for emotional interaction needs, such as the therapeutic robot seal PARO (Robinson et al. 2013). More extensive applications of social robots include robots for medical companionship and service robots for public use. The social robots discussed in this paper are mainly social robots applied among people who need emotional support, such as service robots used for elderly care (Borenstein and Pearson, 2010; Sharkey and Sharkey, 2012) and childcare (Sharkey and Sharkey, 2010). A common feature of these robots is that they all complement the emotional needs of people who are lack of social interaction through intentional anthropomorphization. In this relational sense, human-social robot interactions are generated by a combination of active anthropomorphic commitment from the makers and emotional needs from the users.

Humans' innate ability of anthropomorphization projects emotions to inanimate objects. The anthropomorphization ability manifests itself by assigning human characteristics to inanimate objects and animals and helping humans rationalize the behavior of objects (Duffy, 2003). Gray et al. argue that the core of anthropomorphization is attributing human-like attributes and mental states to non-human agents and objects, in which the core anthropomorphic characteristics include conscious experience, metacognition, and intention (Gray et al. 2007). Epley believes that the core of anthropomorphization should be supplemented with descriptions related to, for example, emotional states, behavioral traits, and human-like forms (Epley et al. 2007). Experiments by Riether et al. suggest that robot members contribute to human task performance (Riether et al. 2012). These social robots share a common characteristic that, regardless of whether they are anthropomorphic or anthropomorphism, they emulate the behavior and features of living entities to meet humans' emotional needs in social interactions. Like humanoid social robots, the animal traits and behaviors of the PARO robot involve a broader concept of anthropomorphization. In this context, the concept of anthropomorphism refers to how we interpret non-human entities through our own understanding, whether by attributing human characteristics to them or other non-human characteristics. Designers of social robots are both intentionally anthropomorphizing social robots for the purpose of satisfying human emotional needs.

However, anthropomorphic social robots face accusations of deception, disappointment, and reverse manipulation, which are considered to lead to a range of moral and emotional risks. Some approaches have tried to address the above risks but have not solved the problem. With this in mind, how exactly should human-social robot interactions be positioned? To answer this question, we explicitly point out that based on the current advancement of technology, there is naturally a virtual interaction environment (VIE) during human-social robot interaction and previously people didn't realize or choose to ignore its existence. We facilitate the realization of VIE indication by discussing the necessary work of social robot producers. In the next section of this paper, we describe the possible deception that anthropomorphic social robots may induce, and the emotional risks that deception entails. Following that section, we will discuss existing solutions to the anthropomorphic risk of social robots and show that all existing solutions face difficulties in addressing these risks. Then we will introduce the central concept of this paper, the "Virtual Interaction Environment", and defend the necessity of a "Virtual Interaction Environment Indication" process. Finally, we will attempt to illustrate the contribution of such virtual interaction environment indication to the reconsideration of ethical guidelines for robot ethics.

Kinds of deception and emotional risks from social robot anthropomorphization

The process by which social robots are perceived as partners, bystanders, and group members is the process by which social robots are personified. The suggestive anthropomorphic language makes it easy for the public to draw analogies between social robots, humans, and natural creatures (Scheutz, 2012). In creating social robots, researchers make promises that social robots can meet emotional needs and design social robots with animal or human like appearances. Even among researchers conducting robotics research, linguistic interpretations that express robot behavior suggest an anthropomorphic tendency of robots, such as the use of adjectives like "with a smile on its face" and "with a sad frown" to describe a robot's appearance (Breazeal 2001, 2002; Proudfoot, 2011). Then, using rationalizations, people try to project such interpretations onto social robots when confronted with their behaviors. In the process of interacting with social robots, users unconsciously increase their expectations on them. This emotional dependence that humans have on robots is a one-way relationship (Scheutz, 2012).

When people interact with social robots, their existing social relationships will also permeate. For example, people are more likely to donate to a museum when the service robot asks in a female voice than in a male voice (Siegel et al. 2009). Authentic social relationships are significant in generating emotions as people interact with social robots. Therefore, people tend to evaluate technological objects in their lives as real people (Reeves and Nass, 1996). The limited autonomy that social robots typically have also affected people's expectations on them (Murphy and Rogers, 2004; Scheutz et al. 2007). The increased expectations lead users to view social robots as real people.

Human users unconsciously perceive social robots involved in their lives as partners rather than tools. The media equation theory leads users to view social robots as real people and to expect more from them (Reeves and Nass, 1996). Studies also show that older adults enjoy and are ready to build relationships with social robots (Pu et al. 2019). However, such expectation creates a gap with the reality of social robot development, where human users expect robots with limited autonomy to have an emotional understanding. The notion that humans may manipulate their perceptions of social robots to satisfy emotional needs aligns with the concept of "suspension of disbelief" (Schaper, 1978), which refers to an audience's willingness to accept the fiction within a narrative, even when they are aware of its fictitious nature. Applying this to our context, individuals might "suspend disbelief" when interacting with social robots, treating these machines as if they possess emotions and sociability. In addition, cognitive dissonance theory suggests that people, when confronted with information that is inconsistent with their personal beliefs, expectations, or values, will change their perceptions in order to minimize this sense of dissonance (Festinger, 1962). Thus, this mental maneuver allows people to derive more emotional value from their interactions, such as companionship, comfort, or entertainment, resulting in the public's preference to imagine that social robots have transcendent abilities and generate consciousness in their dealings with people (Broadbent et al. 2010). Social relationships influence emotional relationships, and social robots promise social illusions without the ability to interact socially. Such social illusions benefit groups in need of emotional companionship, such as elderly people living alone (Broadbent et al. 2010; Robinson et al. 2014) and children with autism (Diehl et al. 2012). Based on competence and trust, human-social robot interactions become similar to human-human interactions; people see social robots as their partners, children, or servants, and just ignore the lack of real comprehension ability and emotional competence of social robots.

