Ethik in der Medizin (2023) 35:221–246
https://doi.org/10.1007/s00481-023-00760-y

ORIGINAL ARTICLE

Can robots be trustworthy?

Reflections about responsive robots and trust as a human capability

Ines Schröder · Oliver Müller · Helena Scholl · Shelly Levy-Tzedek · Philipp Kellmeyer

Received: 3 October 2022 / Accepted: 8 February 2023 / Published online: 24 April 2023

Abstract

Definition of the problem This article critically addresses the conceptualization of trust in the ethical discussion on artificial intelligence (AI) in the specific context of social robots in care. First, we attempt to define in which respect we can speak of 'social' robots and how their 'social affordances' affect the human propensity to trust in human–robot interaction. Against this background, we examine the use of the concept of 'trust' and 'trustworthiness' with respect to the guidelines and recommendations of the High-Level Expert Group on AI of the European Union.

✉ Ines Schröder
Philosophisches Seminar, Albert-Ludwigs-Universität Freiburg, Platz der Universität 3, 79098 Freiburg, Germany
E-Mail: ines.schroeder@philosophie.uni-freiburg.de

Prof. Dr. Oliver Müller
Professur für Philosophie der Gegenwart und Technik, Albert-Ludwigs-Universität Freiburg, Freiburg im Breisgau, Germany

Helena Scholl · Dr. med. Philipp Kellmeyer
Human-Technology Interaction Lab, Klinik für Neurochirurgie, Universitätsklinikum Freiburg, Freiburg im Breisgau, Germany

Prof. Dr. Shelly Levy-Tzedek
Recanati School for Community Health Professions, Department of Physical Therapy, Faculty of Health Sciences, Ben-Gurion University of the Negev, Be'er Scheva, Israel

Prof. Dr. Shelly Levy-Tzedek · Dr. med. Philipp Kellmeyer
Freiburg Institute for Advanced Studies (FRIAS), University of Freiburg, Freiburg im Breisgau, Germany

Dr. med. Philipp Kellmeyer
Institut für Biomedizinische Ethik und Geschichte der Medizin, Universität Zürich, Zürich, Switzerland

Medizinische Fakultät, Albert-Ludwigs-Universität Freiburg, Freiburg im Breisgau, Germany

Arguments Trust is analyzed as a multidimensional concept and phenomenon that must be primarily understood as departing from trusting as a human functioning and capability. To trust is an essential part of the human basic capability to form relations with others. We further want to discuss the concept of *responsivity* which has been established in phenomenological research as a foundational structure of the relation between the self and the other. We argue that trust and trusting as a capability is fundamentally *responsive* and needs responsive others to be realized. An understanding of *responsivity* is thus crucial to conceptualize trusting in the ethical framework of human flourishing. We apply a phenomenological–anthropological analysis to explore the link between certain qualities of social robots that construct responsiveness and thereby simulate responsivity and the human propensity to trust.
Conclusion Against this background, we want to critically ask whether the concept of trustworthiness in social human–robot interaction could be misguided because of the limited ethical demands that the constructed responsiveness of social robots is able to answer to.

Keywords Social robots · Care · Responsivity · Trust · Capabilities

Können Roboter vertrauenswürdig sein?

Reflexionen über responsive Roboter und Vertrauen als menschliche Fähigkeit

Zusammenfassung

Definition des Problems Dieser Artikel setzt sich kritisch mit dem Begriff des Vertrauens im ethischen Diskurs um künstliche Intelligenz speziell im Kontext von sozialen Robotern in der Pflege auseinander. Zunächst versuchen wir zu bestimmen, in welcher Hinsicht wir überhaupt von *sozialen* Robotern sprechen können und wie sich die „social affordances“ von Pflegerobotern auf die menschliche Vertrauensbereitschaft in der Mensch-Roboter-Interaktion auswirken. Vor diesem Hintergrund wird im Folgenden die Verwendung der Begriffe von Vertrauen und Vertrauenswürdigkeit näher untersucht, insbesondere im Hinblick auf die Empfehlungen der „High Level Expert Group on AI“ der EU.
Argumente Vertrauen wird dabei als mehrdimensionales Phänomen analysiert, das vorranging ausgehend von der menschlichen Fähigkeit („capability“) des Vertrauen-Könnens verstanden werden muss. Vertrauen-Können ist ein essenzieller Teil der menschlichen Grundfähigkeit, Beziehungen mit anderen einzugehen und aufzubauen. Wir möchten in dieser Hinsicht weiter das phänomenologische Konzept der Responsivität diskutieren, das in phänomenologischer Forschung als grundlegende Struktur der Beziehung zwischen dem Selbst und dem Anderen entwickelt wurde. Wir argumentieren, dass Vertrauen und Vertrauen-Können als Fähigkeit aufgrund seiner relationalen Aspekte grundsätzlich *responsiv* ist und responsive Andere benötigt, um als Fähigkeit aktualisiert zu werden. Ein Verständnis von Responsivität ist daher zentral, um das Vertrauen-Können innerhalb eines Ansatzes des guten Lebens begrifflich fassen zu können. Wir möchten mithilfe einer phänomenologisch-anthropologischen Analyse die Verbindung zwischen gewissen Qualitäten sozialer Roboter, die Responsivität konstruieren und simulieren, und der menschlichen Vertrauensbereitschaft untersuchen.

Konklusionen Vor diesem Hintergrund wollen wir kritisch fragen, ob das Konzept der Vertrauenswürdigkeit in der Mensch-Roboter-Interaktion problematisch sein könnte, weil technisch konstruierte Responsivität nur begrenzt den ethischen Ansprüchen menschlicher Responsivität genügt.

Schlüsselwörter Soziale Roboter · Pflege · Responsivität · Vertrauen · Fähigkeiten

Introduction

Nations worldwide consider robots part of the package in dealing with societal challenges of aging societies such as increased demands for care and health care provision. As technology advances, pilot projects have emerged in many countries[1] that test and embed robots in care and healthcare environments. This is in line with the agenda of the German Ministry for Education and Research (BMBF) to pursue assistive technology as a possible solution in the care of older adults (BMBF 2018). Integrating robots into care routines, however, defines a substantial change in an understanding of care as a human to human, social engagement in which ethical sensitive actions and phenomena occur. This leads to an advanced discussion of core social and relational concepts that try to capture these phenomena. One example is the concept of trust, which has been discussed as essential for the success of therapeutic interventions (Thom et al. 1999) and human–robot interaction (Langer et al. 2019; Koren et al. 2022). In this paper, we wish to address the ongoing need for ethical and normative reflection on the 'question of trust' (Kellmeyer et al. 2018) on the basis of current but also possible future (e.g., complemented by neurotechnological and artificial intelligence [AI][2]) applications of care robots for the older population. In asking if robots could be trustworthy, we highlight and criticize the philosophically insufficient conceptualization of trust and trustworthiness as a measurable design feature.

Trust and trustworthiness should not be reduced to mere means for facilitating the acceptance of socially assistive robots (SARs). Thus, the main aims of this paper are to describe trust as an essential capability of living beings to generate relations with others and to analyze the ethical implications of robots that can simulate core dimensions of this capability to engender trust in humans.

We argue that care situations in particular show this relational condition for trust in the necessity for *responsive others* as indispensable for care. Someone who does not respond to any of the trustor's needs could not be considered caring, even if that person is a nurse. We want to take this observation further and argue that the practice of *responding* is not only a prerequisite for care but also for trust to occur at all, which explains why human beings seem to be more incited to trust someone

[1] For example: ACCRA (Agile Co-Creation of Robots for Ageing); CARESSES (Italy) (caresses-robot.org/en/); ReThiCare (Weimar), MARIO Project (Ireland, Italy) (http://www.mario-project.eu/portal/. Accessed: 23 June 2022).

[2] For background on the convergence of AI and neurotechnology, e.g., in brain–computer interfaces for operating robots, see Kellmeyer (2019a).

