AI & SOCIETY
https://doi.org/10.1007/s00146-025-02274-1

MAIN PAPER

Trusting the (un)trustworthy? A new conceptual approach to the ethics of social care robots

Joan Llorca Albareda[1] · Belén Liedo[2] · María Victoria Martínez-López[1,3]

Received: 18 June 2024 / Accepted: 19 February 2025

Abstract
Social care robots (SCR) have come to the forefront of the ethical debate. While the possibility of robots helping us tackle the global care crisis is promising for some, others have raised concerns about the adequacy of AI-driven technologies for the ethically complex world of care. The robots do not seem able to provide the comprehensive care many people demand and deserve, at least they do not seem able to engage in humane, emotion-laden and significant care relationships. In this article, we will propose to focus the debate on a particularly relevant aspect of care: trust. We will argue that, to answer the question of whether SCR are ethically acceptable, we must first address another question, namely, whether they are trustworthy. To this end, we propose a three-level model of trust analysis: rational, motivational and personal or intimate. We will argue that some relevant forms of caregiving (especially care for highly dependent persons) require a very personal or intimate type of care that distinguishes it from other contexts. Nevertheless, this is not the only type of trust happening in care spaces. We will adduce that, while we cannot have intimate or highly personal relationships with robots, they are trustworthy at the rational and thin motivational level. The fact that robots cannot engage in some (personal) aspects of care does not mean that they cannot be useful in care contexts. We will defend that critical approaches to trusting SCR have been sustained by two misconceptions and propose a new model for analyzing their moral acceptability: sociotechnical trust in teams of humans and robots.

Keywords Social care robots · Ethics · Trust · Trustworthiness · Sociotechnical trust

1 Introduction

Are robots ethically apt for performing in care contexts? The literature on the ethics of social care robots (SCR) have provided different arguments to try to answer this question. Such an answer should include a satisfactory ethical account of care and a solid framework for understanding how care is performed nowadays.[1] In this paper, we want to address the problem by focusing on an underdeveloped issue: trust.

In contemporary Western societies, the capacity of social structures in place for providing adequate care for all is in question. Despite the crisis of care in which we are immersed (Tronto 2015; Boada et al. 2021; Ausín et al. 2023), it must be analyzed whether the care provided by robots would not impoverish the tasks of care in such a way that this activity would lose its most significant features. While some authors have emphasized the universal condition of the human need for care (Butler 2004; Fineman 2008; Dodds 2013; Kittay 2019), it is also true that some situations in life imply a higher need of being cared for by others such as old age, infancy, or cognitive diversity, among others. These states do not in themselves imply frailty (Rodríguez Díaz et al. 2014b) but can generate a high dependency in

✉ Joan Llorca Albareda
joanllorca@ugr.es

[1] Department of Philosophy I, University of Granada, University Campus of Cartuja, 18011 Granada, Spain

[2] Institute of Philosophy, Spanish National Research Council, Albasanz Street, 26-28, 28017 Madrid, Spain

[3] Department of Nursing, University of Granada, Health Technology Park, 18016 Granada, Spain

[1] It is important to differentiate between various types of care, such as "professional health care," which involves specialized assistance in the field of health; "care for activities of daily living," which refers to assistance with daily tasks; and "dependent care," which encompasses care for people who need help, either temporarily or permanently, among others. In this article, when we speak of "care" we refer to care in a broad sense. In those cases in which we are referring to care in a specific sense (formal, informal, for caregivers, for vulnerable, sick or dependent persons), it would be specified in due course.

Published online: 17 May 2025

the performance of activities of daily living (ADLs) that are related to higher levels of emotional distress in these people (Rodríguez Díaz et al. 2014a) and imply a high demand for care. More caregivers are increasingly needed despite the fact that, paradoxically, care conditions are precarious and caregivers are often in a highly vulnerable[2] position (Copé 2021). This situation calls for solutions that provide the highly dependent population with the necessary care and caregivers with better conditions.[3] In this context, the recent developments of social robotics seem to make possible the introduction of SCR in the care systems, with the aim of contributing to tackling this care crisis.

Assistive care robots are those devices that have functions related to care and/or protection in clinical, welfare or social environments. The actions they can perform include health care, physical and cognitive rehabilitation, activities related to daily domestic life, and educational activities in different environments (hospitals, nursing homes, homes, and schools) (Boada et al. 2021). In this work, we refer in particular to those robots with linguistic and social capabilities due to their potential to establish human-like relationships (social robots), that are embedded in caregiving context, i.e., the so-called SCR. The ethics of care robotics raises concerns about the moral implications of implementing these technologies. The critical views highlight the limitations of social robots in providing good care because of their inner capacities, that is, robots are not able to show true concern or develop empathy. We aim to turn the structure of conventional understandings on its head: we will defend that what matters in the implementation of care robotics lies not in the properties of these entities, but in the relationships we maintain with them (Llorca Albareda 2023, 2024). In particular, we will show how an analysis of the possibility of trusting SCR constitutes a powerful approach to answer the question about their moral acceptability.

The contributions to the discussions about the moral acceptability of the development and implementation of SCR are growing in number (Vallor 2011; Sharkey and Sharkey 2012; van Wynsberghe 2016; Hämäläinen 2020; Martínez-López et al. 2024, Liedo et al. 2024, Liedo 2024) and important arguments have been made, both in a more optimistic aim (Borenstein and Pearson 2010; Sorell and Draper 2014; Coghlan 2022) and in a more reticent view (Sparrow and Sparrow 2006; Turkle 2011; Sparrow 2016). Regarding trust, there has been extensive discussion about the type of trust in digital environments (Taddeo 2009; Buechner et al. 2013; Tavani 2015) between humans, between humans and artificial systems, and between artificial systems (Coeckelbergh 2012; de Laat 2016; Grodzinsky et al. 2020). To our knowledge, discussions about trust in SCR have been mainly dedicated to two types of questions: (i) what perspective does trust bring when designing care robots (Yew 2021) and (ii) what kind of trust is developed in actual encounters between humans and care robots (Poulsen et al. 2018a, b; Song 2020). The first consideration is normative and presents itself as an interesting theoretical tool for understanding what kind of ethical criteria should govern the programming of a care robot; and the second is factual in the sense that it asks what trust relationships are possible between humans and SCR. We believe, however, that next step in the analysis has not yet been taken: the idea that the moral acceptability of SCR is intimately linked with the desirability of the kind of trust that one can have in these artificial entities.

For investigating this issue, we present a categorization of trust relationships. We defend that, although personal trust relationships with robots are untrue and morally troublesome, this is not a reason to deny the moral acceptability of SCR on the whole. Robotized caregiving relationships can involve morally acceptable forms of trust that can enhance existing personal trust relationships between human caregivers and cared-for persons. Clarifying what kind of trust we should maintain with SCR will allow us to overcome the criticisms and delineate which types of activities and relationships should be attached to these entities. We will show that SCR are trustworthy at the rational and thin motivational level and we will offer a new model for analyzing trust in care relationships: socio-technical trust in teams of humans and robots.

2 A new model for trust: the context of care

New AI systems acquire a certain degree of independence from their programmers, thereby differing from other artifacts because their actions and processes cannot be fully predicted (Floridi and Sanders 2004; Matthias 2004). This makes the question of which and how much trust we can have in these systems gain weight: it seems that we can trust them in a different sense than other artifacts, but without reaching the levels of trust in humans (Tavani 2015).

An important distinction made in philosophical analyses of trust is that which differentiates *trust* from *trustworthiness* (Hardin 2006; Tallant and Donati 2020).

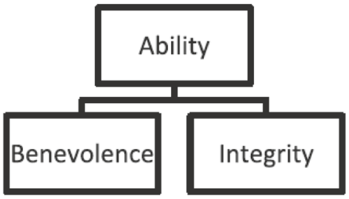
[2] Needing care is not limited to any particular so-called vulnerable population. Likewise, the need for care does not necessarily imply any specific vulnerability if the needs are met through good care. However, the lack of appropriate care, especially when it comes to people highly dependent, is a harm. At the same time, because of precarity, caregivers are also subjected to specific vulnerability to being abused, overwhelmed, underpaid and other problems (Llácer et al. 2007).

[3] We are referring to better working conditions in the case of formal care in institutional or domestic settings and a reduction in the overburdening of the caregiver's role in the case of informal care, mostly carried out by the family environment of the cared-for person. Often, this overburdening of the caregiver's role has a major impact on their health in the medium and long-term (Akalin 2007; Bravo-Benítez et al. 2019).

Trust refers to the *attitude* of the trustor who believes that the trustee will perform as expected; whereas trustworthiness refers to the *qualities* of the trustee that make her worthy of trust. Applied to the debate on SCR, we note that this distinction is fundamental. Because of our tendency to anthropomorphize, we know we can trust robots *qua attitude*; those attitudes have been suggested to be qualitatively different from those commonly established with other artifacts (Poulsen et al. 2018a). However, on an ethical level what is relevant is not whether it is possible to have this attitude, but whether SCR possess the necessary qualities to be trustworthy.

The qualities required to be trusted can be very different according to the reasons why we trust. This raises another distinction. Hawley (2012) distinguishes *trust* from *reliance* in the following sense: when we rely on a person because of the skills she has, we may not care whether her motives are good or bad. We rely on her because she can do an action effectively and there is a high probability that she will do it. In contrast, trust is much more demanding. We often trust people because we know they have good motives and are morally of sufficient integrity. They must have certain skills to perform a certain action, but what concerns us most is that they are genuinely interested in our well-being.