Rodogno argues that the interaction between social robots and people is similar to the interactive process of reading a book or watching a movie (Rodogno, 2016). Emotions are insufficient to constitute deception because the sadness disappears after we realize that it is fictional. The generation of real emotions of individuals reflected in social robots implies that such emotional reactions are motivated by imagination, therefore, social robots should not face moral accusations (Rodogno, 2016). However, the emotional satisfaction based on active deception weakens the individual's responsibility to understand the world (Sparrow, 2002). Social robotics researchers have been subjected to unethical torture because they not only support deception but also promote and encourage it in the design process of social robots. The designability of robots means that "active deception" can occur directly in the design and development of social robots. Frequent intimate interactions with social robots are feared to be detrimental to human emotional and social development and lead to attachment problems (Sharkey and Sharkey, 2010). When anthropomorphization, or rationalization of behavior, exceeds certain boundaries, anti-human technological systems may arise (Royakkers and van Est, 2015).

The emotional disappointment after being actively duped can turn people off against the "social illusion" of social robots. When people face emotional crises in virtual social relationships, they may not feel fulfilled in actual social relationships and become further immersed in the "social illusion" created by social robots, which ultimately affects the healthy development of the individual's mind. For example, this can lead to dependence on robots in children who have not yet developed full cognitive abilities, which can affect the development of empathy (Severson and Carlson, 2010). Furthermore, Elster argues that if users adapt their preferences to get along with their robot companions, real preferences will be replaced by unreal ones (Elster, 2016). Coeckelbergh argues that robots as companions can deprive potential human companions of service (Coeckelbergh, 2009). Prolonged single interactions with social robots can affect an individual's cognitive socialization. When the virtual immersion is externally threatened, the rebound emotional response manifests itself in doubting the trust basis of genuine social relationships, thus presenting a kind of illusory dependence. Such an illusory dependence is based on the social illusion given by social robots, and once formed, it becomes challenging to accept genuine social relationships.

None of the existing artificial intelligence agents meet the most basic requirements of meaningful sociality. In this sense, social robotics still has a long way to go. There is no evidence that today's social robots have role-specific emotional representations. The emotional feedback of social robots is also incomparable with that of pets, which are able to perceive human emotions, while current social robots do not have such level of delicate cognition. Based on the current level of intelligence of existing social robots, the emotional expectations of them are destined to be illusions rather than real "family members".

The risk of one-way emotional ties lies in that psychological dependencies can be exploited for the inculcation of specific values and the reverse manipulation of people (Scheutz, 2012). Reverse manipulation refers to using emotional dependencies by social robots to persuade users to make decisions, for example, convincing the user to buy more products from the producer (Scheutz, 2012). Danaher distinguishes between external state deception, surface state deception, and covert state deception. The distinction between surface state deception and covert state deception is interesting, mainly because a covert "understanding but pretending not to understand" is more ethically disturbing than the surface state "not understanding but pretending to understand" (Danaher, 2020a). In the case of social robot

deception, some believe it is necessary, for that deception in the service of a higher purpose is morally justified (Wagner and Arkin, 2011; Shim and Arkin, 2016; Wagner, 2016; Isaac and Bridewell, 2017). Others believe that if we need social robots to perform social functions, we must allow robots to deceive (Wagner and Arkin, 2011; Shim and Arkin, 2016; Wagner, 2016). However, value inculcation is manifested in the fact that designers can embed negative or immoral value judgments into the behavior of social robots. Malicious or unethical value orientations threaten social robots' companionship with children. It may also monitor users, harm them, or instill radical ideas.

Compared with social robots, an intact living pet is not subject to original value inculcation issues. Although one opinion is that pets are also induced by their original birth environment, the induction is not the same as the design of a robot, which includes more of a "puppet show" element in its design process. As a product, a robot as a puppet operates more with an unchangeable program, while a living pet has its own natural nature, which ensures that it retains more possibilities and cannot be easily manipulated by some intentional "coding". The deviation caused by social robots is more generally manifested in the difficulty of translating the emotional expectations of their human users into reality.

On the one hand, considering the silence towards anthropomorphization in public space, we need to think about how to make the public aware of the existence of anthropomorphic non-human agents. On the other hand, we need to fully understand the emotional deception of social robots and avoid "active deception". People should be aware of the risks of active deception, reposition the relationship between humans and social robots, and consider how to deal with the differences between interpersonal interactions and human-social robot interactions.

An analysis to existing approaches addressing the risks of social robot anthropomorphization

Several approaches claim to address the risks faced by humans in their relationships with social robots. We will show that each of the pathways faces some problems.

Jackson et al. propose to give the robot the ability to reject inappropriate commands so as to reduce the possibility of unreasonable user expectations from the beginning. Through rejection, the user understands that the robot possesses special autonomy and therefore dissipates their own excessive expectations. Ultimately, the robot is removed from moral accountability in the face of anthropomorphic disappointment (Jackson et al. 2021). However, rejection in a moral sense confronts a more profound problem of anthropomorphism, which Epley found to lead to a general increase in anthropomorphism when non-human agents violate human expectations (Epley et al. 2007). Consider a scenario where a person is using a washing machine and the machine stops because of a malfunction, but the clothes inside are not yet finished. In response to the "malfunction" of the washing machine, one would think that "the machine is consciously working against me", and this understanding would lead to an increase in anthropomorphism when human expectations are violated.

Moreover, the rejection approach does not solve the problem of deception but creates more deception. A robot's ability to reject or violate a task can lead to more expectations of anthropomorphic and emotional capabilities from users and thus make the development of social robots problematic. The rejection response will also circumvent the real purpose people have in interacting with social robots, which is to expect them to provide service. Responding to the anthropomorphic risks of social robots with rejection is an escape from the problem and does not really

address the risks faced. Furthermore, the social robot's rejection is not a rejection in the usual sense, but rather more of an internal programming result that indicates a malfunction in many cases. If one's perception of a social robot's rejection is misplaced, this would suggest that the effect of introducing rejection for a social robot will not make the interaction more explicit, but rather introduce confusion between the real internal structures of the social robot and the anthropomorphic ones. Therefore, the rejection approach gives neither a measure that facilitates the tackling of risks, nor does it address the problem of relationship repositioning for human and social robots, but simply ignores the problem itself by adding more behavioral descriptions.