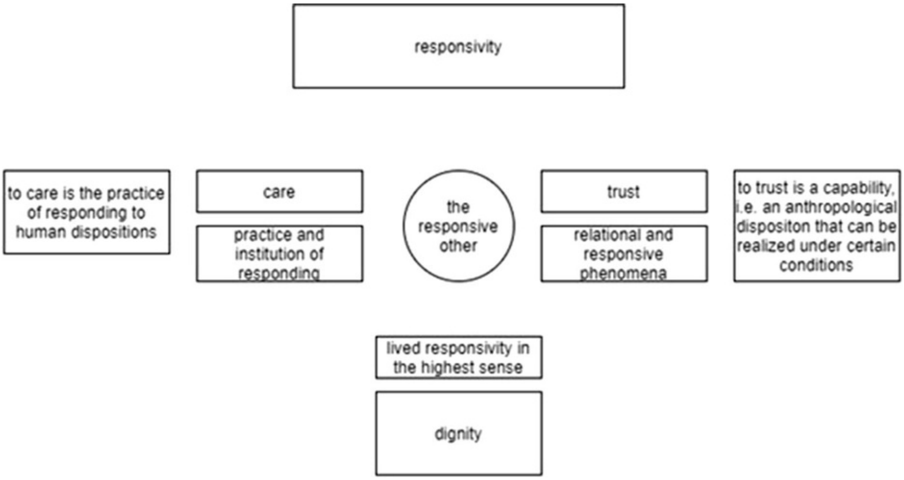
responsivity
to care is the practice of responding to human dispositions
care
practice and institution of responding
the responsive other
trust
relational and responsive phenomena
to trust is a capability, i.e. an anthropological dispositon that can be realized under certain conditions
lived responsivity in the highest sense
dignity

Fig. 1 Illustration of the relationship between care, responsivity, dignity, and trust

they perceive as caring, i.e., *responding* in a caring way. To foster the argument, we will proceed in four steps:

1. The analysis of trust and trusting in a caring context elucidates the bidirectionality of the concepts of care, trust, and responsivity (Fig. 1) from a phenomenological–anthropological and situational perspective. Trust, in our definition, is a responsive phenomenon that emerges from the human capability, i.e., the anthropological disposition and realized functioning to trust under certain conditions. Responsivity is structurally essential for trust and for care, while trust and care interrelate on a qualitative scale.
2. The phenomenological–anthropological analysis addresses the structural phenomenological level of responsivity: We start from the premises that all human responsive phenomena are grounded in responsivity, which serves at the phenomenological level as the basic meaningful relational engagement with 'others'. Responsivity has been studied in phenomenological research as a foundational structure that is constitutive for responding to the (ethical) demand of the other (Waldenfels 1987, 1994, 1997). Without responding, care would not be different from just a 'procedure' or 'following a protocol'. In that, care is a relational engagement defined by responding. In our adaptation of the Capability Approach by Martha C. Nussbaum (2000, 2011), responding in a caring way to the human dispositions and capabilities means supporting human flourishing. In the following, we will make the case that without responsivity, relations would not account for the possibility of trust and trustworthiness.
3. By introducing the phenomenological structure of responsivity to the ethical debate, we want to propose a different approach for evaluating trustworthiness. The analysis of the phenomenological–anthropological conditions for trust leads to the question in which case a human being might be incited to trust but *should not* since the robot does not (and cannot) respond to inherent demands that human

responsivity allows for with respect to the realization of human capabilities. As we understand trusting to be a scalable engagement that relates to the scope *and quality* of the responsive interaction, we argue that good care means to respond to the existential complexity of the human being (e.g., its dignity, vulnerability, lived body experience, situatedness, and capabilities). The trustworthy other is therefore someone who lives up to these complex ethical demands of responsivity.

4. Based on a deeper understanding of the alterity-relation in human–technology interaction—that robots could be 'others' to which we can relate in complex ways (Müller 2022)—, we want to examine the capability to trust as the ability of living beings to generate relations with diverse forms of responsive others. Robotic systems that are constructed to react to this human disposition, e.g., through AI-complemented social functions, could thus be part of a responsive trust situation and relation. In looking at instances of responsive behavior in humans and robots in order to create a preliminary hypothesis on possible similarities and differences, we still see the need to limit the scope of this paper to the theoretical analysis that should be expanded by in-depth empirical research. The phenomenological–anthropological approach, however, allows for the account of the first-person subjective experience to be integrated into the theoretical work. We will argue that robots should not be designed as too trust-inciting because the constructed responsive qualities do not suffice for a reciprocal and dignified relationship since constructed responsiveness in robots is not accompanied by a sense of responsibility necessary to meet the existential complexity of the human being.

In the following, we will first ask the question what it means for a 'social' robot to be social and offer a short survey on existing and future (AI-complemented) applications of SARs in care. Further, we want to take a closer look at the multifactored embeddedness of the phenomenon of trust in care situations and on how the qualities and features of the robot interplay with the human responsive capability for trusting. After discussing concepts of trust and trustworthiness in philosophy and research as well as existing guidelines for technology development along the example of AI, we will conduct a preliminary analysis of the responsive qualities of SARs and how they are perceived as trust-inducing social affordances. In the last section, we will discuss the ethical and normative dimension of the potential for deception when designing 'responsive machinic-others' for human users.

Social robots and the need for trust

Social robots are considered 'social' because they are designed to interact with people "in a natural, interpersonal manner—often to achieve social–emotional goals" (Breazeal et al. 2016). Part of this interaction is 'natural' communication "using both verbal and nonverbal signs", and engaging as a 'partner', "not only on a cognitive level, but on an emotional level as well" (Breazeal et al. 2016, p. 1935). This leads to the goal of designing so-called "empathetic" robots (Misselhorn 2021; Henschel et al. 2021) which are able to detect human emotions and reply with a programmed emotional reaction-schema (e.g., AI robot systems like MABU).

In the context of care, the assistive functionality of robots has priority. Robots may be categorized as "contact assistive robots" (Feil-Seifer and Mataric 2005) like robot arms, which aid and support a human user by close contact to the body, and 'non-contact assistive robots' for which their communicative abilities can become more defining. However, they are not mutually exclusive as the pet robot PARO demonstrates which is contact assistive but communicates nonverbally. From a descriptive perspective and depending on their abilities, we can speak of different degrees of social features that are used to facilitate the general assistive purpose.

Yet, categorizing social robots remains difficult: As PARO also demonstrates, the label 'socially assistive' does not only imply the means (having social features like communication) but also the purpose (assisting with social needs [e.g., reducing apathy and enhancing interaction of the patient]). Purpose and means tend to blend. SARs provide special assistance through these social engagements (Feil-Seifer and Mataric 2005), which can be individualized and adapted to a patient's needs.

Why is trust relevant for interaction with social robots in care?

The discourse about patients' needs in care highlights the role of trust for successful therapeutic and care interactions (Pellegrini 2017; Greene and Ramos 2021; Dinç and Gastmans 2013, 2011; Peter and Morgan 2001). Trust in general is regarded as highly relevant for successful patient–physician relationships (Montgomery et al. 2020; Ridd et al. 2009; Thom et al. 1999). The quality of the patient–physician relationship and the ability of the physician to ameliorate patient cooperation refer to trust as part of the bond in the therapeutic alliance or working alliance (Müller et al. 2014; Bordin 1979, 1974), as well as to psychological findings about the human capacity to form trusting relationships as part of early child development (Koepke and Denissen 2012; Erikson 1950).

In the case of assistive social robots, the robot is being integrated into the therapeutic or care interaction. The understanding of socially assistive robots (Lewis et al. 2018) in care starts from the premise that the robot, too, needs to induce trust in order to successfully fulfill the role of a social interaction partner in care. To explain the effects of the robotic interventions, the phenomenon of trust is considered to be an important factor as it emerges in and from social interactions.

The functionality of social robots thus adds a certain urgency to the question of trust. Current models differ from the historically prototypical 'industrial robot' in that they are often described as 'intelligent'.[3] The metaphor refers to an enhanced computational functionality made possible by the implementation of machine learning, such as artificial neural networks for deep learning, sensory endowment, and data 'memory'.[4]

[3] The enhanced computational functionality of robots by AI systems makes trust an issue in the first place, which is more relevant for interaction with AI systems as compared with a non-intelligent technical artifact like an electrical doorbell.