Nickel et al. (2010) understand reliance as *rational trust* and trust as *motivational trust*. Rational trust is concerned only with the efficacy of a given entity to perform a given action. They emphasize that this view does not differentiate between humans and artifacts: rational considerations about the efficacy and probability with which an action is performed do not attend to internal motives. Motivational trust does focus on internal motives since it understands that these are the basis of true trust. Therefore, the requirements for trustworthiness vary depending on whether we are trusting rationally or motivationally. Both conceptions assume that, in trusting another entity, we are assessing whether that entity has the qualities necessary to be trustworthy, i.e., whether we have reasons to trust rationally or motivationally. Figure 1 shows both conceptions of trust.



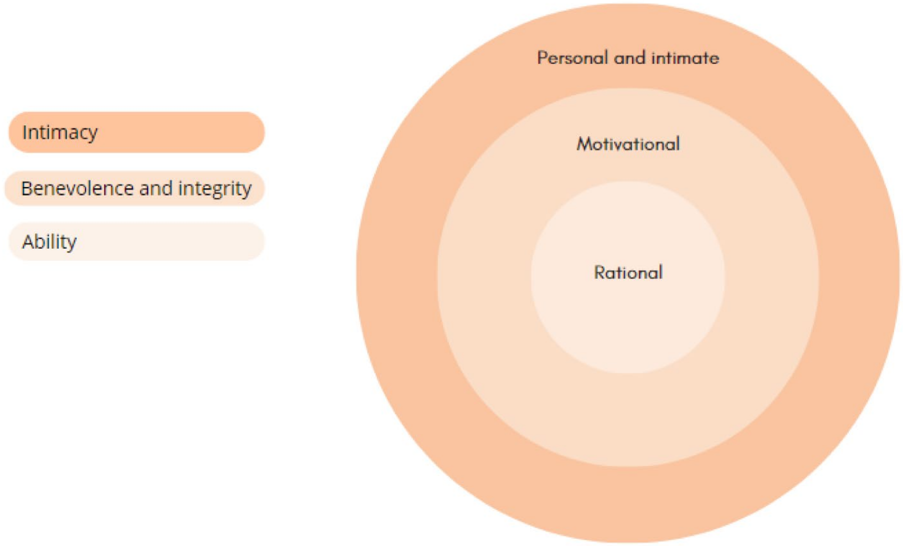
Trust is articulated, first, on the basis of certain abilities that make it possible to carry out an action and, second, on the intentions and character of the trustees. Both dimensions show the difference between rational and motivational trust and how the latter requires the former.

Fig. 1 Conceptions regarding trust (in Sica and Sætra 2023)

While this distinction is of great help for discerning the moral significance of the trust people tend to build towards robots and their trustworthiness, we consider that this categorization can be further nuanced to better acknowledge the role of trust in care contexts. The logic behind many trusting attitudes does not exactly respond to this scheme. Sometimes, people will not be able to give clear reasons for their trusting attitudes, whether motivational or rational. Often, the individual enters into a certain context in which there are certain trust relationships that she implicitly assumes and reproduces. Two concepts help to grasp this phenomenon: *zones of trust* (Walker 2006; de Laat 2016) and *relational trust* (Coeckelbergh 2012). The zones of trust refer to the different ways of trusting that are articulated in different contexts and human activities, and relational trust to the default mode of trusting certain people, institutions and artifacts with which we tend to relate. In this sense, in each context we tend to trust rationally or motivationally without explicitly elaborate the reasons why we are trusting certain entities. For example, in a context in which journalism has a good reputation, people will tend to trust the media by default.

In the context of caregiving, trust relationships are peculiar. While technical skills and good motives are needed to care, it is also marked by the requirements of the ethics of care, which can be claimed to be qualitatively different from other areas of life (Kittay 2019) and potentially problematic when it comes to robots. This is due to the fact that some relevant aspects of caregiving relationships require much more "human" entrustment than in other domains, as it is crucial that the caregiving relationship arouses authentic emotions, and that there is no objectification (Sparrow and Sparrow 2006; Sharkey and Sharkey 2012). The complexity of cared-for persons' interests and the possible vulnerability of a person in need of care (especially those who are in high need) suggest that the relationships between both parties are likely to be significantly intimate and personal. It is probable that people involved in a care relationship need to know each other quite well and probably some of this knowledge concerns intimate areas of their lives. In some care relationships, such as child raising, people need to personally trust each other for the relationship to be successful. Therefore, the rational and motivational requirements for trust do not seem to exhaust all the possibilities of trustworthiness: there is a condition of intimacy or very personal trust that exists in close relationships or relationships that necessarily concern intimate areas of

Fig. 2 Levels of trust



These three levels are classified according to the stringency of the reasons and expectations concerning trust. Every level presupposes the previous one. Motivational trust, even if its motives shift the gaze from technical skills, demands such skills. Without those skills, no matter how morally upright the entity might be, it could not perform the action properly. The same is true of the condition of intimacy. The entity with whom one has a very personal or intimate relationship is believed to have the skills to perform the action, and be able to be morally upright and seek the good of the trustor.

life (such as friendships or love relationships). In Fig. 2, we propose a three-level understanding of trust.[4,5]

While this humane aspect of care has been a main concern for the most critical commentators, we consider that, in terms of trust, it requires further nuancing. Our idea of "intimate trust" is aimed to acknowledge the type of trust required for those aspects of care characterized by intimacy and bonding. We consider personal/intimate trust as a relatively relevant trait of care because (i) it is present in a sufficiently significant number of care practices, (ii) intimate trust is relatively much more important in those care practices in which it happens, compared to other areas of life, and (iii) (personal) trust has been theorized as a main ethical feature of care ethics, as opposed to more rationalistic, abstract moral theories (Baier 1986; Held 1990). It is, therefore, worth exploring whether robots are trustworthy at this level to establish the role they should play in care contexts.

Notwithstanding, we highlight that not all care relationships require strict personal trust to be successful. Admittedly, some care activities happen without personal trust; take, for example, the role of an urgent care nurse who is only deemed to extract some blood from the patient and will never see her again. This activity is always subjected to the standards of care and nurse ethics, but the patient does not need to personally/intimately trust the nurse. The patient may not trust her nurse personally but she can trust in her professionalism or in the system, as trust within the healthcare system (and in other areas) is part of a network where different agents interact as a web (Martínez-López et al. 2023). In our analysis, we will consider the adequacy of robots regarding all types of trust happening in care.

[4] It is important to emphasize the distinction between trust and trustworthiness. This figure shows the possible trust relationships that can be maintained in a care environment. However, as we have pointed out and will continue to develop in subsequent sections, we can trust an entity in a certain manner without deserving it. This is why it is so significant that, in addition to highlighting the types of trust that occur de facto in certain contexts, an analysis should be made of whether the recipients of the trust are really worthy of it.

[5] Another fundamental aspect of our categorization is that the more demanding levels presuppose the less demanding ones. This requires two clarifications. First, there may be tradeoffs between different levels. For example, if I trust someone at the motivational level, this trust will also include the rational level. However, there may be other people or artifacts that are better recipients of rational trust, either because they have better skills or because they perform tasks more effectively. Since my trust is motivational, I can compensate for a loss in rational trust by a greater concern for my welfare or a more morally upright character. But I need the entity I trust at the motivational level to be minimally suitable or skilled to perform the activity I want them to perform. The same goes for the personal level and its motivational and rational presupposition. This brings us to the second point. Each level points to the main reasons and motivations for trusting. The presupposition does not tell us that our motives contain other levels, but rather that, in order to trust in the more demanding levels, minimum trust requirements must be met at the previous levels. To trust personally, the entity must be minimally skilled in the performance of its activity and capable of having concern for others and a morally upright character.

To summarize, we have distinguished between three basic types of trust: (1) rational, (2) motivational, and (3) personal/intimate. While (1) and (2) have been theorized by previous literature, we have added (3) in an attempt to grasp the specific type of trust happening in some practices of care, arguing that it requires a high degree of vulnerability, intimacy, and embodied knowledge of each other. We do not claim that all care relationships need personal trust, but we do claim that it is a type of trust sufficiently relevant in sufficiently common practices of care that we can consider it a distinctively important aspect of care. Each type of trust presupposes the previous one(s). Furthermore, we have introduced the concepts of zones of trust and relational trust to acknowledge the ways in which trust attitudes tend to happen in social settings. In what follows, we further delve into the implications of this categorization for the account of trust in roboticized care environments.

3 Uncovering robot trustworthiness

Care is fundamentally relational and the care relationships can be constituted by different actors and be of different nature. In many caring relationships, there are one or more people in a situation of specific vulnerability and need of care, for example, if they are infirm. In this kind of relationship, trust is at the core. An important part of this relationship is, on the one hand, the belief in what the other party is and can do (trust); and, on the other hand, whether the belief regarding the qualities of the other party is true (trustworthiness). Consequently, we propose that the ethical appraisal of trustworthiness in care contexts should focus on the kind of relationships that would be desirable when a robot enters the scene. Let's see how this approach applies to the general discussion on the ethics of SCR.