The second approach handles the anthropomorphic risks through warning statements and built-in distancing effects (Verfremdungseffekt), where Bendel proposes to indicate the unsuitability of robots to participate in social life by reminding users of their limited autonomy (Bendel, 2019). Weber-Guskar's rebuttal to this view starts from the failure of the deception explanation, arguing that the anthropomorphization of social robots is an imaginative perception instead of a deception. Therefore, warning statements and built-in distancing effects are unnecessary (Weber-Guskar, 2021). Moreover, the warning approach does not effectively guarantee that people will no longer have anthropomorphic expectations or one-way emotional attachments to social robots. Warning statements and built-in distancing effects confront risk through morally overly rigid means, in which the meaning of social robots is limited when people use them only to obtain responses that have no relevance to the people themselves. The descriptions of the robots' capabilities required in some relevant systems face an accusation of such one-way self-disclosure. An interpretable and ethical social robot does not simply rest on mere self-disclosure by the robot but requires a repositioning beyond the meaning of human-robot interaction. The explication of such a repositioning cannot be done through the unilateral disclosure of people or robots but requires a state explication of the interaction between people and social robots. Based on the current advancement of related technology, the relationship between humans and social robots is very different from the relationship between humans and pets, or the relationship between humans. The difference is directly reflected in the meaning of the virtual environment that exists in the interaction between social robots and humans, which we will discuss in the next section.

The third approach is the matching hypothesis, where Goetz et al. propose that instead of intentional anthropomorphic appearance design, user trust is gained by achieving a matching between robot appearance and tasks (Goetz et al. 2003). There are three problems with this solution. First, the matching hypothesis only sees functional mapping but ignores the real emotional needs of people. Achieving a match between a robot's appearance and its task only completes the functional part but fails to respond to users' emotional needs on social robots. The presence of such emotional needs may make the producer more willing to convince users, through some kind of implication, that there is a person hiding inside their robot. As a result, this approach ultimately fails to properly reposition the relationship between humans and social robots. Second, the matching of robot appearance and tasks does not account for the virtual nature of human-social robot interaction, which refers to the cognitive bias in people's emotions and understanding of social robots compared with the social robot's reality. People give social robots anthropomorphic expectations. The expectations in the sense of cognitive bias have not been truly fulfilled, but social robots are given anthropomorphic appearances in a form that fits such expectations, resulting in a failure to recognize the real problem of the virtual nature of the relationship between humans and

social robots. Third, appearance-task matching does not have adequate explanatory validity to account for the human-social robot relationship. This explanatory validity focuses on matching appearance to social robot function and merely attributes the explanatory responsibility to the implementation of social robot function without providing an explanation of the human-social robot relationship. Since no relational account is given, the matching hypothesis will not achieve the purpose of effective explanation to the users.

In addition to the aforementioned approaches to address anthropomorphism risks, Proudfoot proposes to make social robots make mistakes to improve human recognition of social robots, including reducing human expectations by using "deliberate misspellings", "lack of common-sense knowledge" or "unreasonable conversations" (Proudfoot, 2011). The major problem with the error-making approach is that users interact with social robots for the purpose of obtaining a complete and high-quality service, but social robots that are prone to make mistakes fail to do so. The initial service goal is sacrificed in order to improve the recognition of social robots, turning the deployment of social robots into an "either/or" problem, which is not in line with our original vision. What's worse, the error-making approach will also lead to disappointment in the performance of social robots, which is eventually detrimental to the development of social robots.

Virtual Interactive Environment Indication

An intentional account of the virtual interactive environment between humans and social robots can ameliorate problems associated with social robot anthropomorphism and yield a diversity of social robot image-building within a broader range of scenarios. In this section, we attempt to defend the cognitive repositioning of the virtual interactive environment for human-social robot interaction and, on this basis, consider the practical applications of virtual interactive environment indication in the design and development of social robots.

The virtual environment of social robots is based on virtual interactions, which are mainly manifested as virtual emotional relationships shaped by human-social robot interactions. Sweeney interprets the relationship between humans and social robots as a "fictional" emotional relationship, arguing that the relationship between humans and social robots is a relationship to an object with a fictional overlay (Sweeney, 2021). Contrarily, interaction in a virtual environment, as we have argued, is the idea that from the moment a person interacts with a social robot, the person is already involved in a virtual environment, in which the person acts as the protagonist and is engaged in an interplay with other characters, i.e., the social robot. Therefore, the emotional relationship between people and social robots should be interpreted as an interactive relationship within that virtual environment.

When we engage with a compelling narrative, we may feel "transported" into the world of the story, becoming fully immersed in the situation and developing deep emotional resonance and identification with the characters. This is captured by narrative transportation theory (Green and Brock, 2000). A similar "transport" can occur during interactions with social robots, drawing us into a virtual interaction environment. This virtual interaction environment is shaped by our psychological engagement and emotional responses. Our understanding and emotional involvement in the narrative situation can create a vivid and realistic virtual interaction environment (Radford and Weston, 1975). However, this environment is not static but evolves with the intensity and duration of our emotional involvement. It's worth noting that the virtual interaction environment does not exist in isolation from the real world.

The virtual interactive environment is achieved through transport borne by the sense of immersion. By adhering to the principle of interactivity, users are transported into a virtual environment where vividness and authenticity are entirely dependent on their sense of immersion. This sense of immersion carries the so-called "transportation" process, which is an emotional flow and transformation between the virtual and the real. During the interaction process, the emotional engagement and transference of individuals shape and enhance the immersive effect of the virtual environment, and the human imagination and emotional involvement play a pivotal role in constructing and maintaining such a virtual environment. Each virtual interactive environment has its peculiarities. They have a normative commonality, mainly reflected in the deep emotional involvement and the construction of imagination. This is like when we read novels or listen to stories. Although each story has its unique plot and characters, reading or listening to stories requires our emotional involvement and imagination. It stimulates our emotional involvement and imagination by referring to real scenarios, producing a profound experience, and gaining emotional comfort and satisfaction through interactions with social robots. Creating a virtual environment through the principle of interactivity can be understood as a continuous, dynamic interactive process where human-machine interaction shapes and continuously reshapes the environment.