[4] We speak here of metaphor because human qualities and abilities like 'intelligent' behavior are transferred to the machine. This metaphorical understanding stems from the 'brain–computer analogy', an influential paradigm in cognitive science and neuroscience. It can be traced back to model building in cybernetics, i.e., early computer science in the 1950s. See Bringsjord and Govindarajulu (2018).

This increasingly automatized and adaptive functionality makes it more likely that the robots are used *without a human supervisor directly present*. This may create situations of exposedness to and in some cases dependence on the robot. This is especially critical in care, where the robot assists in actions and tasks that the patient might not be able to perform alone anymore. At the same time, the robot might create the situational atmosphere of 'caring' by asking about well-being and needs.[5]

When used in rehabilitation, robots are also programmed to prompt and monitor goal-oriented co-operation. One open empirical question is whether and to what extent humans who interact with these robots form specific beliefs about the 'good' intentions (and their truthful realization) of actions that are announced by the robot, or whether they are operating from a baseline level of trust without speculating (implicitly or explicitly) about the robots' intentions. These aspects underlie the specific relevance of trust in human–social robots interaction in that the machines are implemented and used to create a situation and interactional setting in which the human propensity to trust is activated out of vulnerability, dependency, and susceptibility to constructed social cues (Baier 1986; Ryan 2020; Hoff and Bashir 2015), which is particularly different to other human–AI interactions in which the social dimension is not essential to the interaction (Duran and Jongsma 2021).

Current research, however, neglects this social situational embeddedness and its phenomenological–anthropological dimensions. We agree with Gille et al. (2020, *cp*. 1) that trust is "relational, highly complex", "situational and difficult to develop as a general concept", but we link the problem to generalize to the tendency of epistemological approaches to reduce trust to its cognitive aspects. Assessing the probability for justified trust (Starke et al. 2021) does not consider the embodied and subconscious processes involved in trust. Updating or not-updating a belief on the trustworthiness of an actor or economizing on the monitoring (Ferrario et al. 2021) would reduce the cognitive aspects of the trust relationship. Rational trust theories tend to understand interaction as a *transaction* (exchanging control or complexity for trust), turning trust into a form of capital. We want to point out that the social dimension in care interactions exceeds the transactional paradigm. This will be demonstrated by a phenomenological–anthropological analysis and qualitative interpretation of the interrelation between the structure of responsivity that underlies human being-in and being-to-the-world and the capability to trust in respect to the design of the robot and its social functionality. As trust is also used in other emerging contexts of human–technology interaction, a brief examination

[5] The following examples show how intelligent systems enhance the appearance of autonomy and partnership: Robots like MABU can 'read' the emotions of patients by facial recognition. The device collects and stores data and reminds patients to take their medicine. The program CARESSES, which can be implemented in the robot PEPPER, includes cultural knowledge that can be used to start a conversation. It 'learns' the habits and practices of its users by 'remembering' information that it has been told, thus, adapting and individualizing its functions. Robots that include machine learning algorithms tend to optimize their functionality by using the collected data to 'adapt' to their environment. Social robots also give information about their own 'state': They use voices to talk, hand and arm gestures and facial expressions to simulate human behavior.

of conceptualizations, here along the example of the EU guidelines for AI, will be helpful.

Guidelines for trust and trustworthiness with respect to AI

According to the EU HLEG guidelines on AI (HLEG AI 2019), trustworthy AI should be: "(1) lawful, complying with all applicable laws and regulations, (2) ethical, ensuring adherence to ethical principles and values, (3) robust, both from a technical perspective and social perspective" (HLEG AI 2019, p. 2). The guidelines themselves do not offer any definition or operational description of the concept of trust or trustworthiness that goes beyond the (not binding) regulatory aim.[6] Instead, they add more information on key requirements that an AI system should meet "in order to be deemed trustworthy": "AI systems should empower human beings", they need to be resilient, safe, secure, human-monitored, accurate, reliable, and reproductive. They should ensure "full respect for privacy and data protection", and they should be transparent, accessible, fair, sustainable, environmentally friendly, and auditable.[7]

The trustworthy AI-enhanced robot, so to speak, satisfies these attributes, which are partly *technical* (robust, reliable, accessible, sustainable, safe, secure) and partly *procedural* (transparent, respectful of laws, regulations, privacy and data protection, as well as respectful of ethical principles and values). The first set of attributes, in our view, refers to the stability of the operation and is better described as a set of features that signal the reliability of the robot since these are expectations about the technical construction. The robot should run smoothly and without unexpected disturbances. It is clear that these guidelines aim at AI developers. Hence, the responsibility for the promised functionality lies with them. They should design AI systems in compliance with these properties, and at the same time, make AI systems comply with these properties. But why do these attributes render an AI system automatically trustworthy?

Structurally, the argument for appealing to the semantics of trust and trustworthiness usually derives from the idea that a 'trust situation' involves risk. Describing something as safe, secure, robust, and reliable, thus, includes the claim to have reduced risk to the minimum by applying control. Technically speaking, the chance of system failure can never be eliminated. In this case, trustworthiness seems to be a declaration of limited liability: The engineers have done everything in their power but there is no total security. To be more precise, AI systems most of all create a 'risk situation' that the EU guidelines try to solve by appealing to the semantics of trust, mixing the technical with desirable procedural attributes.

Most of all, the EU guidelines seem to adopt the language of consumer trust without reflecting on it. We agree with Gille et al. (2020) that the conceptualization of trust and trustworthiness in the EU guidelines lack coherence. By calculating that

[6] There is, however, an additional 'Assessment List for Trustworthy Artificial Intelligence' (EU HLEG) that offers a list of yes or no questions for self-assessment.

[7] Summary taken from the official website of the EU HLEG on AI: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (Accessed: 24 May 2022).

trustworthiness is a condition for trust in AI and, further, that trust is a condition for beneficial use and acceptance of the AI product, proposing a list of desirable attributes (*that are then called trustworthy*) seems like a marketing strategy to mask the risk situation, and not a policy to prevent the risk.[8]

Meanwhile, it facilitates the false impression of the AI system as having a moral character that respects (i.e., 'cares') for human values by anthropomorphizing the technology. Yet, it is the developer who is implicitly responsible for engineering the 'risk situation' by implementing the technology. Thus, it is also still the developer who should *guarantee* and (publicly) testify to the user and the community that these guidelines are met. Here, 'trustworthiness' could be considered an 'empty' concept that does not translate into clear requirements for laws and regulations that are legally (and not only morally) binding and would offer a societal tool to manage the consumer–developer relation and to govern innovation. Instead, the EU guidelines create a paradoxical constellation in which a possible patient would need to believe that the developers respected the regulations without having any direct personal contact or means of interacting with them.[9] The guidelines therefore serve as a rhetorical document for dissolving public responsibility into private.

Trust is often seen as a design component for successful social interaction and should thus be enhanced to optimize the utility (Kuipers 2022; Billings et al. 2012) as well as the acceptability (Whelan et al. 2018; Siau and Wang 2018) of the robot. As a consequence, trust is operationalized into a measurable parameter for evaluating the design and the probability of the human to use the device (Hancock et al. 2011), making 'social' an experience that can be constructed and consumed. Hence, to ensure that the innovation and use of robots for care and healthcare is not equally misguided by underspecified conceptualizations of trust and trustworthiness, a critical reflection and conceptual improvements are necessary.

Trust in technology and trust in robots: conceptual discussion

Trust and trust in technology: historical and conceptual foundations

The question of trust in social robots and artificial intelligence is part of a wider debate around trust in technology. While trust as a general research focus had its first heyday in the 1980s, introducing a sensitivity to the ethics and morality of trust (Baier 1986, 1991), the development of artificial intelligence brought the still unresolved issues of what trust actually is to the foreground again. Today, trust appears in the context of key issues within the philosophy of technology, such as automation and control.

[8] We thank an anonymous reviewer for the remark that a list in itself is not the problem. As Gille et al. (2020) pointed out, the EU HLEG on AI was overly manned by representatives from industry reflecting the interests of industry for an operational concept of trust that suits market purposes. The reductive tendencies hint towards an idea of consumer trust whose normative framework is limited.