Sharkey and Sharkey (2012) have identified six major ethical problems with care robotics: (1) lack of opportunities for human contact; (2) objectification of humans in care tasks; (3) loss of privacy; (4) restriction of personal freedom; (5) deception and infantilization of vulnerable groups; and (6) potential responsibility gaps. The problems related to (3) privacy, (4) freedom and (6) responsibility are not specific to care robotics. There may be deprivation of personal liberty through Internet of Things (IoT) systems or mobile device surveillance. The loss of privacy stems from the pervasiveness of certain technologies and their ability to collect sensitive information, not from their link to care (Véliz 2021). Finally, the responsibility gap is based on the fact that machine learning systems make decisions that are not previously programmed, but these entities do not possess mental properties necessary for them to be attributed responsibility, which is a danger consubstantial to all machine learning systems (Llorca Albareda 2025; Matthias 2004; Sparrow 2007). All of these ethical concerns may possibly be aggravated in dependent care environments, as they are a population potentially vulnerable to certain types of harm, but they would not be qualitatively different to other AI ethics discussions. Because of that, in what follows, we will focus on the problems more specific to SCR since we consider that those are the ones that can be benefited by our account of trust in care.

There are problems identified by Sharkey and Sharkey (2012) that might be specific to the use of SCR for dependent care and deserve more careful attention. These are the ethical problems related to (1) the lack of opportunities for human contact, (2) the objectification of human beings in care work, and (5) deception and infantilization. Substitution of human care by a robotic one could be effective in the designated task (e.g., feeding) but in some cases people dependent to perform activities of daily living find themselves suffering loneliness, in institutionalized and/or hostile environments, and human and friendly contact is reduced to these daily tasks. Decreased human contact for these activities could have serious consequences on the mental health of these people. Moreover, caregiving seems to include a type of knowledge specific to the cared-for person that cannot be codified, as their needs are changeable and contextual, so there is a serious danger of objectification (Sparrow and Sparrow 2006). Finally, the mental capabilities that SCR seem to exhibit are not real and success in performing their tasks depends on the deception and infantilization of users (Sparrow 2016).

These three ethical problems specific to care robotics are clearly relational: robots can deceive us, objectify us and make us lose human contact because we believe that they have the necessary qualities to perform care tasks in a similar way to humans, asking them for more than they can give. We consider that a more in-depth account of trust and trustworthiness of SCR would benefit the debate around these three issues. Indeed, we could have relationships with robots that are appropriate to their capabilities, trusting them to do tasks for which they are suited. However, is caregiving one of them? We will discuss below in what sense we can say that robots are trustworthy.

3.1 Rational trustworthiness

Trust in a purely rational sense is based on the subjective attitude of the trustor regarding the expected consequences of the occurrence of a given outcome (Coleman 1990). The entity performing the action must have certain properties that make it likely and effective (Nickel et al. 2010). And usually this type of trust is domain-specific, since it depends on the performance of one or a particular set of actions (Sica and Sætra 2023).

As we exposed above, this notion of trust is reduced to what is understood as reliability (Whittingham 2003; Hawley 2012). According to Nickel et al. (2010) reliability is

> a characteristic of a person, expressed by the probability that the person will perform his/her required function under given conditions for a stated time interval (…) From a qualitative point of view, reliability (is) defined as the ability of the person to remain functional. Quantitatively, reliability specifies the probability that no operational interruptions will occur during a stated time interval. (pp 433–434).

What is crucial in this understanding of trust is that one does not have to trust on the basis of certain internal qualities of the entity, i.e., it does not imply trusting another entity because of the mental properties it possesses. What is essential is that it performs what is expected in a probable and effective way. In fact, internal qualities can affect negatively. Some authors have pointed out that emotions make human agencies much more unpredictable (Nadeau 2006; Arkin 2009) and, therefore, would be less reliable.

From this paradigm, SCR can be trustworthy. So can other types of artifacts. As long as they meet the stipulated conditions, artifacts and people will be worthy of the same kind of rational trust. There is no difference between the simplest artifacts, SCR and people. The determining properties of each are only important if they help or hinder the probability and effectiveness of the outcome taking place.

3.2 Motivational trustworthiness

It could be contended that rational trust does not advance our argument much. Hawley (2012) defends that the difference between trust and reliability lies precisely in the fact that the former is normative and the latter is not, i.e., only trust is morally relevant. This is because we place our expectations and hopes in agents who have values and the ability to do the morally right thing. Artifacts merely perform a limited set of operations likely and effectively but have no such thing as a morally upright character or emotions and desires for our welfare. Hence, a distinction has been made in the literature between predictive trust and normative or affective trust (Hollis 1998; Faulkner 2007).

Sica and Sætra (2023) argue that the two aspects of motivational trust, as introduced above, are moral integrity and beneficence—two characteristics that are not domain-specific. On the one hand, moral integrity responds to a character or ethical articulation that can only occur in a moral agent, a condition that is not specific to a particular activity but derives from the properties of the agent. Although it is also possible to be morally upright according to the principles of a particular practice, a high degree of generality is still required. On the other hand, beneficence is usually understood as a universal interest in the welfare of others, not reducible to a specific context. One wants the good of another person in different spheres. Its specificity lies in the adequate knowledge of the interests of others: interests are contextual and are articulated differently according to the environment in which they are found. In this sense, current AI systems do not seem to be enabled for the realization of activities that go far beyond the domain-specific character and, in order to have these capabilities, the possession of internal properties seems to be necessary. By extension, it seems that only if SCR possessed internal capabilities could then be trustworthy in a normative sense.

There is no consensus in the literature about the possibility of AI systems being moral agents (Llorca Albareda et al. 2024; Floridi and Sanders 2004; Himma 2009; Laukyte 2017). And by this we do not mean that they may possess the properties linked to moral agency in the future. On the contrary, AI systems today or similar to those already present can be understood as such. Moor (2006) offered a famous formulation of the different ways in which AI systems could be moral: (i) their operations can have moral impact; (ii) they can be implicit moral agents; (iii) they can be explicit moral agents; (iv) they can be full moral agents. The first is limited to the moral consequences that any artifact can have, but the second and third already introduce new modalities of moral agency, other than the human, the fourth. Implicit moral agents are those that possess an internal mechanism that limits certain morally dangerous functions or uses (e.g., the lock on the gun or the airbag in the car). The explicit moral agents, however, are those that can reason ethically, even though they lack properties such as conscience or common sense. Their programming can incorporate certain types of ethical reasoning (Wallach and Allen 2009).[6]

Tavani (2015) has taken up Moor's classification and argued that normative trust is not a dichotomous issue, but a gradual one: there are various ways in which we can trust an artificial entity and each type of trust has particular normative implications. In this sense, explicit moral agents may be motivationally trustworthy, since they can reason ethically given certain programming and in a very context-appropriate manner. That is, they can be morally upright within certain domains. Moreover, they can also be beneficent. They can identify the interests of the people with whom they interact and help to promote them through their ethical programs. While they cannot achieve the universally reaching capabilities that full moral agents possess, explicit moral agents can be morally upright and beneficent in a domain-specific

[6] Van Wynsberghe (2016) has taken up the discussion for the specific case of care robots. She has defended that robots should be considered moral agents only in the operational sense. Indeed, their actions will be morally appraisable, thus the discussion should focus on how to assess the responsibility of actions performed by a care robot, considering the centrality of responsibility in care practices.

sense. For these reasons, Nickel et al. (2010) have spoken of *thin motivational trust*, i.e., one that incorporates beneficence and moral integrity at a narrow level.

Therefore, not only can SCR be trustworthy in a rational sense, but they can also be trustworthy at a thin motivational level. It is not the case, hence, that we should not trust robots. At this point, we should raise the question whether SCR are trustworthy on a personal or intimate level. The context of care is characterized by a strong emphasis on the latter and a possible refusal could undermine the prospect of introducing these entities in care environments.

3.3 Personal trustworthiness

The search for the specific good of others and the development of intimate relationships involve mental capacities that care robots do not have (Sparrow and Sparrow 2006). They cannot know what exactly our good is or cannot fully empathize with us: they cannot *care about* us. So, all those behaviorally emotional responses that care robots display seem to hide their true nature (Scheutz 2011). In this sense, simulating the capacity of engaging in trust relationships can imply ethical problems such as deception.

Nonetheless, it must be noted that the absence of the qualities to be trustworthy can be a technical, but not a metaphysical limit. The debate about the current mental capabilities of robots and their role in the interaction between persons and robots is not straightforward (Bryson 2012; Nahmias 2016; Frank and Nyholm 2017). The possibility of sustaining intimate relationships with robots has been analyzed mainly from two kinds of debates: the feasibility of romantic (Levy 2008; Richardson 2016; Danaher and McArthur 2017; Nyholm and Frank 2019; Gordon and Nyholm 2021) and friendship relationship between humans and robots (Marti 2010; Danaher 2019; Nyholm 2020; Prescott and Robillard 2021; Ryland 2021). Although care is not the same as a relationship of love and friendship, the analysis of the latter allows us to assess the reasons why it is or is not possible and desirable to maintain an intimate relationship with a robot and, therefore, to have personal trust in it. We will discuss this question from Danaher's (2019) argument in defense of robotic friendship.