Three characteristics are exhibited in the interaction between humans and social robots within the virtual environment. The first characteristic is manifested in that the interaction between humans and social robots in the virtual environment is an interactive relationship that is positioned from a human perspective. By treating social robots as objects that can participate in virtual interactions, a unique virtual interaction paradigm is developed between humans and social robots.

The second characteristic is reflected in the distance between the virtual nature of human-social robot interaction and the instrumental nature of the social robot itself. In reality, a machine without intentional constructs is considered a tool, whereas in the interaction with a social robot, a virtual environment is created between humans and social robots. Social robots commit to virtual interactions out of a need to satisfy human emotional companionship, which determines that the users involved in the interaction will get unconsciously involved in the environment shaped by this virtual interaction. Such a virtual interaction differs from reading a novel in that, while people are emotionally touched by the experiences of fictional characters in a novel, they remain conscious of the fictional nature of the novel when reading it, but they are usually not aware of such virtuality when interacting with social robots. As discussed in Section "Kinds of deception and emotional risks from social robot anthropomorphization", people are thus more likely to experience emotional deception in their interactions with social robots, which makes it more necessary to help people truly understand human-social robot interactions as virtual interactions.

The third characteristic is reflected in the stage-specific nature of people's understanding of the virtual interaction between humans and social robots, which consists of three progressive stages. The first stage is at the stage of interaction in which people try to believe that social robots inherently have an understanding of relationships through their perception of social robots' external behaviors, or in Danaher's words, at the stage of ethical behaviorism (Danaher, 2020b). In this stage, people willingly ignore the truth that social robots are not emotionally capable, yet in fact, social robots as tools are unable to engage in deep human interactions as expected. In the second stage, when people are more engaged in interacting with social robots, they start to invest more emotions and resources in them. This will lead producers to recognize the influence of social

robots on people's actual consumption behavior, which in turn will give them a greater incentive to further intentionally shape social robots to be more anthropomorphic in virtual interactive environment, resulting eventually in the prevalence of emotional deception. When people interact with social robots, a feedback loop is formed between humans and social robots, in which people's imagination that social robots do have generated realistic feedbacks dominates their interactions. The significance of helping people gain a "Virtual Interaction Environment" perspective on human-social robot interactions is to shift their cognition of the interactions from an "active deception" state in the second stage to a "lucid dream" state in the third stage. In this third stage, the understanding of the virtual interaction environment facilitates a lucid awareness of the existence of virtual interaction between humans and social robots, reaching to a level similar to that of the cognitive awareness when reading a novel. Just as although there may be emotional highs and lows, one does not perceive that one is actually living in the events of the novel when reading it (Sweeney, 2021), the "lucid dream" metaphor illustrates that one has a lucid understanding of the state and possible impact of the ongoing virtual interaction between oneself and the social robot and does not feel of being intentionally deceived by the social robot or its producer. The understanding of the internal mechanisms of human-social robot interaction thus becomes clearer and a more complete understanding of the role of social robots is achieved. Therefore, it is crucial to emphasize the virtual nature of the interactions when introducing the service functions of social robots.

We argue that adding elucidations and indications for virtual interaction environment is a realistic implementation solution to promote people's understanding of virtual interaction environment. Virtual Interactive Environment Indication (VIEI), by definition, is the process in which the virtual nature of the human-social robot interaction is clearly identified and declared during the deployment and application of social robot products. Through the Virtual Interactive Environment Indication process, human users should be well-informed and aware that they are participating in a virtual interaction with social robots, and thus should pay due attention to the potential risks of deception from anthropomorphization. The intentional construction of an image of a social robot in a virtual interaction environment is a cautionary tale for our dealing with social robots. The non-anthropomorphic virtual environment cognition of social robots drives the conceptual interpretation of the relationships between humans and social robots. It illustrates the risks involved in promoting the social adaptation of social robots.

The advantage of applying Virtual Interactive Environment Indication is that by explaining the existence of virtual interactions, users do not perceive the social robot as a living "person" but rather take a more moderate stance in understanding. Central to avoiding active deception from social robots is that people are aware that their interactions with social robots are taking place in a virtual interactive environment, thereby reducing their own expectations that should not be misplaced on social robots. On the other hand, prior descriptions of the interactive nature of social robots will also circumvent the emotional disappointment that people might feel about social robots. In addition to eliminating deception and avoiding disappointment, the malicious exploitation of social robots can also be regulated by the intervention of Virtual Interactive Environment Indication, which requires that social robots be used "only in the sense of a companionable virtual interactive environment". Thus, the applying of a Virtual Interactive Environment Indication process helps to promote a more normative development of social robots.

Moreover, in contrast to digital environment, the perception of virtual environment of human-social robot interactions is an

explanatory concept for the description of the real situation. The digital environment is often portrayed as a result of the simplification and digitization of the physical world, based entirely on digital features and representations of data. However, when it comes to virtual environment perception for social robots, the process is quite different. The virtual environment perception of social robots is not based on pure representation, but interprets real-life situations through the principle of interactivity. The interaction between humans and social robots is not limited to a one-way or predetermined reaction-based process; rather, it is a feedback-dependent, mutually adaptive, and continually evolving process. The principle of interactivity underscores this ongoing, bidirectional exchange of information and feedback. Applying this principle of interactivity allows us to grasp the dynamics of interactions within virtual environment between humans and social robots. Within such environment, social robots transcend their roles as mere tools executing predetermined actions; instead, they actively engage as participants in an evolving interaction with humans. In this manner, it becomes evident that human-robot interaction is not confined to a static, predefined virtual environment. Rather, it unfolds within a dynamic, adaptable virtual interactive environment shaped by emotion and imagination. This virtual interactive environment, rooted in the principle of interactivity, encompasses not only the interactions between humans and social robots but also encompasses the formation and evolution of emotional relationships between them. Thus, the Virtual Interactive Environment Indication process will provide us with a new perspective on the relationship between humans and social robots, potentially enriching our understanding of reality and helping us better understand what potential benefits and challenges social robots may bring in the future.