[9] This cascade of trust relations also applies to the clinician or family members. Trust is mediated and influenced by the trust of other people. If the clinician trusts the robot, and the patient trusts the clinician, the patient might be more likely to trust the robot; similarly with family members.

In everyday understanding, trust is a basic human experience and psychological propensity or disposition that is prototypically actualized in interpersonal relations with other people: “We trust our partners to be faithful, we trust that our friends will keep our secret, and we trust our family members to stand by us in difficult times and situations” (Ryan 2020, p. 3). Trust in technology, and especially emerging technologies, differs in the way that the interpersonal model of trust is being transferred onto a nonhuman technical artifact. The transfer of the interpersonal model onto human–robot interaction takes into account the human disposition in the interaction but has been criticized for eliminating the central aspect of value-rich human interpersonal relationality. The phenomena of promise and betrayal that make up for a relation between moral agents, differentiate trust in people from reliance on tools and objects (Baier 1986; Holton 1994). “Most philosophers interested in characterizing the nature of trust regard it as a species of reliance” but not “‘mere’ reliance” (Goldberg 2020, p. 97), which leads to further distinctions of trust in different contexts such as e-trust (trust in digital environments; Taddeo and Floridi 2011).

The prototypical ‘trust situation’, however, is characterized by a lack of guarantees (McLeod and Ryman 2020). Epistemologically speaking, trust is thus a belief or supposition (Baker 1987). Trust is directed towards the future and the possible, positive outcome of a pronounced action or the truthfulness of a proposition. On the one hand, this can refer to trust as a positive-affective stance or attitude as a generalized positive belief that can encompass the goodwill and competence of the trustee (Jones 1996). This belief can be informed by emotions, but also by reasons and past experience. Rationality-based accounts of trust highlight the fact-based decision by means of calculating the trustworthiness of the trustee, as well as costs and benefits of the cooperation: “Following the rational choice paradigm, actors are self-interested, and trust is the rational outcome of the imperfect estimations of a trustor’s perspective on the trustworthiness of a trustee” (Hardin 2002). However, the rational calculus rather seems to evaluate how reliable (as opposed to trustworthy) the information about the trustee is and how to justify the evaluation (O’Neill 2018).

More generally, trust has been conceptualized by Luhmann as a conscious surrender to the incompleteness of information and a way of handling contingency (Luhmann 1968). Trust for Luhmann is a reaction to insecurity on different levels, from trust in persons to trust in institutions. Trust, in summary, reduces complexity and enables the subject to act. From the more abstract perspective on social systems, trust and technology converge with respect to human vulnerability, when one condition for trust to occur is a general situation of uncertainty where neither knowledge nor control are possible (Luhmann 1968). Attempting to isolate trust from its situative embeddedness, however, fails to recognize the full complexity of the phenomenon and its anthropological significance.

A ‘trust situation’, therefore, is also constituted by the risks that emerge from the basic dynamics of committing something to someone. Given that trust for Luhmann is a condition that enables the subject to act, wrongfully placed trust may have dire consequences. In the context of human–robot interaction in care, the analysis needs to look at what exactly is at stake. For Esther Keymolen (2016, p. 15), trust is an ability to expose oneself to the vulnerability and uncertainty regarding the future and the agency of other people, instead of “avoiding or diminishing vulnerability”. This leads to another structural tension because trust “solves a basic problem of social relations without *eliminating* the problem” (Möllering 2006, p. 6).

Hence, the concept of trust can be understood as a multidimensional and multilevel concept: at the individual level, it comprises intrapersonal and interpersonal psychological components, complex mechanisms of risk assessment and decision-making as well as relational and social aspects; at the societal level, trust is also influenced by accepted norms, rules and complex aspects such as social hierarchies and other sociocultural conditions. The conceptual work on trust as part of a transactional game–theory relation, however, is insufficient to capture the full ethical dimension of an alterity relation in a social situation.

For understanding trust in the context of human–technology interaction in care, a substantial challenge is to map these more or less well understood dimensions of human–human relations and interactions onto, e.g., human–robot interactions. In a care situation, the robot might be only ‘used’ as a tool by the care facility administration to fill an economic gap. The patient, nevertheless, might not only ‘use’ the robot, but socially interact and form a relation with it. How the patient conceptualizes the robot, i.e., understands the robot as a tool or a social interactional partner, needs to be taken into account. This can also be expressed in the way in which the robot appears to the patient as a machinic or “virtual” other (Coeckelbergh 2011) but lacks important prerequisites (e.g., in terms of moral agency).

In the following, we examine how phenomenological and anthropological aspects of human–human interactions, especially the notion of *responsivity*, may help to conceptualize human–robot interaction and the notion of trust in care situations.

Trust from a phenomenological–anthropological perspective

In the phenomenology of Alfred Schütz, the *natural attitude* in the lived world forms the starting point of human interaction, and as a consequence, also of theory formation. For Schütz, the lived world is characterized by an implicit familiarity and obvious self-evidence (Schütz and Luckmann 2003). Schütz and Luckmann highlight the significance of this primary natural attitude towards the world for the development of trusting relationships. For Thomas Fuchs, the process of original apprehension of the world is a process of *settling into* the world, what he calls *oikeiosis* (Fuchs 2015, p. 102), which has two main aspects: The experience of a fundamental self-familiarity with the body forms the foundation for the belief in the stability and continuity of our sensual perceptions and of a shared reality. Embodied experiences of safety and belonging emerge from interpersonal and social interactions with primary caregivers that consistently repeat and build a habitual practice of basic trust. This belief consists of an affective relation to the future (Fuchs 2015, p. 104) and can develop into a positive expectation towards the goodwill of another person. Familiarity with the world and trust as a human functioning evolve reciprocally. Trust is thus not just a social phenomenon but has to be understood as a basic human capability. To be able to trust is not only a human need but a potential and behavioral propensity that can be actively realized into a functioning. It is part of the basic capability to form relations, and in our interpretation specifically, to form relations with responsive others which is in turn part of a good and dignified life (Nussbaum 2000)[10].

Thomas Fuchs' analysis sheds light on the fact that basic trust is fundamentally connected to the practice and experience of *being cared* for. Embodied experiences of safety and belonging emerge from interpersonal and social interactions with primary caregivers that consistently repeat and build a habitual practice of basic trust. This interpersonal paradigm cannot be reduced to the logic of transaction or the concession of control. The early, embodied, and situational priming with close caregivers informs the human practice of trusting in other situations. In our view, to trust always involves the forming of and entering into a responsive relation with the trustee which marks the specific quality of trusting in respect to responsivity. We trust when someone cares about us.[11]

As trust is considered crucial for the success of the therapeutic or assistive intervention, the design of the robot come into the foreground again. It seems to follow that the design of social robots should avoid creating an atmosphere of danger or look like it would pose a threat but rather display a 'responsive appearance' that makes human patients interact with the machine as if it were 'caring' or in other words: trustworthy. When Coeckelbergh (2012, p. 56) speaks of a "default mode" of trust, we can link his social–phenomenological frame to the human disposition of trusting in the phenomenological–anthropological sense of a capability, an understanding that also undergirds our framework here.

Responsivity and the machinic other: a preliminary phenomenological–anthropological analysis of trust in social robots in care

The social robot as machinic other

As has been discussed by earlier phenomenological analyses, the robot may appear as a 'quasi-other' that portrays 'virtual intentionality' (Ihde 1990; Coeckelbergh 2011, 2010). In his postphenomenologically oriented relational ontology, Ihde identifies 'alterity relations' alongside 'embodiment relations', 'hermeneutic relations' and 'background relations'.[12] With the concept of alterity (also inspired by Levinas 1992), Ihde explores whether and to what extent we can speak of 'technology-as-

[10] Here, we refer again to Martha C. Nussbaum's capability approach which offers an understanding of human capability as a potential for a humanly good life. Martha Nussbaum's concept of dignity is—in our interpretation—fundamentally a relational one since the functionings–capabilities complex is embedded in a relational and situative frame of human learning. Nussbaum herself refers, for example, to "affiliation" as "being able to be treated as a dignified being" (Nussbaum 2006, p. 77).