Danaher departs, like many other theorists who analyze the question of robot friendship (Elder 2014; Nyholm 2020; Ryland 2021), from Aristotelian categories. Aristotle articulated three types of friendship: utility-based, pleasure-based, and virtue-based. The first two do not consider the good of the other and are imperfect forms of friendship, which suit self-interest. On the contrary, virtuous friendship is a fulfilling form of friendship that takes into consideration the good of the other and that is behind our highest ideals of friendship. Virtuous friendship is only possible if it satisfies four conditions: (i) mutuality; (ii) authenticity; (iii) equality; and (iv) diversity of contexts. Friends must reciprocally desire each other's good, experienced in a sincere and authentic manner, considering each other on an equal footing and whose profound relationship crosses different human spheres and is not restricted to one or a few. Danaher considers that all these conditions can be fulfilled by robots, or at least as well as by humans. The last two conditions (iii and iv) do not take place perfectly among humans. With respect to the first two conditions (i and ii), the objections seem more serious. It does not appear that robots can currently have genuine emotions or reciprocate human affection. Danaher argues, on the contrary, because of the epistemological impossibility of accessing the mental states of entities other than oneself, the only way to ensure that both conditions are met lies in attending to the behavior of the entity. If the entity corresponds to us behaviorally in a consistently reciprocal and authentic manner, then that entity can be our friend. Since robots can have these consistent behaviors, then they can be our friends.

The conclusions drawn by Danaher are crucial to our discussion. The conditions of equality (iii) and diversity (iv) are not particularly enlightening for the case of caregiving, since dependent care sometimes involves asymmetrical relationships in terms of vulnerability, dependency and power, and is usually restricted to a specific sphere of activity, e.g. healthcare. However, acknowledging each other's vulnerabilities, needs and desires requires a certain degree of mutuality (i) and authenticity (ii). If epistemologically we lack the means to know whether SCR possess inner qualities, then we must accept a caregiver whose behavior is consistently reciprocal and authentic. If the trust conditions that enable a friendship can be condensed into i–iv –i.e., I should only trust that it will be my friend if it can fulfill those conditions–, then the (personal) trust conditions that enable care can be considered i and ii. Therefore, i and ii, according to Danaher's behaviorism, can be satisfied by SCR, then it seems that we would be able to have intimate and personal trusting relationships with robots.

But we are not, for the following reasons. First, current SCR do not engage in consistently reciprocal and authentic behavior, not even successfully mirror it. On the contrary, although we find important developments of robots that correspond verbally and non-verbally to human emotions (Breazeal 2004; Coghlan 2022), they are not able to fully mimic human relationships as a whole. This is primarily due to a lack of a type of knowledge: robots and other AIs do not possess the kind of contextual, intuitive, and embodied knowledge that characterizes humans (Dreyfus 1992, 2007). Precisely, this kind of embodied and practical knowledge is very important in care (Mol 2008). For this reason, van Wynsberghe (2022) has argued that we should reconsider the sense in which we want SCR to be reciprocal: being designed for this goal may have the sole purpose of the robot gaining greater social acceptance and may undermine current practices of reciprocity between human caregivers and care-for persons. These robots should seek to maximize bonds that are truly bidirectional.

Second, Danaher (2019) omits to incorporate in the conditions of virtuous friendship a fundamental one: friends should be good and seek our good. This is true for friendship, but also for care, and it is furthermore a relevant trait of wide motivational and personal trustworthiness, as stated above. Robots have no character and cannot develop the virtues related to caring. This leads to two problems. On the one hand, neither do robots have the internal mental capacities necessary to be good nor do they display consistent behaviors that demonstrate this (Nyholm 2020). On the other hand, a fundamental aspect of caregiving is neglected: the virtues and goods associated with the practice of caregiving (Vallor 2011). Caregiving does not take place in isolation or in a non-reciprocal manner. The very exercise of care is not only valuable for the person being cared for, but also for the caregiver herself, thus building a network of relationships that perform a function of supporting the social fabric essential for the maintenance of any human group.

Third, the Aristotelian–Danaherian view of friendship postulates a sort of "equal vulnerability" between the parties, that is, we are thinking of two people whose situation regarding their power to hurt one another is roughly similar. Care ethics approaches human relationships with a greater attention to vulnerability and dependency inequalities. When talking about dependency care, healthcare or child-raising, among other care relationships, the vulnerability of the parties is unevenly distributed. People in a situation of high dependency, can be more likely to be deceived and thus harmed (Scheutz 2011). The dangers of deception and objectification are much greater than in other types of relationships. Children, for example, may place a very high level of trust in robots without being fully aware of what they can actually do for them. As Annette Baier (1986) argued in her classical feminist report of trust, traditional accounts of trust are too focused on relationships among equals who voluntarily enter into mutually beneficial pacts. Many relationships do not thrive in such conditions but are characterized by a higher degree of dependence and asymmetry of power. This is often the case in care.

In sum, if we understand the conditions of personal trustworthiness that make care possible to be (a) authenticity, (b) goodness, and (c) mutuality, then SCR are hardly suitable entities to trust personally. They lack the kind of knowledge necessary to perform good care and do not have the mental capacities or the consistent behaviors that display this kind of knowledge. In addition, end-users who lack a specialized knowledge on the functioning of a robot can be more likely, without the necessary measures and safeguards, to be misled about the capabilities of such artificial entities.

However, the fact that we should not place personal trust in them does not mean that they are not trustworthy in other senses as we point previously. The question then arises as to whether this condition is sufficient to reject the ethical acceptability of care robotics. Following the categorization of trustworthiness developed in this section, we have elaborated Fig. 3 to illustrate the extent to which SCR would be trustworthy within each of the proposed levels of trust. In what follows, we show the implications of our model to the general debate on the moral acceptability of SCR.[7]

4 Normative and practical implications of trustworthy human–robot care relations

4.1 Debunking misconceptions about trusting social care robots

There have been significant objections raised about the use of robots in care contexts (Sparrow and Sparrow 2006; Scheutz 2011; Turkle 2011; Sparrow 2016). It is generally argued that robots lack the necessary properties to perform adequate care work, and therefore, they are not suitable for engaging in relationships with humans that mirror those which are typical to human care settings. In this article, we have argued that this approach is limited: we defend that, while discussions about the inner capacities of robots can be enlightening in some aspects, an approach that focuses on the kind of relationships people should have with robots can be much more fruitful in the case of SCR. One crucial aspect of this relationship is trust: i.e., what we believe they can do for us and what they can actually do for us. It may be that we can maintain beneficial and ethically acceptable relationships with them in a caring context.

Critical authors have arguments that point out the negative consequences of introducing SCR. This is due to what Danaher (2019) has called the *corrosive effect*: while SCR can be used to complement or enhance human activities, they will progressively lead to replacing humans in social and inter-individual relationships. This total displacement generates a major problem. Since SCR allegedly lack

[7] An interesting aspect of these three ways of being trustworthy is the particular object of trust, i.e., what specific element or part of the entity is involved in the trust relationship. We can identify three: actions, dispositions, and mental states. As we have shown, trust is usually articulated in specific contexts and occurs by default, making it difficult to point to a particular element. Nevertheless, each of these forms of being trustworthy emphasizes specifically one of these elements: rationally, we trust effective and probable actions, beyond the nature of the entity; motivationally, we trust dispositions, ethically relevant forms of character and attitudes towards our good; and personally, we trust that the entity has mental states and that these are authentic and reciprocal.

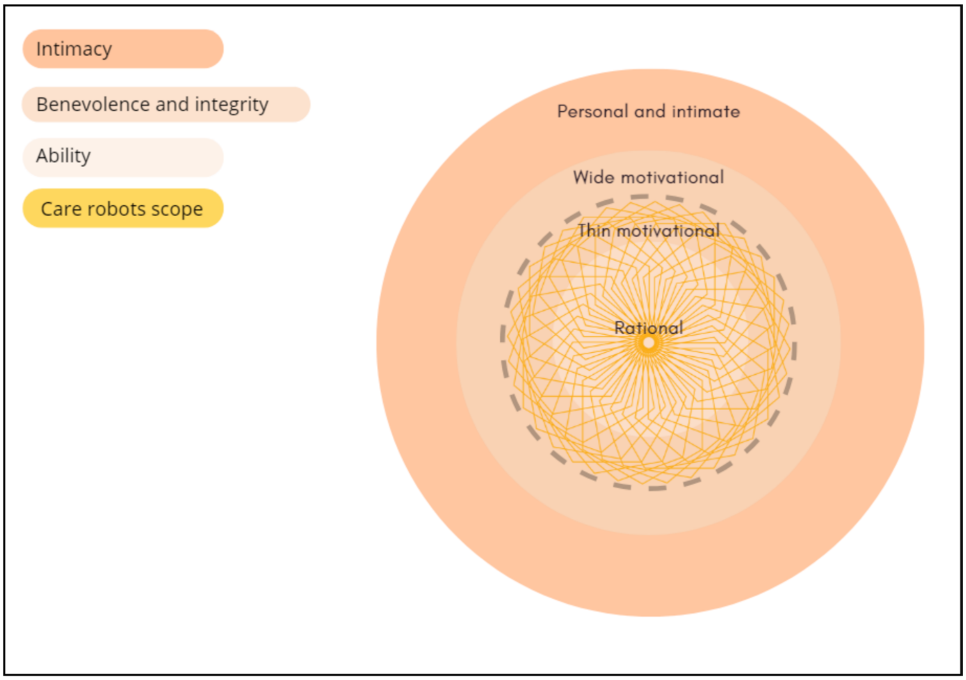


Fig. 3 Trustworthiness of social care robots

adequate capabilities to fulfill the conditions necessary to perform good care as a whole and to be proper recipients of personal trust, they would impoverish the social ideals of care. Sætra (2022) raises this issue in the context of prospective loving relationships with robots. If these were to take the place of human-to-human relationships, our ideas about love and sex would lose many goods we hold dear. They would be much less rich and fulfilling relationships. In the case of care, human contact would be lost; the treatment would be much less personal, intimate, and contextual, so that we might tend toward objectifying treatment of the persons cared for; and they would be relationships characterized by deception.[8] In what follows, we will reframe these discussions following our characterization of trust and show how it can help to clarify some of the moral pivotal points.