In the previous section, we have analyzed the current approaches in addressing the anthropomorphization risks of social robots and elaborated on their limitations. We propose that the introduction of Virtual Interactive Environment Indication effectively tackles these issues. Firstly, Jackson et al.'s suggestion that social robots reject inappropriate commands to mitigate anthropomorphization risks (Jackson et al. 2021) does not directly eliminate the risk and might even amplify it when robot responses contradict human expectations. In contrast, our proposed solution of Virtual Interactive Environment Indication proactively establishes the virtual nature of the interaction environment, guiding users to understand the distinction and avoid excessive anthropomorphic expectations. Secondly, Bendel's strategies involving warning statements and built-in distancing effects (Bendel, 2019) do not sufficiently prevent the formation of one-way emotional dependencies on social robots, stemming from misconceptions about interaction. Through Virtual Interactive Environment Indication, we redefine the human-robot interaction paradigm, offering an environment that enables users to grasp the significance of social robots beyond mere emotional projection. Regarding the third strategy, Goetz et al.'s matching hypothesis, which focuses on aligning robot appearance and tasks, lacks consideration for users' emotional needs (Goetz et al. 2003). The introduction of Virtual Interactive Environment Indication allows users to realize their participation in an immersive narrative, reducing one-way emotional dependencies and deceptions. Lastly, Proudfoot's suggestion of social robots intentionally making mistakes to address issues (Proudfoot, 2011) undermines the high-quality user experience, potentially leading to disappointment towards social robots in general. In contrast, the introduction of Virtual Interactive Environment Indication clarifies the true state of social robots during interactions, addressing the disappointment issue at its core. The manufacturers will illuminate the relationship between people and

social robots, eliminating deception and disappointment issues inherent in the aforementioned methods, thus preserving the high-quality user experience of social robots.

Virtual interactive environment indication promotes the reconsideration of ethical guidelines for social robots. A cognitive repositioning of the image of social robots can help us re-examine the existing robot ethics guidelines. Instead of simply considering in terms of providing companionship, an ethical guideline for social robotics should clearly define the responsibility of producers in terms of the "Virtual Interactive Environment Indication" requirement.

First and foremost, by introducing the requirement, the responsibility of social robot producers can be reconsidered and initiatives to address the risk of social robot deception can be developed based on the redefined responsibility. Virtual Interactive Environment Indication is instrumental in protecting the emotional rights of various vulnerable groups, such as children, by clarifying in what sense the developers are held responsible. As argued in "Why Robots Should Not Be Treated Like Animals", the difference between social robots and animals is that the animal and animal trainer relationship is purer (Johnson and Verdicchio, 2018), whereas the relationship between robot and "robot trainer" also contains an artificial component, which makes the responsibility of social robots more the responsibility of the manufacturer. Thus, the distinction between the responsibility of social robots and animals arises in this sense, suggesting a third-party subject of responsibility when considering the responsibility of social robots, i.e., the producer of the social robot. The relationship between users, social robots, and their manufacturers is different from the relationship between humans and pets, and the hidden "unnaturalness" of this relationship makes it impossible to draw a single division of responsibility.

In economic and organizational ethics, responsibility is often seen as a key concept. As Friedman and Miles argued in "Stakeholders: Theory and Practice", the responsibility of an organization's behavior is not merely to meet legal requirements, but its impact on society (Friedman and Miles, 2006). In other words, manufacturers of social robots should be accountable for the impact of their products on users, not only because it is legally required but also because their products may have substantive effects on users' lives (e.g., potential emotional harm). Therefore, manufacturers should explicitly perform VIEI-related work. As Donaldson and Dunfee pointed out in their research on economic ethics, corporate social responsibility is multilevel, encompassing responsibility towards stakeholders, society, and the global community (Donaldson and Dunfee, 1999). When considering emotional risks, such as deception, disappointment, and reverse manipulation that social robots might bring, the responsibility of manufacturers should extend to users (i.e., stakeholders), society, and the global community. Moore believes that responsibility should be allocated based on the actions of various parties that influence the outcome (Moore, 1999). In our context, this means that if the automation of social robots leads to a certain consequence, manufacturers should be held responsible; however, if this consequence is an allegation of emotional deception that arises after the manufacturer has explicitly stated the existence of the virtual interactive environment within the human-social robot interaction, then to some extent the manufacturer has fulfilled its duty to inform, which means that the user also needs to bear some responsibility. Thus, we need to reallocate responsibility between manufacturers and users to reflect their respective impacts on the outcome.

Traditional social robots with limited autonomy face the problem of unlimited liability for the manufacturer. When we

view a social robot as a virtual external environment, it becomes possible to divide the liability of the social robot between that arising from automation and that arising from the virtual interaction environment. When facing a liability dilemma, if the consequences are caused by the automation of the social robot, it is necessary to pursue liability in the automation cause of liability, i.e., to deal with the manufacturer of the social robot. If the consequences are caused by the inappropriate perception of the virtual interactive environment between the human and the social robot, then the responsibility can be pursued in a sense including the interacting parties. This understanding of liability helps to move away from a mere "unlimited automation responsibility" and start from the categorization of environmental interactions to seek a more precise attribution of responsibility for the negative consequences of social robots.