[11] In cases where the formation of a responsive relation with the potential trustee is impaired or impossible, the semantics of trust and trustworthiness is misguiding.

[12] For example, while embodiment relations describe a form of incorporation that usually does not happen with a robot, it is useful when looking at prosthetics or exoskeletons. A special case might also be an advanced wheel-chair. The robot might also offer hermeneutic relations, if it mediates between the world and the patient (e.g., analyzing and interpreting health data, or reading from a screen).

other' or also of a 'quasi-otherness' in the encounter with machines. These three relations form the dimensions of 'technological intentionality' (Ihde 1990) that inform the appearance of technological artifacts 'as if' they own a form of subjectivity.

The prominence of trust and trustworthiness in the debate about AI-complemented social robots thus highlights this ambiguity: It circles around the blurry ontological status of robots which oscillates, in western philosophy, between the categories of object and subject (tool and agent), where the latter has traditionally been limited to European, white, male, human beings and framed as active, rational and autonomous as highlighted by critical discussions in feminist philosophy (Loh and Coeckelberg 2019). Critiques of the notion of technological artifacts as (potential) subjects come to the surface again by automation and AI, when the appearance and functionality of the robot suggest a form of agency.[13]

As an alternative to these debates, a phenomenological–anthropological account focuses on the phenomenal appearance of the robot while describing the structure of experience that emerges in the interaction between human and robot and its situative embeddedness with respect to human dispositions and capabilities.

Responsivity in caring for others

In care, particularly health care, robots are expected to be more than mere tools: they should recognize our needs and respond to them in appropriate ways. In the absence of a human doctor, nurse, or therapist, the patient's counterpart is the artificial replacement. The robot is assigned the position of the 'caring other' by its function as replacement of the human caregiver as well as by the situational social dynamic, including the normative dimensions of what the robot is ought to do and whatnot as a socially assistive robot (Vaesen et al. 2013).[14]

"Care is a practice of awareness and relatedness, which includes self-care and small gestures of attention the same way as nursing and providing interactions as

[13] Without entering the philosophical debate about agency in general (Schönau 2019) and robot agency in particular (Jackson and Williams 2021), some of the still debated questions are worth mentioning: Can the machine be assigned empathy (Misselhorn 2021), creativity (McCormack 2008), or even authorship (Gervais 2019)? Looking at how the actor is situated in the network can lead to a political understanding of robots as complex sociotechnological systems (Latour 2006), or of the robot as the carrier of values and subject of responsibility in the sense of being a moral patient (Loh 2019). How the robot agency is designed also evokes normative discussions about what constitutes a "good" or even "moral" agent (Floridi and Sanders 2004; Sullins 2006). To speak of a tool ignores the complexity of the interactive features and competences the machine displays.

[14] Relating this point to the idea of giving "discretionary authority" (Nickel 2022, p. 1) to the AI system, we argue that the care-situation creates a certain expectation by the care-receiver (and the care-practitioner) towards the 'caring function' of the robot in the assistive context, too. However, in contrast to answering diagnostic questions, or analyzing data or offering a judgment (see example in Nickel 2022), care cannot be experienced without the affective and normative dimensions. Trusting ("giving discretionary authority") that the social robot fulfills its function is thus fundamentally connected to the constraints of the situation—which is not a purely epistemic situation but a social, relational, and *responsive* one. The normative dimensions of care can thus not be just simply transferred by implying ethical obligations "on the part of the AI practitioner" (Nickel 2022, p. 5) toward the care-taker.

well as collective activities" (Conradi 2001, p. 13, translated by IS)[15]. Caregiver and care receiver are in a relationship that is often characterized by asymmetry and dependency.[16] Yet, this dependency is one of reliance inside a "world comprised of relationships rather than of people standing alone" (Gilligan 1982, p. 29). Central aspects of a care orientation are attentiveness and awareness, aid, and responsiveness to needs (Bubeck 1995).

Hence, caring for others is an intrinsically relational process that requires certain capacities for social interaction such as language, but also 'theory of mind'—the ability to see the world from someone else's perspective—and empathy. Together, these capacities enable us to specifically *respond* to the needs of someone. This human capacity for responding seems fundamental for successful interactions, yet it has not been a very prominent topic in recent ethical scholarship on trust and trustworthiness concerning human–technology relations.[17] We want to connect the anthropological observation of a human capacity or *capability* to respond with the 'responsive appearance' and the phenomenon of being perceived as responsive (virtual responsivity) and take a closer look at the phenomenological structure underlying this behavior. Here, we suggest that philosophical theories of *responsivity* (*Responsivität*), particularly from the phenomenology of Bernhard Waldenfels, could provide key insights for understanding human–robot interactions and, ultimately, the characteristics of trustworthy relations to robots and other machines. Therefore, to answer the question whether robots can *truly* care, we need to understand the role of responsivity, particularly in relation to trust.

The interest in responsive behavior already elicited empirical studies. However, without an ethical and phenomenologically informed analysis. Psychological research has been interested in responsive behavior in close human relationships with respect to intimacy (Reis and Clark 2013; Reis 2014), clustering a certain behavioral morphology as 'responsiveness' or 'perceived responsiveness' towards the partner.[18] Here, we aim to extend the literature on SARs by a phenomenolog-

[15] "Care ist eine Praxis der Achtsamkeit und Bezogenheit, die Selbstsorge und kleine Gesten der Aufmerksamkeit ebenso umfasst wie pflegende und versorgende menschliche Interaktionen sowie kollektive Aktivitäten" (Conradi 2001, p. 13).

[16] This fact also diminishes the scope of contractualist theories of trust since they focus mainly on trust between equals.

[17] Cappella and Pelachaud argued in 2002 that the "science of relationships—especially human interaction in relationships—needs to be imported into the science of modeling interactions" (Cappella and Pelachaud 2002, p. 5). While the definitions of responsiveness used by Cappella and Pelachaud (2002) do not offer much in terms of understanding the quality of responsive behavior or the ethical implications of responsive social interactions, the approach to analyze the basic pragmatic and sequential structure of 'being responsive' in human interaction corresponds with the need for empirically accessible instantiations of responsiveness that (also) serve as the foundation of a phenomenological description. Further findings in social psychology have been translated into "responsiveness cues" by Birnbaum et al. (2016a, p. 416) and implemented into a robot to study human bonding.

[18] For example: "visible displays of attention and comprehension of a communication", "openness and nondefensiveness" or "feeling understood, validated, and cared for" (Reis and Clark 2013, p. 7). Many of the listed definitions in Reis and Clark (2013) include performed 'responsive' engagement (e.g., assistance and emphatic gestures of caring) that also support the idea that 'responsiveness' is an aspect of the more complex capability to respond and the phenomenologically relevant structure: *responsivity*.

ical–anthropological account of the interrelation between responsiveness, trust and care.

Responsivity, virtual responsivity and constructed responsiveness

Responding to the needs of another being, answering to the demand of the other, is a special form of caring responsivity: "[C]are is integrated into our dealing with things, others and ourselves. [...] It appears in everyday form, for example in the form of carefulness and in institutional form as care for children and old people, as preventive medicine, a public welfare, as custody of children, as religious or secular pastoral care" (Waldenfels 2020, p. 189). The "aim of therapy consists in enhancing or restoring responsivity" (Waldenfels 2020, p. 196) which is in line with the medical anthropology of Viktor von Weizsäcker (1987) and Kurt Goldstein (1934). We would submit that responsivity is a key phenomenon for human–robot interaction, too, when therapy, for Waldenfels, "responds by getting the other to respond" (Waldenfels 2020, p. 203)—which is a stated goal in the case of, for example, pet robots for patients with dementia. Indeed, it is evident that when robots are perceived as responsive, they have a stronger effect on the human user, be it a greater reduction in the perception of pain (Geva et al. 2022), a decrease in salivary cortisol levels (Tanaka et al. 2012), or an increase in prevalence of positive emotions (Crossmann et al. 2018).