As we have shown, each human sphere has its own particular trust relationships (trust zones) and these are typically assumed by default when participating in these spheres (relational trust). In care spheres, personal trust is salient. By default, the types of relationships that take place in care settings require the presence of mental life and emotions in the trusted entities. In this sense, it would seem that, by the very nature of care environments, robots would not be able to participate otherwise than in relationships of a very personal kind. Therefore, if the care environment demands this type of trust and robots cannot provide it, the values and ideals of care would be impoverished by the implementation of such entities.

Notwithstanding, we consider this rationale to be flawed for two reasons: (i) it erroneously assumes that the type of trust relations placed depends only on the type of entity, i.e., the properties it possesses; (ii) it states controversially that each context has a specific type of trust and that of the sphere of care is personal trust.

On the one hand, we challenge the idea that the type of entity is the crucial element that determines the type of trust that can be placed on it. Although we do not want to establish an exhaustive index of the conditions that create trust, we do want to highlight that the entity trusted does not determine by itself the type of trust people can place on it. We do not always trust human beings on a personal basis. When we go to a shop, we do not trust the seller to have a fair price and a safe product because of the intimate relationship we have with her, but we mainly trust the internal mechanisms of the market and the regulatory entities that ensure the safety of the goods under consideration. Nyholm and Smids (2020) have defended this idea applied to social robots in the workplace. In the world of work, we do not need our relationships with our coworkers to be personal. We need them to achieve work purposes, have pleasant and informal conversations, treat colleagues well, or be sensitive to the work of colleagues and adjust to their rhythms. And this can be done by an entity with which we can maintain thin motivational trust. For this reason, while the properties of an entity determine the maximum degree of trust we can have

[8] Farina et al. (2022) have made a similar argument from a virtue ethics perspective: we need to reflect on which social goods, such as love or friendship, are weakened or impoverished by the introduction of AI into certain domains of our lives.

in it, we can have different degrees of trust with a given type of entity. The adequacy of each type of trusting relationship depends, at least partly, in the requisites of the context. The requisites in the case of care leads us to our second argument here: non-personal trust plays a role in care.

On the other hand, it is also clear from the objections to care robotics that the context of care does usually include a very personal or intimate type of trust. Each context has its own zones of trust (Walker 2006) and it can be argued that the context of care has very exhaustive trust conditions. Personal or intimate trust is conceived as the defining kind of trust in care environments and, since robots cannot be trustworthy in this sense, then they cannot participate in this context. However, we believe that here we are mistaking the part for the whole: care has personal trust as a fundamental trait, but it is neither the only type of trust that functions in care practices, nor the only one that should exist. Thin motivational and rational trust has also a place in care, and robots can fulfill the requirements for being recipients of these kinds of trust. In similar terms, Coghlan (2022) has argued that, while robots cannot provide fully humanistic care, they can provide an expressivist variant of it. SCR should not replace human care, but they could constitute an element that could improve the totality of care. There may be other types of non-personal trust with other entities that help and enhance personal trust relationships. Therefore, different types of trust can occur in the same context and, in the case of caregiving robotics, appropriate trust could even help us to improve the current relationships that exist between caregivers and cared-for persons.[9]

4.2 Sociotechnical trust: a new understanding of trust in robot care and its implications

Robots need not degrade the current values and practices of care. As we saw in the previous section, the causal link between their introduction and the impoverishment of care can be understood as related to two misconceptions of trust. We have argued that care robots may be trustworthy at a rational and thin motivational level. In this section, we will show how these types of trustworthiness would fit into a care environment, where personal or intimate relationships are highly salient. A more complex understanding of the types of trust that can function in a care context will help to consider what roles a SCR may successfully play.

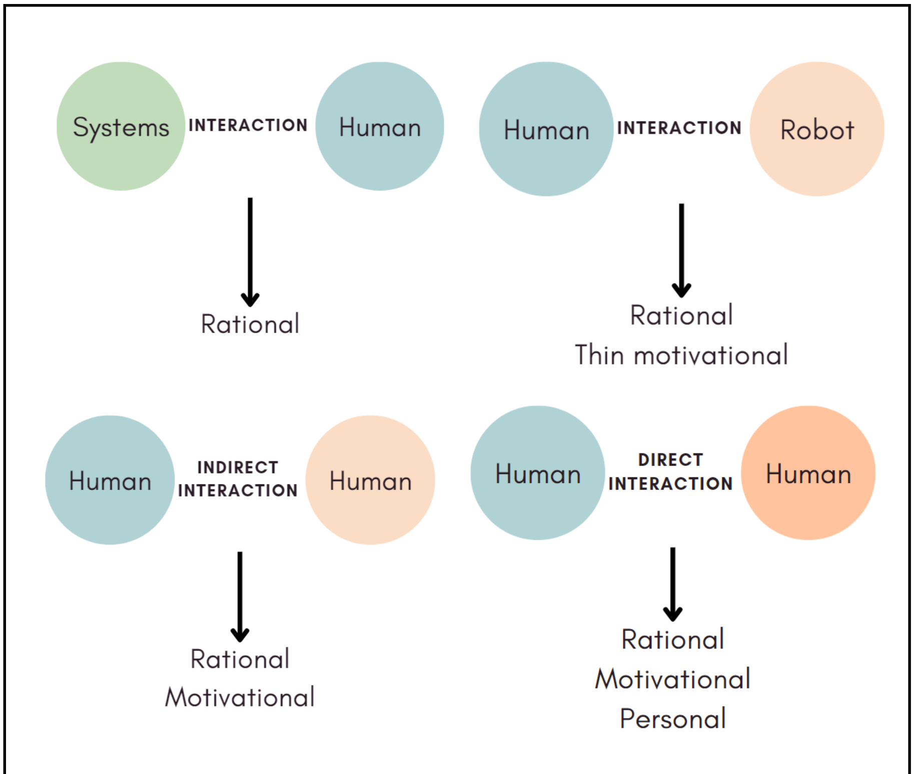
Robots can play a very relevant role in care settings, but this role will only be ethically acceptable to the extent that the participation of human care remains significant. Thus, we believe that the central element is not so much that dyadic relationships of trust between caregivers (human or robotic) and cared-for persons are acceptable, but that the network of relationships in caregiving contexts includes personal trust, as well as other types of trust when it is adequate. For further unpacking our proposal, we will now present an understanding of trust relationships inspired by the notion of socio-technical system (Cooper and Foster 1971; Latour 1992). This concept has been proposed from Science, Technology and Society studies to account for the collective intertwining between human and technical activities. And this is precisely what we intend to do: to show how the moral acceptability of care robotics and the enormous benefits it can bring derive from a good coordination between human care and organization, and the care provided by robots.

Some authors have begun to apply this idea to certain AI ethics debates. Nyholm (2018) has argued against the responsibility gap of new AI systems on the basis of the idea of teams of humans and machines. While AI systems do not meet the agential requirements to be responsible, their actions, although independent and not fully controlled by designers and users, should be understood as executions of commands and instructions by humans. Just as a child who performs an action at the request of her parents should not be held responsible and her parents should be, designers and users should be held responsible when placing or using these entities in certain contexts under certain instructions. In this sense, we should not reject a certain entity because it lacks the agential requirements to be held responsible, but rather we should form appropriate teams in which full moral agents take charge of other minimal agencies.

We also find proposals along these lines applied to the automation of care environments, from the emphasis on how technology is phenomenologically experienced by people defended by Coeckelbergh (2009) or the early approach to "robot caregivers" focused on their interaction with people in care relationships (Borenstein and Pearson 2010). Vallès-Peris and Domènech (2023) argued for a "Robot Embedded in a Network", inspired by the Actor Network Theory and aimed to have a more context-aware and nuanced approach to the incorporation of robots to care systems. In the same vein, van Wynsberghe and Li (2019) designed the "Human–Robot-System Interaction" (HRSI) as a way to enhance the capacity of traditional Human–Robot interaction approach to fully acknowledge all the relationships

[9] We should make two clarifications regarding Coghlan's argument. On the one hand, his use of the term expressivist has a different meaning from that traditionally attributed to it in ethical theory. He means that while SCR cannot perform humanistic care, they can *express* a "humanistic kind of care that the recipient knows emerges from human life and peculiar human behaviors" (p 2103). On the other hand, Coghlan does not develop his argument in terms of trust. His thesis is that while SCR cannot provide fully human care, they can deliver an expressive variant of it. That is, they can provide companionship and support in a more adequate way than other technologies, but without reaching human levels. In our view, it follows from his thesis that, once we are aware of the limitations and benefits of robotic care, we can establish appropriate trusting relationships based on what they can actually do for us.

Fig. 4 Sociotechnical trust: trusting teams of humans and robots



The figure reflects the types of relationships maintained in care settings and which kinds of trustworthiness are involved in each of them. The kinds of trustworthiness are distinguished according to the type of entity being trusted (a human being, a robot or the system) and the type of interaction (whether it is direct or indirect).

and actors implied in healthcare relationships. As they themselves suggest, the HRSI model is better equipped for understanding how a robot may impact in the trusting zone of healthcare: “When introducing the robot in between the care receiver and the health care system, the question is whether the care receiver is being asked to trust the health care system, the robot, or the third party involved in the robot’s implementation” (p 18).