Second, the introduction of "Virtual Interactive Environment Indication" illustrates a pathway to circumvent disappointment in social robots. The current social robots are not yet truly intelligent, but more of a seemingly intelligent tool. Instrumental intelligence cannot meet people's expectations for real social interaction and will affect how social robots are perceived in society. The emergence of disappointment can significantly affect the expected social robot paradigm as well as people's expectations of the whole robot industry. The more human disappointment accumulates, the more it will affect the adoption of social robots. The need to re-establish new images of social robots that are different from the instrumental ones arises in this background. To avoid disappointment in the anthropomorphization of social robots, diverse social robot images should be constructed during the development of social robot technology. Considering the interaction between social robots and people as Virtual Environment Interaction implies a reconstructing of current image of social robots and a repositioning of the human-social robot interaction, which is a first step of such efforts to open up a much broader technological future for social robotics. For instance, consider a scenario where a social robot serves as a lifestyle assistant for the elderly, tasked with reminding medication, prompting physical activity, and facilitating simple social interactions. Traditional methodologies in robot ethics might design this robot to imitate human behavior as closely as possible to cater to the elderly's social requirements. This approach, however, could lead to unrealistic expectations from the robot, such as the anticipation that the robot will comprehend and respond to all their emotional needs, which is currently beyond the reach of robotic technology. Such misaligned expectations could result in feelings of disappointment, or even deception, towards the social robot. At this point, bluntly reminding the elderly that the robot lacks the ability to respond emotionally seems harsh, and allowing the robot to make errors fails to meet their desire for a beneficial user experience. However, if we conduct VIEI process at the beginning of the interaction, clearly stating to the elderly that this social robot is a "reading-enabled" machine available for their use, but simultaneously providing narrative-like virtual interactive characteristics for explanations, their expectations will adjust accordingly. They will understand the functional limitations of the social robot, realizing it is more akin to a "character in a story" rather than a tool or a real human. This way, they would be less likely to develop unrealistic expectations towards the social robot, thus reducing potential feelings of disappointment and deception.

Third, the realization of "Virtual Interactive Environment Indication" facilitates a more rational understanding of humans towards social robots. Experiments such as "Uncanny Valley" have shown that the understanding of social robots in society significantly influences individuals' emotions toward social robots, whether pity or panic, due to anthropomorphic imagery

(Mori, 1970). In this sense, a "Virtual Interactive Environment" cognition can avoid misunderstandings towards social robots. People's emotions towards social robots that originate from the fear of the unknown can also have a shift in the sense of such "Virtual Interactive Environment" cognition. Such cognitive transformations will make social robot images more diverse in people's imagination, thus also giving freedom to the technical imagination of social robots in the interactive sense.

Fourth, the implementation of "Virtual Interactive Environment Indication" will promote people's understanding of the internal mechanisms of social robots. Schneidernnan argues that anthropomorphic social robot design leads to unpredictability and ambiguity about social robots (Schneidernnan, 1988). The internal mechanisms of social robot systems, as a Blackbox, are not intuitively available to the public, making it impossible for users to know the internal processes by which social robots provide services. The lack of knowledge of the internal mechanisms of social robots and the anthropomorphic design aggravate the misconceptions about social robots. The "Virtual Interactive Environment" cognition of social robots takes into account the open iterations of the robot's internal mechanisms, and is consistent with the "interpretability" requirement of robot ethics guidelines. Thus, it promises a complementary provision that the lack of human understanding of the internal mechanisms of social robots can be changed in the context of specific human-social robot interactions. Additionally, the allegations of deception discussed earlier can be mitigated and resolved through VIEI. With an understanding of the VIE, users can accurately situate and recognize their relationship with social robots as a distinct interaction from the outset. Consequently, when utilizing a social robot, accusations of deceptive behavior are precluded. This promotes the virtual interactive environment, sparing users from grappling with a momentary negative perception of interacting with a social robot and instead establishing a proactive pre-engagement agreement. For instance, when an elderly individual interacts with a social robot, the instructions regarding the virtual interactive environment enable them to engage with the social robot without being troubled by potential deception and disappointment. Instead, they approach the interaction with explicit awareness that they are engaging in a "special reading" mindset. As a result, the efficacy of VIEI exceeds that of straightforward "artificial" indications due to the deeper level of comprehension it affords. VIEI accentuates the contextual and virtual aspects of human-social robot interactions. Rather than bluntly informing users that the robot is manufactured, it proves more constructive to enable them to comprehend their involvement within a virtual interaction environment through the retrospect of the emotional responses evoked during novel reading. This comprehension aids users in distinctly recognizing that their interactive counterpart is artificially constructed, thereby mitigating emotional risks stemming from discrepancies between expectations and reality.

Fifth, a new image for social robots constructed by "Virtual Interactive Environment Indication" can contribute to their better adaptation to social mechanisms. The elucidation of virtual interactive environment promotes thinking about the ethical guidelines of social robots. The Bottom-up approach for modeling human moral faculties into machines (Wallach and Allen, 2008) and the idea of Virtual Interactive Environment is internally consistent and will help social robots learn ethical models of society. Social robots acquire experience that aids them in adapting to social mechanisms. Social robots relying solely on external appearances come across as disconnected entities. However, with the integration of VIEI, these robots gain an extended dimension to their representations. Specifically, social robots are perceived as engaging with users much like reading a

novel. This interaction symbolically casts the social robot as an alternate reader, participating with users within a narrative virtual interactive environment. This significantly reduces the perceived unfamiliarity of these robots, promoting their acceptance. This newfound familiarity, akin to the immersion experienced while reading novels, transforms the perception of virtual social robots from something alien to something relatable, drawing on individuals' past reading experiences. Over time, social robots evolve into familiar entities for people, presenting a novel image of social robots. Grounded in this new perception, the cognitive load people experience when interacting with social robots lessens, contributing to a more positive societal response while delving deeper into the ethical considerations associated with these robots.