As discussed above, trust differs from reliance in the belief that the person we are trusting acts with a specific attitude of goodwill towards us (Baier 1996). The expectation of goodwill, or at least the expectation of a sense of obligation, as a (habitual) affective attitude derived from positive relations with primary caregivers, includes an expectation that the trustee 'cares about' the trustor, i.e., *responds* and not only reacts. Hence, responsivity is crucial in a bi-directional manner: Without responsive others, human beings would not be able to develop the capability to trust. As trust is closely connected to responsivity, trust will be understood particularly as the capability to form affective relations with responsive others.

We argue that the appearance of caring displayed in the functionality of the robot is achieved by the *responsive qualities* that the robot exhibits achieved by an engineered or constructed responsiveness in the technical sense.[19] In line with Coeckelbergh's terminology, we could call that phenomenon 'virtual responsivity' in the form of constructed responsiveness.[20] Virtual denotes the experience of pos-

[19] We want to thank an anonymous reviewer for the following phrasing: It could be said that the constructed responsiveness of the robot (generated by the social affordances it provides by the AI-enhanced speech recognition and its physical mobility) gives the robot the appearance of 'virtual responsivity' as perceived by the patient.

[20] Terminological consistency with respect to the debate in the phenomenology of technology lead to the choice of distinguishing 'virtual responsivity' as the perceived potential for responding experienced by the human patient from the technical notion of 'constructed responsiveness' as part of the technical functionality of the artifact. We like to point out that further phenomenological research is needed to systematically describe the differentiation of the first-person experience of the responsive appearance of social robots in care, for example, by qualitative interviews with patients to support or falsify the claim. However, the latter accounts for the human responsive need for trusting to be realized in its full existential dimension as a capability. The subjective, first-person element of perceived potential for responding to the

sibility, i.e., something exists in virtue of its potential.[21] Virtual responsivity also refers to the interplay between the perception of something as something and the thing that appears as something. The entity or thing in question offers affordances (in the sense of James J. Gibson 1979)[22], options to act upon and interact with, that are perceived as stimulus and attraction in a certain way, e.g., as inviting, as threatening or—as trustworthy. Affordances constitute a "symbolic excess" (Waldenfels 2019, pp. 376–377); they can inspire a certain action *and more* in terms of a creative openness. Affordances suggest ways of acting and using, which in the case of social robots could be framed as 'social affordances'. These include not only interactional prompts but also the potential for bonding experiences. A social robot in a care situation, thus, 'invites' the patient to socially respond to the affordances on a relational and possibly emotional level. For determining virtual responsivity, it is not essential whether the robot *truly* cares (as this would lead also into complex discussions about robot consciousness, etc.), but how and why the patient responds to the social affordances of the robot resulting in the subjective *impression* of being cared-for.

Basic dimensions of human responsivity

The capability to respond relates to a more fundamental phenomenal structure that needs further clarification. The phenomenology of responsivity is rooted in the challenge that "the other" poses to the self. This challenge is a moment of alienation, experienced as a 'call upon the self' (Waldenfels 2011, p. 36), which imposes on the self a doubled pathos and a demand. This pathic dimension of challenge, provocation, withdrawal and defiance constitutes a crucial asymmetry of call and response for Waldenfels (cp. 2011, p. 37). In responding, the self is already "incited, attracted, threatened, challenged and appealed" (Waldenfels 2010) by something different from itself that calls upon it but never becomes quite normalized. The call, or demand (*Anspruch*) "is directed *at someone*" and has "a claim *to something*" (Waldenfels 2011, p. 37). It starts with looking-at and listening-to (cp. Waldenfels 2011, p. 37) and leads to the invention of an answer that meets the "invite" of the other (Waldenfels 2011, p. 38). Responsivity is thus a "basic trait present in all our behaviors towards things, towards ourselves and towards others" (Waldenfels 2010).

From there follows a scantly noticed anthropological description: "The human being is an animal which responds" (Waldenfels 2011, p. 38). Responding, however, does not only constitute a specific speech act, it constitutes a basic dyadic phenomenological motif in the structure of human experience that can also be described in philosophical–anthropological manner as a capability to receive and answer. There exists for Waldenfels (2011, p. 38) a necessity and obligation to

subjective needs and existential situation, in our account, is the evaluation of the trustworthiness of a given other.

[21] In "Das leibliche Selbst", Waldenfels uses the term "virtuality" (Virtualität) to describe the human sense of possibility (Sinn für das Mögliche), (Waldenfels 2021, p. 199). For further readings about virtuality in the sense of digital or cyber virtuality see Kasprowicz and Rieger (2020).

[22] The US American psychologist James J. Gibson (1979, p. 127) introduced the concept of affordances as implicit calls for actions that an object imposes on someone.

answer, evoked by the presence of the other. Besides the above-mentioned re-interpretation of Husserl's concept of intentionality (Husserl 2009), the argument can be interpreted by structuring it into four aspects of responsive relationality: the symbolic–expressive dimension, the embodied dimension, the situative embeddedness in the life-world, and the ethical dimension.

1. The symbolic–expressive dimension: Language is a powerful medium to convey meaning and generate sense, and yet communication and semantic copractices go beyond solely linguistic features. The symbolic functionality of human language has been identified as the anthropological difference to the stimulus–reaction schema of other beings, conceptualized as 'responding' rather than reacting (Cassirer 1944).

   Language serves as the prime medium for responsive expressionality in the form of question and answer (Waldenfels 2007). Similar, but less anthropocentric and more relational conceptualizations can be found in Donna Haraway's 'response-ability' (2003) aiming at a reciprocal relationship between humans and nonhumans that roots in responsive interaction.
2. The embodied dimension: Responding for Waldenfels is part of the "embodied responsory" (*leibliches Responsorium*) that is not limited to language (Waldenfels 2019, p. 255) but includes emotions, moods, and expressive movement. While *looking at* and *listening to* constitute a prototypical embodied-responsive stance to the world, gestures of giving and receiving hint towards an elementary normativity, an ethos of the senses and the body (Waldenfels 2019, p. 388).[23]
3. The situative embeddedness in the life-world: Situations constitute the frame of action and understanding. People can get accustomed to situations that occur repeatedly, following scripts and rules that help to deal with ambiguity. However, human beings can be characterized by their ability to transform this frame of action and understanding, which Waldenfels calls the "sense of possibility" or "virtuality" (*Virtualität*; Waldenfels 2021, p. 199). The sense of possibility inherently accompanies human perception and interaction with other beings and objects, offering potential threats or welcome chances (cp. Waldenfels 2021, p. 204). The given might not be "taken as that which it really is, but it is already viewed in light of its possibilities" (Waldenfels 2021, p. 204).[24] The sense of possibility (or virtuality) can thus be linked to the question of the human condition, expressed as a basic and fundamental situation embedded in the life-world with respect to the actualization of potential.

[23] In terms of enactive agency, the lived body offers various additional forms of responsive interactions: moving in synchrony (e.g., in dance), in recreational games and play, in music (e.g., jazz improvization), in navigating and other structured forms of interaction. Through this enactive agency, the lived body already engages in an interpreting way—perhaps in the sense of 'corporeal hermeneutics'—with its environment. Organs and limbs are not just tools to achieve a goal that has been set by a decoupled mind following the model of a straight line from aim to action. Human responsivity is a disposition that can be realized through embodied and responsive agency in dynamic interactions with the environment and other beings.

[24] "Der andere Pol wäre die Hypostasierung des Möglichkeitssinnes: das Gesehene wird nicht genommen als das, was es ist, sondern es wird immer schon im Hinblick auf sein Anderssein gesehen" (Waldenfels 2021, p. 204).

4. The ethical dimension: The ethical dimension follows the basic conception that responding as such starts from “somewhere else”[25] (Waldenfels 2019, p. 255). The initiative arises from the other, from an “alien impulse” (Waldenfels 2019, p. 255). This line of thought is rooted in the phenomenology of alterity of Emmanuel Levinas, who argued for a priority of the ethical dimension over the self-constituting processes of the subject in traditional Western philosophy (Levinas 1969). In fact, the self is pre-ontologically dependent on the other, its social directedness, so to speak, is an alterity-oriented, i.e., responsive, one.[26]

The social robot as virtual responsive other

The dimensions of human responsivity relate back to the care-situation and constitute the phenomenological–anthropological foundations of the experience of being cared-for—and as a consequence, the capability and willingness to trust the robot. As the basic trait of interacting with and experiencing the world, human beings respond to the incentives of the robot from their human disposition for trust if the robot appears to be responsive.