Indeed, for the introduction of a robot into a caregiving environment to be ethically acceptable, not only is trust at a rational or thin motivational level in this entity necessary. In addition, the cared-for person should trust the human caregivers with whom one has personal trust, the professionals in charge of advising the implementation of the robot and the care system as a whole. We propose, in this way, to understand trust relationships with care robots sociotechnically from the notion of teams of humans and machines at the three levels outlined above—rational, motivational and personal. SCR will be ethically acceptable as long as we trust them in a rational or thin motivational manner, which implies that, for these trust relationships to take place properly, their development and implementation must be supervised by human professionals and complemented by human caregivers in whom we can have personal trust. The coordination of all these elements will enable rational trust in the system. See Fig. 4 for a visual exposition of our proposal.

From these coordinates, sociotechnical trust in teams of humans and robots would consist of four elements. First, we would find trust relationships with SCR, which, as we have argued, should only be of a rational and thin motivational type. Second, direct trust relationships within humans, which can incorporate personal trust in care environments, would be highly relevant. This is not to say that direct human relationships have to be solely based on personal trust, since other types of trust can also take place, but to raise awareness on the fact that introducing a robot can affect the personal trust between humans. Third, we would also encounter indirect relationships among humans, mainly those that would be maintained without physical or virtual contact with the professionals in charge of supervising the deployment of robots in care settings and ensuring the well-being of patients. Fourth, the combination of the three previous dimensions would lead to the articulation of trusting relationships with care systems, understood as the coordination of teams of humans and robots, in a rational sense. That is, if the three previous elements work and complement each other adequately, we will have trust in the system working effectively with a high probability.

The three main ethical topics specific to care robotics that we stated above, following Sharkey and Sharkey (2012), could benefit from our approach to trust. In this section, we have proposed a new view of trust in roboticized care places. We will now turn to show how our approach can contribute to these three topics. First, we propose that our categorization and our turn to a more system-focused approach would slightly shift the framework of both loss of human contact and objectivization problems. Second, we will briefly engage in the debate about deception and show the new ideas that our categorization can provide.

(a) Objectification and loss of human contact

We have shown that, since care practices are usually formed by complex and intricate webs of relationships, we should be concerned about the impact a robot may have on the system as a whole and not only according to their capacity to enter in one-to-one relationships with particular persons. Critical authors are concerned that, due to the corrosive effect, robotic care will ultimately cover the totality of care practices. This would cause caregiving to lose two fundamental components: human contact and the complex and deep knowledge of the cared-for person's interests. Both problems would lead to a possible objectification of dependent persons.

Our notion of sociotechnical trust in teams of humans and robots accommodates this multidimensional understanding of care. Humans and robots are part of an environment made up of multiple types of trust relationships that must be articulated based on the nature of the interactions and the type of agents involved. Robots can participate in relationships that do not involve personal trust if there is good supervision and coordination. Then, will robots replace direct caring relationships where personal trust is fundamental? Our response refers to the general structure of trust in teams of robots and humans. To the extent that direct interaction with human caregivers is not lost and adequate human supervision and coordination occurs, objectification and lack of human contact need not to occur. In this sense, the robot should avoid interfering in human relationships in ways that can undermine the possibility of establishing adequate trust between the parties, notably personal trust. Hence, the healthcare system is responsible in two ways. First, robots should not imply a crisis of trust in the system, since trust is a relevant factor in care. Second, the robotic market should not take advantage of the trust people already have in the health or care system, if it exists, and be sure that robots are actually reliable and adequate for a context of care.

We believe that such considerations regarding trust help to shift the focus of the discussion regarding both loss of personal contact and objectification. Direct contact with robots need not to provoke those harms. However, since they are arguably grave in a moral sense, developers and institutions should be mindful of the ways a robot impacts the trust system happening in a care setting and avoiding the potential deleterious consequences on its quality that could lead, in the long run, to these two problems. The attention to the effect on trust within the system can be a useful proxy for this precaution.

(b) Deception

We have stated that personal trust requires some kind of reciprocity of which robots are not capable. Therefore, people who place some kind of personal trust in robots would have a wrong appreciation, as authors such as Sharkey and Sharkey (2012) and Turkle (2011) have already pointed out. Since deception has been the topic of extensive debates, we will briefly present the implications of our proposal for the matter. Regarding robots, it is clear that the mere belief that a robot is "competent", that is, that the robot is capable of performing certain tasks satisfactorily, corresponds to reliance or rational trust. Of course, one can be mistaken in their expectations about the capacities of a robot, but this error does not seem to be very different from overestimating the capacities of someone. As we have stated before, when it comes to rational trust, there is no relevant difference between a person and a machine.

A different matter is the case of the care systems generating some expectations that are in fact unattainable. In the case of robots, the more salient possibility is that users believe that the robot possesses some kind of mental or emotional capacities it does not, notably those that would allow to establish a relationship of personal trust. Since, as mentioned, trust qua attitude do not fully depend on rational, informed deliberation, the care systems implied in caring for people should be in some ways responsible from the reputation and expectations raised. Note that the public arguably lacks the sufficient familiarization with SCR due to their lateness, so there is a high risk that people decide to place their trust (or mistrust) on robots depending on irrelevant considerations, such as ideas learned through sci-fi cinema and literature.

Critical views argue that deception constitutes a morally pernicious harm. First, it prevents us from seeing the world as it is, which constitutes a moral failure. Our actions should be guided by a true knowledge of the world (Sparrow and Sparrow 2006). Second, our subjective preferences are not the only thing that counts in our well-being (Sparrow 2016), including the values of good care (van Wynsberghe 2015). Other authors have ease the worries about deception, arguing that it is not always a negative phenomenon in moral terms

(Coeckelbergh 2016) or that veracity is not an absolute value that should be always protected in SCR (Segers 2022).

Another important aspect of deception is its emergence. Some studies have shown that anthropomorphism in social robotics generates higher levels of trust (Hoff and Bashir 2015). This includes resemblance of appearance and social cues such as facial expressions (Kühnlenz et al 2011) or empathy gestures (de Kervenoael et al 2019). However, it seems that anthropomorphism induces a higher degree of deception (Sharkey and Sharkey 2011) which makes it morally problematic for two reasons. On the one hand, private companies can take advantage of the situation to obtain sensitive user data (Scheutz 2011). On the other hand, given the apparent human tendency to anthropomorphize robots, deception would happen even without the intentionality of designers, which makes it more difficult to avoid (Sharkey and Sharkey 2021).

We do not take a definite stance on the moral acceptability of robot deception, but we highlight that the attention to the impact of robots in trusting the sociotechnical system also shifts the main focus on the debate on deception. First, the implementation of SCRs does not have to involve deception. As we have shown, if the right kinds of trust relationships are developed, we will not be deceived about their capabilities. We can maintain relationships that do not overestimate what they can do for us.[10] Second, we align ourselves with the work of Sætra (2021), who investigate the impact robots may have in the culture of trust needed for successful social practices:

> The argument is that in societies built on trust and the expectancy of truthful signals of both superficial and hidden states, repeated deception will erode this trust and change the culture and social norms. This, I argue, is one reasons why robot deception is problematic in terms of cultural sustainability. (p 282)

Following Sætra's argument and our categorization, in the case of care settings deception should be addressed from the point of view of the impact it may have in the general trust relationships that are relevant for the success of care practices. Deception that happens through a robot interaction but whose full agency can be found elsewhere (for example, in the actions of those developers who purposefully want to provoke emphatic sentiments in end-users), the ethical appraisal of the phenomenon can be equivalent to other types of betrayal and it should take into account the type of trust relationship in which the deception is happening and the impact on the system as a whole. In some cases, it could be considered that deception involves a clear harm to users and this evaluation can be informed by the effect the deceptive machine has in how we trust the sociotechnical system.

5 Conclusions

Debates in the ethics of care robotics have discussed at length about the properties of robots and how they may prove beneficial or detrimental to care robotics. In this article, we have proposed a new relational approach, based on the trusting relationships that are possible and ethically acceptable in caregiving environments. To this end, we have proposed a new model of trust that includes three levels: rational, motivational and personal or intimate.

First, we have analyzed whether SCR are trustworthy at any of these levels. They are not trustworthy in the personal sense in the current state of the technology. The internal capabilities or consistent behaviors required for this type of trust to take place are not found in contemporary social robots. However, we found that they are in fact trustworthy at the rational and thin motivational level, other types of trustworthiness that are not as demanding on the internal qualities of robots. We can continue to benefit from the advantages of these new artificial systems without placing inordinate expectations and beliefs in them. In this sense, we may be able to maintain beneficial and ethically acceptable relationships with them in a caring context.

Second, we have argued that critical authors have based their objections on two misconceptions of trust. They have believed that the implementation of care robots would impoverish the ideals and values of care. However, this need not be the case once we understand trust in a more complex manner. On the one hand, trust relationships do not depend only on the type of entity being trusted, on the properties it possesses. The properties determine the maximum degree to which an entity is trustworthy, but we can trust it in other, less exhaustive ways. On the other hand, caring environments incorporate personal trust relationships as a fundamental, but not sufficient, condition. Other forms of rational and thin motivational trust may be required to complement personal trust.