VIEI represents an interaction-based approach to comprehend and interpret the relationships between social robots and humans. It diverges from conventional methods of robot ethics, emphasizing the process of interaction rather than focusing only on the internal logic and decision rules of robots. Similar to the emotional states evoked by reading a novel, VIEI refers to an individual's ability to create a virtual emotional experience by projecting emotions onto a social robot, drawing on the imaginative power of storytelling. Applying this concept to interactions with social robots could enhance the understanding and acceptance of the human-robot relationship. Furthermore, manufacturers carry the responsibility of mitigating potential emotional risks by considering their product designs and usage. This responsibility-driven approach could prove more effective than existing robot ethics methods, as it takes into account not only robot behavior but also people's interpretations and comprehension of these behaviors. The identity of social robots is understood as a source of inclusive values. As with Rolston's understanding of environmental value, the value of nature is derived from nature itself, and the cognition of the Virtual Interactive Environment can similarly enable social robots to derive their value from themselves (Rolston, 1988). Considering the interaction between social robots and humans as a virtual environment interaction is conducive to advancing further research on the issue of moral competence of social robots. We can advance the interactive knowledge of social robots' ethical practices through the virtual environment interactions between individuals and social robots, in which the shift in cognition contributes to our broader portrayal of robot ethics.

Sixth, the Virtual Interactive Environment cognition of social robots helps us to circumvent the role crisis of considering social robots as specific roles. The one-way attachment to social robots tends to view social robots as having specific roles in social relationships, and such cognition of social robots as socially embedded "individuals" and the resulting accusations of deception are also associated with role-specific attachment to social robots, which distances the social robots from the norms of practice. For example, a young girl using a social robot may perceive the robot as her best friend, and the fixed feedback of the robot may cause the girl to have the illusion of understanding the robot, which may lead to accusations of disappointment and deception. The significance of VIEI is that the little girl explicitly recognizes at the outset that the "illusion" of social robot interaction is due to the fact that she has entered a virtual interactive environment and is participating in a particular narrative, which means that the little girl no longer sees the social robot as her best friend, but as a particular character in a storybook. Thus, by shifting the cognition of social robots to a Virtual Interactive Environment, the appeal to role-specific social robots is circumvented, and the nature of human-social robot interaction, i.e., interactions taking place in a virtual scenario, is appropriately understood.

Conclusion

With a discussion of the emotional risks associated with social robot anthropomorphization, and a detailed analysis of the limitations of existing coping methods, the concept of Virtual Interactive Environment proposed in this paper is an interpretation of human-social robot interaction in the specific social robotics context. The identification and elucidation of Virtual Interactive Environment by social robot producers will help promote more clear attribution of responsibility, increased interpretability, reduced public disappointment with social robotics technology, reshaping the cognition of social robot images, and exploring the construction towards better robot ethics guidelines. As a response to the existence of Virtual Interactive Environment, Virtual Interactive Environment Indication (VIEI) is an interpretive clarification to quell allegations of active deception from social robots. Social robot producers are liable to go through an explicit Virtual Interactive Environment Indication process during the production design and deployment of social robots.

The use of the Virtual Interactive Environment concept and its specific applications need to be further explained in future work to seek better ways to help people understand the interaction processes that are taking place. While interpretive indication of Virtual Interactive Environment is thought to facilitate and contribute to the understanding of the virtual interactions that are occurring between humans and social robots, how to specify and regulate such an indication process also needs to be further explored in future work.

References

Bendel O (2019) Sexroboter aus Sicht der Maschinenethik. Handbuch Maschinenethik 335–353

Borenstein J, Pearson Y (2010) Robot caregivers: harbingers of expanded freedom for all? Ethics Inf Technol 12:277–288. https://doi.org/10.1007/s10676-010-9236-4

Breazeal C (2002) Designing sociable robots (Intelligent robotics and autonomous agents). MIT Press, Cambridge

Breazeal C (2003) Toward sociable robots. Robot Autonom Syst 42:167–175. https://doi.org/10.1016/S0921-8890(02)00373-1

Breazeal C (2001) Affective interaction between humans and robots. In: European conference on artificial life. Springer, pp 582–591

Broadbent E, Kuo IH, Lee YI et al. (2010) Attitudes and reactions to a healthcare robot. Telemed e-Health 16:608–613. https://doi.org/10.1089/tmj.2009.0171

Coeckelbergh M (2009) Personal robots, appearance, and human good: a methodological reflection on roboethics. Int J Soc Robot 1:217–221. https://doi.org/10.1007/s12369-009-0026-2

Danaher J (2020a) Robot Betrayal: a guide to the ethics of robotic deception. Eth Inform Technol 22:117–128. https://doi.org/10.1007/s10676-019-09520-3

Danaher J (2020b) Welcoming robots into the moral circle: a defence of ethical behaviourism. Sci Eng Eth 26:2023–2049. https://doi.org/10.1007/s11948-019-00119-x

Diehl JJ, Schmitt LM, Villano M, Crowell CR (2012) The clinical use of robots for individuals with autism spectrum disorders: A critical review. Res Autism Spectr Disord 6:249–262. https://doi.org/10.1016/j.rasd.2011.05.006

Donaldson T, Dunfee TW (1999) When ethics travel: The promise and peril of global business ethics. Calif Manag Rev 41:45–63

Duffy BR (2003) Anthropomorphism and the social robot. Robot Auton Syst 42:177–190. https://doi.org/10.1016/S0921-8890(02)00374-3

Elster J (2016) Sour grapes: studies in the subversion of rationality. Cambridge university press, New York (N.Y.)

Epley N, Waytz A, Cacioppo JT (2007) On seeing human: a three-factor theory of anthropomorphism. Psychol Rev 114:864. https://doi.org/10.1037/0033-295X.114.4.864

Festinger L (1962) A theory of cognitive dissonance, vol. 2. redwood

Fong T, Nourbakhsh I, Dautenhahn K (2003) A survey of socially interactive robots. Robot Auton Syst 42:143–166

Friedman AL, Miles S (2006) Stakeholders: Theory and practice. OUP oxford

Goetz J, Kiesler S, Powers A (2003) Matching robot appearance and behavior to tasks to improve human-robot cooperation. In: The 12th IEEE International Workshop on Robot and Human Interactive Communication, 2003. Proceedings. ROMAN 2003. Ieee, pp 55–60

Gray HM, Gray K, Wegner DM (2007) Dimensions of mind perception. Science 315:619–619. https://doi.org/10.1126/science.1134475