Most of the robots are still only accessible in experimental trials and research contexts, so a full integration into the life-world of people in need of care is not a reality yet. However, video documentations, ethnographic observations, and qualitative interviews (e.g., Koren et al. 2022) can be used as source material for a preliminary phenomenological–anthropological description of the phenomenal presence of the robots and their effects on the human disposition for trust, which could be complemented by an in-person analysis in the future. As Hancock et al. (2011) categorized, the “social character” of the robot is the second most important factor in their meta-analysis of trust in human–robot interaction. This social character of the robot, we argue, is in part due to its ‘virtual responsivity’ that encompasses appearance and functionality. Virtual responsivity aims to capture the perceived potential of the robot to respond (i.e., its constructed responsiveness) that is informed by the human sense of possibility. In the following, the specific appearances that constitute the constructed responsiveness of the robot will be shortly described.

1. *Dialogic responsiveness*: Robot systems like LIO offer dialogic communication with the patient they serve. Most prominently, social robots ask questions like ‘What can I do for you today?’ or ‘How can I help you?’ The question is a proposal of options (Waldenfels 2007, p. 191), while adhering to the laws of polite turn-taking. Birnbaum et al. (2016a) named positive responsive speech acts like “You must have gone through a very difficult time” (Birnbaum et al. 2016a, p. 419) or “I completely understand what you have been through” (Birnbaum et al. 2016a, p. 419) as linguistic codes for responsive behavior. The robot might also comment

[25] Deutsch im Original: “Die Pointe besteht kurz gesagt darin, dass das Antworten als solches anderswo beginnt” (Waldenfels 2019, p. 255).

[26] In seeing the face, the irreducible face (*Antlitz*), the Other shows and offers its vulnerability and invokes an ethical involvement and responsibility (Levinas 1969, p. 197).

on the activities of the human as if it was concerned ('Did you sleep well?')[27], offer "confirmations" (Hoffman et al. 2014) or explicit acknowledgement of the previous speech segment (Birnbaum et al. 2016b). Adding language, speech acts, and dialogically fitting reactions like summarizing, affirming, and repeating to the embodied gestures of understanding enhances the constructed responsiveness of the robot.

2. *Embodied responsiveness*: Attributing competences to the robot seems to rest on the robot's appearance as an entity that can *enact* these competences physically. Primarily, it is movement and seemingly autonomous movement that evokes qualities of a living being and animacy (Plessner 2016, pp. 179–180)[28]. The materiality and superficial appearance of the robot is also the zone of contact between human and machine. Enhanced sensors, cameras, and microphones create a data-sensitive interface to the environment. Understanding trust in human–robot interaction has to consider this sensual encounter with the robot's 'body' and its mobile functions in respect to its sensors.[29] This dynamic physicality of the robot leads to the impression that the robot has received and *internally* processed the environmental data and interactional cues and *responds* to them in an active and agentive way.[30]
3. *Social responsiveness*: Social robots also offer social affordances by their movements that may generate interactions like play, dance, cooperation and appreciation: The human user is invited to repeat or imitate what the robot is doing. Prompting imitation, human and robot enter a synched and inter-coordinated interaction in which the robot responds to the human moving, giving feedback and correcting the movement if need be—or the other way around. *Leading* and *following* are the corresponding responsive social phenomena. Accompanied by communication and esthetic features that enhance familiarity, these socially oriented cues reinforce the impression of alterity and social roles.[31] Combined, these cues tap into our inherent propensity to trust based on social signals.

[27] Quote translated from the developers (F&P Robotics) presentation about robot LIO: Lio—Ein Tag in der Stiftung Rossfeld: https://www.youtube.com/watch?v=7Nm-2UmBYOo (24.05.2022): 0:52 min.

[28] For Plessner, a movement that is alive presents itself phenomenologically by a tendency towards its own fulfillment. Living movements appear as if they 'could have proceeded differently' (Plessner 2016, p. 180).

[29] For example, the pet robot PARO 'responds' to contact by signaling pleasure with animal sounds as if it can feel being touched. Robots like PEPPER, NAO, PALRO, LIO, HUGO or RUDY move and navigate around objects and appear as autonomous and self-moving. They are 'responsive' to their surroundings in the sense that they can *react* and *adapt* their physical movements to spatial conditions.

[30] In Birnbaum et al. (2016a, b), the robot TRAVIS was explicitly programmed to display embodied responsiveness by maintaining a "forward focus towards the participants, gently sway back and forth to display animacy, and nod affirmatively in response to human speech" (Birnbaum et al. 2016a, p. 419). Embodied responsiveness here is connected to the directing of the gaze and moving the head in the direction of the speaker or action. This creates the appearance of awareness. PEPPER includes arm and hand gestures simulating 'active listening' and its widely opened eyes signal attention. The embodied expressivity of the social robot creates an appearance of an 'outside' and an 'inside' or 'inner life' (that is perceived as the cause for the movement), which in turn relates further to the responsive dimension of 'understanding' in Maisel et al. (2008).

[31] In most designs, the robot's appearance is adapted to human esthetic preferences for human-likeness (so-called anthropomorphic design). The design of the robot PEPPER evokes associations with a child due to its height, round forms, big eyes, and disproportionately big head. Although the arms appear mus-

Ethical implications of trusting robots

Following the preceding analysis, it could be said that the constructed responsiveness of the robot (generated by the social affordances it provides), the AI-enhanced speech recognition and its physical mobility, gives the robot the appearance of 'virtual responsivity.' Analogous to the phenomenological structure of responsivity and its dimensions in human lived experience, this raises questions about potential ethical pitfalls, tensions, and challenges of SARs, especially in the context of medical applications.

From a principlist, deontological perspective, interacting with robots may affect a patient's personal autonomy (e.g., by carrying out actions on behalf of a patient) and we need to balance the potential positive impact of interacting with a robot (beneficence), for example, for therapeutic purposes, with the inherent risks (safety, medical, psychological) to avoid harm to patients (nonmaleficence; Beauchamp and Childress 2001). Another important principle is justice, i.e., how can we ensure equal access and use to complex and costly technologies such as robots, especially in economically and/or technologically constrained settings, given that this technology often requires an advanced digital ecosystem. In addition to these classical principles, we want to especially highlight three ethically relevant dimensions related to responsivity that also need to be considered in human–robot interaction in medicine: vulnerability, dignity, and the ethical dimensions of the responsive capability to trust.

Vulnerability can be understood as an anthropological foundation of human existence that, in its most basic form, denotes a human's propensity to experience harm based on their certain characteristics, for example, based on specific medical conditions but also—importantly—because of group-based (e.g., gender-based) discrimination or other forms of structural inequalities and injustices (Herzog et al. 2022). In the context of biomedical ethics, vulnerability is a multidimensional concept denoting various yet specific ways in which a human being can be vulnerable[32]. Vulnerability increases in a situation where responsive interaction is particularly and explicitly needed, for example, when the patient's own capability to respond is limited or impaired due to sickness, circumstances (isolation), or age[33].

SARs are meant to facilitate and improve care situations and they succeed in that their design interplays with the dimensions of responsivity shown above. There is, however, a potential for deception when the ethical and normative significance of responsivity is not appropriately reflected. The constructed responsiveness and resulting 'virtual responsivity' might make it difficult to see that the robot is used as

cular, suggesting strength, the form of the waist is female-like. Combined, PEPPER seems androgynous, almost without gender, reducing sexual connotations to a form of innocence. The robot is designed to have a character that helps to design social roles, for example, the role of the 'butler' for LIO.

[32] For example, a child with developmental disabilities that lives under conditions of extreme poverty would be vulnerable because of their age, their underlying condition as well as their socioeconomic situation.

[33] As robots, AI systems and other digital technologies play an increasing role in medical care, it has been suggested that interaction with digital technologies could itself be an important emerging source of vulnerability ('digital vulnerability', cp. Kellmeyer 2019b).

a medium and interlocutor that only allows for an indirect form of 'virtual responsivity' without the essential direct reciprocity of ethical responsivity.