Third, we have proposed, given the foregoing argumentation, a new model of trust in care settings: sociotechnical trust in teams of humans and robots. This new conception of trust is articulated on four types of relationships that are differentiated by virtue of the type of entity trusted and the

[10] Our tendency to anthropomorphize can lead us to place personal trust in robots without these entities being trustworthy enough. This tendency is exacerbated, as we have shown, when robots acquire human appearances and gestures. Our model of sociotechnical trust also intends to provide an answer to these threats: to ensure that the system works properly and that trusting personal relationships only take place between humans, there must be good supervision of the relationships between caregivers and SCR, and the latter must be designed with the appearance and safeguards to prevent deception that would lead to personal trust in robots. We thank an anonymous reviewer for encouraging us to clarify this point.

type of interaction. We have shown, finally, how this new model addresses the specific problems of the ethics of SCR: lack of human contact, objectification and deception. While robots cannot provide personal and comprehensive care, we have shown how their implementation is ethically acceptable if conducted within the framework of an appropriate set of trustworthy relationships.

The results of this paper should be supplemented with sound empirical research to show in what sense trustworthy relationships with robots are practically achievable. Moreover, it can serve as an ethical coordinate on which empirical studies can be designed to monitor the implementation of robots in care settings from the perspective of trust. This monitoring would involve taking into account the context in each case, the needs of the users and the consequences derived from the use of these technologies.

Acknowledgements We are very grateful to Francisco Lara and Jon Rueda for their insightful and valuable comments. We would also like to thank the attendees to the Workshop "Ethical and legal aspects of digitized medicine", held in the University of León, and the PDIFF Seminar, held in the University of Granada.

Funding Funding for open access publishing: Universidad de Granada/CBUA. Funding for open access publishing: Universidad de Granada/CBUA. JLA has benefited from a research contract FPU (FPU22/03178), funded by the Spanish Ministry of Education. BL received funding from the Spanish Ministry of Universities under grant FPU19/06017 and from US-Spain Fulbright under a 2023–2024 Predoctoral Research grant. This article is also part of the research project AutAI, funded by the Spanish National Research Agency (PID2022-137953OB-I00).

Data availability There is no data associated with this research.

Declarations

Conflict of interest All authors declare that they have no conflict of interest.

References

Akalin A (2007) Hired as a caregiver, demanded as a housewife: becoming a migrant domestic worker in Turkey. Eur J Women's Stud 14:209–225. https://doi.org/10.1177/1350506807079011

Arkin R (2009) Governing lethal behavior in autonomous robots. Chapman and Hall/CRC, New York

Ausín T, Liedo B, López Castro D (2023) Robótica asistencial: una reflexión ética y filosófica. In: Andreu Martínez MB, Espinosa de los Monteros Rodríguez A (eds) Tecnología para la salud: una visión humanista desde el bioderecho. Madrid, Plaza y Valdés, pp 63-82.

Baier A (1986) Trust and antitrust. Ethics 96:231–260

Boada JP, Maestre BR, Genís CT (2021) The ethical issues of social assistive robotics: a critical literature review. Technol Soc 67:101726. https://doi.org/10.1016/j.techsoc.2021.101726

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

Bravo-Benítez J, Pérez-Marfil MN, Román-Alegre B, Cruz-Quintana F (2019) Grief experiences in family caregivers of children with autism spectrum disorder (ASD). Int J Environ Res Public Health 16:4821. https://doi.org/10.3390/ijerph16234821

Breazeal CL (2004) Designing sociable robots. The MIT Press, Massachusetts

Bryson JJ (2012) A role for consciousness in action selection. Int J Mach Conscious 04:471–482. https://doi.org/10.1142/S1793843012400276

Buechner J, Simon J, Tavani HT (2013) Re-thinking trust and trustworthiness in digital environments. In: 11th computer ethics: philosophical enquiry (CEPE 2013). pp 1–15

Butler J (2004) Precarious life: the powers of mourning and violence. Verso book, London

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

Coeckelbergh M (2012) Can we trust robots? Ethics Inf Technol 14:53–60. https://doi.org/10.1007/s10676-011-9279-1

Coeckelbergh M (2016) Care robots and the future of Ict-mediated elderly care: a response to doom scenarios. AI Soc 31:455–462. https://doi.org/10.1007/s00146-015-0626-3

Coghlan S (2022) Robots and the possibility of humanistic care. Int J Soc Robot 14:2095–2108. https://doi.org/10.1007/s12369-021-00804-7

Coleman JS (1990) Foundations of social theory. Belknap Press of Harvard University Press, Boston

Cooper R, Foster M (1971) Sociotechnical systems. Am Psychol 26:467–474. https://doi.org/10.1037/h0031539

Copé MLR (2021) Empleo digno y de calidad: ¿utopía en el trabajo doméstico. Lex Soc: Revista De Derechos Sociales 11:594–627. https://doi.org/10.46661/lexsocial.5977

Danaher J (2019) The philosophical case for robot friendship. J Posthuman Stud 3:5–24. https://doi.org/10.5325/jpoststud.3.1.0005

Danaher J, McArthur N (2017) Robot sex: social and ethical implications. MIT Press, Cambridge

de Laat PB (2016) Trusting the (ro)botic other: by assumption? SIGCAS Comput Soc 45:255–260. https://doi.org/10.1145/2874239.2874275

de Kervenoael R, Hasan R, Schwob A, Goh E (2019) Leveraging human-robot interaction in hospitality services: incorporating the role of perceived value, empathy, and information sharing into visitors' intentions to use social robots. Tour Manage 78:1–15. https://doi.org/10.1016/j.tourman.2019.104042

Dodds S (2013) Dependence, care, and vulnerability. In: Mackenzie C, Rogers W, Dodds S (eds) Vulnerability: new essays in ethics and feminist philosophy. Oxford University Press, USA, pp 181–203

Dreyfus HL (1992) What computers still can't do. A critique of artificial reason. The MIT Press, Cambridge

Dreyfus HL (2007) Why Heideggerian AI failed and how fixing it would require making it more Heideggerian. Philos Psychol 20:247–268. https://doi.org/10.1080/09515080701239510

Elder A (2014) Excellent online friendships: an aristotelian defense of social media. Ethics Inf Technol 16:287–297. https://doi.org/10.1007/s10676-014-9354-5

Farina M, Zhdanov P, Karimov A et al (2022) AI and society: a virtue ethics approach. AI Soc. https://doi.org/10.1007/s00146-022-01545-5

Faulkner P (2007) On telling and trusting. Mind 116:875–902

Fineman M (2008) The vulnerable subject: anchoring equality in the human condition. Yale J Law Fem 20:8–40

Floridi L, Sanders JW (2004) On the morality of artificial agents. Mind Mach 14:349–379. https://doi.org/10.1023/B:MIND.0000035461.63578.9d

Frank L, Nyholm S (2017) Robot sex and consent: is consent to sex between a robot and a human conceivable, possible, and desirable? Artif Intell Law 25:305–323. https://doi.org/10.1007/s10506-017-9212-y

Gordon J-S, Nyholm S (2021) Kantianism and the problem of child sex robots. J Appl Philos 39:132–147. https://doi.org/10.1111/japp.12543

Grodzinsky F, Miller K, Wolf MJ (2020) Trust in artificial agents. The Routledge handbook of trust and philosophy. Routledge, Milton Park

Hämäläinen A (2020) Responses to vulnerability: care ethics and the technologisation of eldercare. Int J Care Caring 4:167–182. https://doi.org/10.1332/239788220X15833753877589

Hardin R (2006) Trust. Polity. Cambridge University, Cambridge

Hawley K (2012) Trust. A very short introduction. Oxford University Press, Oxford

Held V (1990) Feminist transformations of moral theory. Philos Phenomenol Res 50:321–344. https://doi.org/10.2307/2108046

Himma KE (2009) Artificial agency, consciousness, and the criteria for moral agency: what properties must an artificial agent have to be a moral agent? Ethics Inf Technol 11:19–29. https://doi.org/10.1007/s10676-008-9167-5

Hoff KA, Bashir M (2015) Trust in automation: integrating empirical evidence on factors that influence trust. Hum Factors 57:407–434. https://doi.org/10.1177/0018720814547570

Hollis M (1998) Trust within reason. Cambridge University Press, Cambridge

Kittay EF (2019) Love's labor: essays on women, equality and dependency, 2nd edn. Routledge, New York

Kühnlenz B, Sosnowski S, Mayer C, et al (2011) Improving aspects of empathy and subjective performance for HRI through mirroring facial expressions**.** In: 20th IEEE international conference on robot & human interactive communication (RO-MAN). pp 350–356

Latour B (1992) Where are the missing masses? The sociology of a few mundane artifacts. In: Bijker WE, Law J (eds) Shaping technology/building society: studies in sociotechnical change. MIT Press, Cambridge, pp 225–258

Laukyte M (2017) Artificial agents among us: should we recognize them as agents proper? Ethics Inf Technol 19:1–17. https://doi.org/10.1007/s10676-016-9411-3

Levy D (2008) Love and sex with robots: the evolution of human-robot relationships. Harper Perennial, New York

Liedo B (2024) Navigating autonomy, privacy, and ageism in robot home care with aged users: A preliminary analysis of ROB-IN. Bioethics. https://doi.org/10.1111/bioe.13340

Liedo B, Van Grunsven J, Marin L (2024) Emotional Labor and the Problem of Exploitation in Roboticized Care Practices: Enriching the Framework of Care Centred Value Sensitive Design. Science and Engineering Ethics 30:42. https://doi.org/10.1007/s11948-024-00511-2

Llácer A, Zunzunegui MV, del Amo J et al (2007) The contribution of a gender perspective to the understanding of migrants' health. J Epidemiol Commun Health 61:4–10. https://doi.org/10.1136/jech.2007.061770

Llorca Albareda J (2023) El estatus moral de las entidades de inteligencia artificial. Disputatio 12:241-249. https://doi.org/10.5281/zenodo.8140967