Green MC, Brock TC (2000) The role of transportation in the persuasiveness of public narratives. J Personal Soc Psychol 79:701. https://doi.org/10.1037/0022-3514.79.5.701

Isaac A, Bridewell W (2017) White Lies on Silver Tongues: Why Robots Need to Deceive (and How), chap. 11

Jackson RB, Li S, Banisetty SB, et al. (2021) An integrated approach to context-sensitive moral cognition in robot cognitive architectures. In: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, pp 1911–1918

Johnson DG, Verdicchio M (2018) Why robots should not be treated like animals. Eth Inform Technol 20:291–301. https://doi.org/10.1007/s10676-018-9481-5

Moore G (1999) Corporate moral agency: Review and implications. J Bus Eth 21:329–343. https://doi.org/10.1023/A:1006020214228

Mori M (1970) Bukimi no tani [The uncanny valley]. Energy 7:33

Murphy RR, Rogers E (2004) Introduction to the special issue on human–robot interaction. IEEE Transac Syst Man Cybern-Part C 34:101

Proudfoot D (2011) Anthropomorphism and AI: Turing's much misunderstood imitation game. Artif Intell 175:950–957. https://doi.org/10.1016/j.artint.2011.01.006

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. Gerontologist 59:e37–e51. https://doi.org/10.1093/geront/gny046

Radford C, Weston M (1975) How can we be moved by the fate of Anna Karenina? Proc Aristot Soc Supple Vol 49:67–93. https://doi.org/10.1093/aristoteliansupp/49.1.67

Reeves B, Nass C (1996) The media equation: How people treat computers, television, and new media like real people. Cambridge, UK 10

Riether N, Hegel F, Wrede B, Horstmann G (2012) Social facilitation with social robots? In: Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction. pp 41–48

Robinson H, MacDonald B, Broadbent E (2014) The role of healthcare robots for older people at home: A review. Int J Soc Robot 6:575–591. https://doi.org/10.1007/s12369-014-0242-2

Robinson H, MacDonald B, Kerse N, Broadbent E (2013) The psychosocial effects of a companion robot: a randomized controlled trial. J Am Med Dir Assoc 14:661–667. https://doi.org/10.1016/j.jamda.2013.02.007

Rodogno R (2016) Social robots, fiction, and sentimentality. Eth Inform Technol 18:257–268. https://doi.org/10.1007/s10676-015-9371-z

Rolston H (1988) Environmental Ethics. Temple University Press

Royakkers L, van Est R (2015) A literature review on new robotics: automation from love to war. Int J Soc Robot 7:549–570. https://doi.org/10.1007/s12369-015-0295-x

Schaper E (1978) Fiction and the Suspension of Disbelief. Br J Aesthet 18:31–44. https://doi.org/10.1093/bjaesthetics/18.1.31

Scheutz M (2012) The Inherent Dangers of Unidirectional Emotional Bonds Between Humans and Social Robots. In: Robot ethics: The ethical and social implications of robotics. MIT Press, Cambridge, Mass, pp 205–221

Scheutz M, Schermerhorn P, Kramer J, Anderson D (2007) First steps toward natural human-like HRI. Auton Robot 22:411–423. https://doi.org/10.1007/s10514-006-9018-3

Schneidernnan B (1988) A nonanthropomorphic style guide: overcoming the humpty dumpty syndrome. Comput Teach 8:9–10

Severson RL, Carlson SM (2010) Behaving as or behaving as if? Children's conceptions of personified robots and the emergence of a new ontological category. Neural Netw 23:1099–1103. https://doi.org/10.1016/j.neunet.2010.08.014

Sharkey A, Sharkey N (2012) Granny and the robots: ethical issues in robot care for the elderly. Eth Inform Technol 14:27–40. https://doi.org/10.1007/s10676-010-9234-6

Sharkey N, Sharkey A (2010) The crying shame of robot nannies: An ethical appraisal. IS 11:161–190. https://doi.org/10.1075/is.11.2.01sha

Shim J, Arkin RC (2016) Other-oriented Robot Deception: How can a robot's deceptive feedback help humans in HRI? In: Social Robotics: 8th International Conference, ICSR 2016, Kansas City, MO, USA, November 1-3, 2016 Proceedings 8. Springer, pp 222–232

Siegel M, Breazeal C, Norton MI (2009) Persuasive robotics: The influence of robot gender on human behavior. In: 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems. IEEE, pp 2563–2568

Sparrow R (2002) The march of the robot dogs. Eth Inform Technol 4:305–318. https://doi.org/10.1023/A:1021386708994

Sweeney P (2021) A fictional dualism model of social robots. Eth Inform Technol 23:465–472. https://doi.org/10.1007/s10676-021-09589-9

Wagner AR (2016) Lies and deception: Robots that use falsehood as a social strategy. Robots that talk and listen: Technology and social impact 203–225

Wagner AR, Arkin RC (2011) Acting deceptively: Providing robots with the capacity for deception. Int J Soc Robot 3:5–26. https://doi.org/10.1007/s12369-010-0073-8

Wallach W, Allen C (2008) Moral machines: Teaching robots right from wrong. Oxford University Press

Weber-Guskar E (2021) How to feel about emotionalized artificial intelligence? When robot pets, holograms, and chatbots become affective partners. Eth Inform Technol 23:601–610. https://doi.org/10.1007/s10676-021-09598-8

Author contributions

AB conceptualized the study, drafted the initial manuscript, and revised the manuscript. YZ conceptualized the study, provided critical feedback, and helped shape the research, analysis, and manuscript writing and revision. EL contributed to the conception of the work, writing and refinement of the manuscript and helped shape the analysis. All authors have read and agreed to the published version of the manuscript. Correspondence to YZ.

Competing interests

The authors declare no competing interests.

Ethical approval

This article does not contain any studies with human participants performed by any of the authors.

Informed consent

This article does not contain any studies with human participants performed by any of the authors.

Additional information

Correspondence and requests for materials should be addressed to Yi Zeng.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.