Llorca Albareda J (2024) Anthropological crisis or crisis in moral status: a philosophy of technology approach to the moral consideration of artificial intelligence. Philosophy & Technology 37:12. https://doi.org/10.1007/s13347-023-00682-z

Llorca Albareda J, García P, Lara F (2024) The moral status of AI entities. In: Lara F, Deckers J (eds) Ethics of artificial intelligence. Cham, Springer Nature Switzerland, pp 59-83. https://doi.org/10.1007/978-3-031-48135-2_4

Llorca Albareda J (2025) Uncovering the gap: challenging the agential nature of AI responsibility problems. AI and Ethics: 1–14. https://doi.org/10.1007/s43681-025-00685-w

Marti P (2010) Robot Companions: towards a new concept of friendship? Interact Stud: Soc Behav Commun Biol Artif Syst 11:220–226. https://doi.org/10.1075/is.11.2.07mar

Martínez-López MV, Díaz-Cobacho G, Astobiza AM, Rodríguez López B (2024) Exploring the Ethics of Interaction with Care Robots. In: Lara F, Deckers J (eds) Ethics of artificial intelligence. Cham, Springer Nature Switzerland, pp 149-167. https://doi.org/10.1007/978-3-031-48135-2_8

Martínez-López MV, McLaughlin L, Molina-Pérez A, Pabisiak K, Primc N, Randhawa G, Rodríguez-Arias D, Suárez J, Wölhke S, Delgado J (2023) Mapping trust relationships in organ donation and transplantation: a conceptual model. BMC Medical Ethics 24:93. https://doi.org/10.1186/s12910-023-00965-2

Matthias A (2004) The responsibility gap: ascribing responsibility for the actions of learning automata. Ethics Inf Technol 6:175–183. https://doi.org/10.1007/s10676-004-3422-1

Mol A (2008) The logic of care: health and the problem of patient choice. Routledge, London

Moor JH (2006) The nature, importance, and difficulty of machine ethics. IEEE Intell Syst 21:18–21. https://doi.org/10.1109/MIS.2006.80

Nadeau JE (2006) Only androids can be ethical. In: Ford KM, Glymour HP (eds) Thinking about android epistemology. AAAI Press, Washington, pp 241–248

Nahmias E (2016) Free will as a psychological accomplishment. In: Schmidtz D, Pavel C (eds) The Oxford handbook of freedom. Oxford University Press, Oxford

Nickel PJ, Franssen M, Kroes P (2010) Can we make sense of the notion of trustworthy technology? Know Techn Pol 23:429–444. https://doi.org/10.1007/s12130-010-9124-6

Nyholm S (2018) Attributing agency to automated systems: reflections on human-robot collaborations and responsibility-loci. Sci Eng Ethics 24:1201–1219. https://doi.org/10.1007/s11948-017-9943-x

Nyholm S (2020) Humans and robots: ethics, agency, and anthropomorphism. Rowman Littlefield, Washington

Nyholm S, Frank L (2019) It loves me, it loves me not: is it morally problematic to design sex robots that appear to "love" their owners? Techné: Res Philos Technol 23:402–424. https://doi.org/10.5840/techne2019122110

Nyholm S, Smids J (2020) Can a robot be a good colleague? Sci Eng Ethics 26:2169–2188. https://doi.org/10.1007/s11948-019-00172-6

Poulsen A, Burmeister O, Kreps D (2018a) The ethics of inherent trust in care robots for the elderly: 13th IFIP TC 9 international conference on human choice and computers, HCC13 2018, held at the 24th IFIP world computer congress, WCC 2018, Poznan, Poland, September 19–21, 2018, Proceedings. pp 314–328

Poulsen A, Burmeister OK, Tien D (2018b) Care robot transparency isn't enough for trust. In: 2018 IEEE region ten symposium (Tensymp). pp 293–297

Prescott TJ, Robillard JM (2021) Are friends electric? The benefits and risks of human-robot relationships. iScience 24:101993. https://doi.org/10.1016/j.isci.2020.101993

Richardson K (2016) The asymmetrical "relationship": parallels between prostitution and the development of sex robots. SIGCAS Comput Soc 45:290–293. https://doi.org/10.1145/2874239.2874281

Rodríguez Díaz MT, Cruz-Quintana F, Pérez-Marfil MN (2014a) Dependencia funcional y bienestar en personas mayores institucionalizadas. Index De Enfermería 23:36–40. https://doi.org/10.4321/S1132-12962014000100008

Rodríguez Díaz MT, Pérez-Marfil MN, Cruz-Quintana F (2014b) Propuesta de plan estandarizado de cuidados para prevenir la dependencia y la fragilidad. Gerokomos 25:137–143. https://doi.org/10.4321/S1134-928X2014000400002

Ryland H (2021) It's friendship, Jim, but not as we know it: a degrees-of-friendship view of human-robot friendships. Mind Mach 31:377–393. https://doi.org/10.1007/s11023-021-09560-z

Sætra HS (2021) Social robot deception and the culture of trust. Paladyn J Behav Robot 12:276–286. https://doi.org/10.1515/pjbr-2021-0021

Sætra HS (2022) Loving robots changing love: towards a practical deficiency-love. J Future Robot Life 3:109–127. https://doi.org/10.3233/FRL-200023

Scheutz M (2011) The inherent dangers of unidirectional emotional bonds between humans and social robots. In: Lin P, Abney K, Bekey GA (eds) Robot ethics: the ethical and social implications of robotics. MIT Press, Cambridge, pp 205–221

Segers S (2022) Robot technology for the elderly and the value of veracity: disruptive technology or reinvigorating entrenched principles? Sci Eng Ethics. https://doi.org/10.1007/s11948-022-00420-2

Sharkey A, Sharkey N (2011) Children, the elderly, and interactive robots. IEEE Robot Autom Mag 18:32–38. https://doi.org/10.1109/MRA.2010.940151

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

Sharkey A, Sharkey N (2021) We need to talk about deception in social robotics! Ethics Inf Technol 23:309–316. https://doi.org/10.1007/s10676-020-09573-9

Sica A, Sætra HS (2023) In Technology we trust! but should we? In: Human-computer interaction: thematic area, HCI 2023, held as part of the 25th HCI international conference, HCII 2023, Copenhagen, Denmark, July 23–28, 2023, Proceedings, Part II. Springer-Verlag, Berlin, Heidelberg, pp 293–317

Song Y (2020) Building a 'deeper' trust: mapping the facial anthropomorphic trustworthiness in social robot design through multidisciplinary approaches. Des J 23:639–649. https://doi.org/10.1080/14606925.2020.1766871

Sorell T, Draper H (2014) Robot carers, ethics, and older people. Ethics Inf Technol 16:183–195. https://doi.org/10.1007/s10676-014-9344-7

Sparrow R (2007) Killer robots. J Appl Philos 24:62–77. https://doi.org/10.1111/j.1468-5930.2007.00346.x

Sparrow R (2016) Robots in aged care: a dystopian future? AI & Soc 31:445–454. https://doi.org/10.1007/s00146-015-0625-4

Sparrow R, Sparrow L (2006) In the hands of machines? The future of aged care. Mind Mach 16:141–161. https://doi.org/10.1007/s11023-006-9030-6

Taddeo M (2009) Defining trust and E-trust: from old theories to new problems. IJTHI 5:23–35. https://doi.org/10.4018/jthi.2009040102

Tallant J, Donati D (2020) Trust: from the philosophical to the commercial. Philos Manag 19:3–19. https://doi.org/10.1007/s40926-019-00107-y

Tavani HT (2015) Levels of trust in the context of machine ethics. Philos Technol 28:75–90. https://doi.org/10.1007/s13347-014-0165-8

Tronto J (2015) Democratic caring and global care responsibilities. Ethics of care. Policy Press, Bristol, pp 21–30

Turkle S (2011) Alone together: why we expect more from technology and less from each other. Basic Books, New York

Vallès-Peris N, Domènech M (2023) Caring in the in-between: a proposal to introduce responsible AI and robotics to healthcare. AI & Soc 38:1685–1695. https://doi.org/10.1007/s00146-021-01330-w

Vallor S (2011) Carebots and caregivers: sustaining the ethical ideal of care in the twenty-first century. Philos Technol 24:251–268. https://doi.org/10.1007/s13347-011-0015-x

van Wynsberghe A (2015) Healthcare robots: ethics, design and implementation. Routledge, London

van Wynsberghe A (2016) Service robots, care ethics, and design. Ethics Inf Technol 18:311–321. https://doi.org/10.1007/s10676-016-9409-x

van Wynsberghe A (2022) Social robots and the risks to reciprocity. AI & Soc 37:479–485. https://doi.org/10.1007/s00146-021-01207-y

van Wynsberghe A, Li S (2019) A paradigm shift for robot ethics: from HRI to human–robot–system interaction (HRSI). Medicolegal Bioethics 9:11–21. https://doi.org/10.2147/MB.S160348

Véliz C (2021) Privacy is power: why and how you should take back control of your data. Melville House, New York

Walker MU (2006) Moral repair: reconstructing moral relations after wrongdoing. Cambridge University Press, Cambridge

Wallach W, Allen C (2009) Moral machines: teaching robots right from wrong. Oxford University Press, Oxford

Whittingham R (2003) The blame machine: why human error causes accidents. Routledge, London

Yew GCK (2021) Trust in and ethical design of carebots: the case for ethics of care. Int J of Soc Robotics 13:629–645. https://doi.org/10.1007/s12369-020-00653-w

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