It is clear that human beings can respond to virtual, i.e. imagined or perceived possibilities and turn them into real possibilities, actual potential and actions or skills. Responsivity structurally includes a sensitivity to the potential of the other, i.e., the "more" of the other. This compels us to take another important ethical consideration into account: The special quality of human responsivity seems to include a claim to dignity. Responsivity not only allows us to answer to the vulnerability of the other person, their fundamental otherness and difference, but also to the perceived potential of the other being as a fundamental capability and important prerequisite for health, wellbeing, and flourishing. Treating someone with dignity means respecting the capacity for responsivity of the person as an actual potential.

One might argue, however, that the virtual responsivity of a social robot alone does not entail specific ethical risks because the robot is not a person, i.e., it does not possess a sense of self, moral consciousness nor moral agency and other characteristics (depending on one's understanding of what makes a person). We would agree with this point but would argue that insofar as the robot displays an operational or functional type of responsivity and acts socially interactive in a persuasive manner, it might be *mistaken* for an 'Other' that possesses a more elaborate moral status than being merely a machine. This elevated moral status might then make a human treat the robot with a kind of dignity that is similar to the dignity they confer to fellow humans. This can be a source of, at least, disappointment if not psychological harm. Imagine a human that is becoming so attached to a robot that they feel sad or lonely if the robot is no longer with them (e.g., because therapy has ended) or because the robot cannot, after all, reciprocate in treating the human in a dignified manner because, ultimately, it lacks the conceptual understanding and hence agency. In situations in which a robot cares for a human, in the sense of providing care work (whether physical/'manual' labor such as lifting, washing etc., or social labor such as communication or other forms of interaction), it might thus indeed matter to the human if the robot *truly* cares in the sense of having the capacity for the required psychological and anthropological features of caring (e.g., empathy, a sense of relationality and full responsivity), all of which are invariably tied to consciousness and personhood, both of which robots in their current form clearly lack. An important empirical question (with ethical implications) at the level of designing and using robots could therefore be to investigate, which types of responsivity displayed by a robot elicit which kinds of feelings (e.g., of trust, feeling respected as a person) in the human 'user' and whether and to which degree this would be considered to be a form of manipulation.

Regarding the ethical implications of trust, we want to propose that trust entails the belief that the other person respects one's own claim to dignity by virtue of their capacity to respond. It is a belief to be treated as a being with an inherent 'virtuality' (in the sense of potentiality) to flourish which forms an intrinsic value[34]. If responsivity is the fundamental phenomenological structure in relation to the

[34] In designing digital technologies for medical applications, human flourishing could be indeed a key concern in the sense of flourishing-oriented engineering (cp. Germani et al. 2021).

other, then it is also a prerequisite for trusting when it includes an answer to the perceived potential and 'virtuality' of the other being as expressed in fundamental capabilities.

Setting the stakes this high for trust and trustworthiness allows us to critically examine the potential of robots with respect to the virtuality and potential of human beings. Especially in care situations, where respecting human dignity still forms an important ethical principle for evaluating actions, we need to reflect on trust as an essential human disposition and capacity to form relations, and trustworthiness as the result of responding to this disposition.

We hope to have shown that our analysis of the complex relationship between trust as a concept and phenomenon in human–human relations and the situative dimensions of care proposes a nuanced answer to the question of whether social robots should be trustworthy. The robot is trustworthy if the constructed responsiveness is not used to deprive patients of their claim to dignified care and if the limitations of the robot's 'virtual responsiveness' are being made transparent to the patient. SARs' constructed responsiveness should be appropriate to the therapeutic intervention and not overact on inducing trusting and bonding behavior by patients. Informed usage should therefore perhaps include disclaimers about the human propensity to trust responsive others and human orientation towards alterity. 'Ethically sensitive responsiveness' could be an important factor in the assessment of social robots. It would include boundaries to the social affordances and can explicitly state reasons why certain responsive features are therapeutically necessary and also, for example, justify the deactivation of responsive features in specifically vulnerable patients. SARs should not be introduced as harmless toys if they are in fact (or are perceived as) 'machinic others' with virtual responsive capacities, especially if they are designed to be placed in and maybe replace a responsive human–human relationship of care.

Funding The research for this paper by author P. Kellmeyer was partly supported by a grant (00.001.2019) of the Klaus Tschira Foundation, Germany. The research by author I. Schröder was financed by the Baden-Württemberg Stiftung in the project AI-TRUST.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflict of interest I. Schröder, O. Müller, H. Scholl, S. Levy-Tzedek and P. Kellmeyer declare that they have no competing interests.

Ethical standards For this article no studies with human participants or animals were performed by any of the authors. All studies mentioned were in accordance with the ethical standards indicated in each case.

References

Baier A (1986) Trust and antitrust. Ethics 96(2):231–260. https://doi.org/10.1086/292745

Baier A (1991) Two lectures on "trust": Lecture 1, "trust and its vulnerabilities" and lecture 2, "sustaining trust". Tanner lectures on human values, vol 13. University of Utah Press, Salt Lake City, pp 109–174

Baier A (1996) Moral prejudices. Essays on ethics. Harvard University Press, Cambridge

Baker J (1987) Trust and rationality. Pac Philos Quart 68(1):1–13. https://doi.org/10.1111/j.1468-0114.1987.tb00280.x

Beauchamp T, Childress J (2001) Principles of biomedical ethics. Oxford University Press, Oxford

Billings D, Schaefer K, Chen J, Hancock P (2012) Human-robot interaction: developing trust in robots. In: Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction (HRI '12), pp 109–110 https://doi.org/10.1145/2157689.2157709

Birnbaum G, Mizrahi M, Hoffmann G, Reis H, Finkel E, Sass O (2016a) What robots can teach us about intimacy: The reassuring effects of robot responsiveness to human disclosure. Comput Hum Behav 63:416–423. https://doi.org/10.1016/j.chb.2016.05.064

Birnbaum G, Mizrahi M, Hoffman G, Reis H et al (2016b) Machines as a source of consolation. Robot responsiveness increases human approach behavior and desire for companionship. In: Proceedings of the 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI 2016) https://doi.org/10.1109/HRI.2016.7451748

Bordin E (1974) Research strategies in psychotherapy. John Wiley & Sons, Hoboken

Bordin E (1979) The generalizability of the psychoanalytic concept of the working alliance. Psychol Psychother Theory Res Pract 16(3):252–260. https://doi.org/10.1037/h0085885

Breazeal C, Dautenhahn K, Kanda T (2016) Social robotics. In: Siciliano B, Khatib O (eds) Springer handbook of robotics, 2nd edn. Springer, Cham, pp 1936–1972 https://doi.org/10.1007/978-3-319-32552-1_72

Bringsjord S, Govindarajulu N (2018) Artificial intelligence. Stanford encyclopedia of philosophy. https://plato.stanford.edu/entries/artificial-intelligence/. Accessed 25 May 2022

Bubeck D (1995) Care, gender, and justice. Clarendon Press, Oxford

Bundesministerium für Bildung und Forschung (BMBF) (2018) Pflege durch interaktive Technologien erleichtern. https://www.bmbf.de/bmbf/de/forschung/gesundheit/pflege/pflege_node.html. Accessed 25 May 2022

Cappella J, Pelachaud C (2002) Rules for responsive robots: using human interactions to build virtual interactions. In: Vangelisti A, Reis H, Fitzpatrick M (eds) Stability and change in relationships. Cambridge University Press, Cambridge, pp 325–354

Cassirer E (1944) An essay on man; an introduction to a philosophy of human culture. Yale University Press, Newhaven

Coeckelbergh M (2010) Moral appearances: emotions, robots, and human morality. Ethics Inf Technol 12:235–241. https://doi.org/10.1007/s10676-010-9221-y

Coeckelbergh M (2011) Humans, animals, and robots: A phenomenological approach to human-robot relations. Int J Soc Robot 3(2):197–204

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

Conradi E (2001) Take care. Grundlagen einer Ethik der Achtsamkeit. Campus, Frankfurt a.M., New York

Crossman MK, Kazdin AE, Kitt ER (2018) The influence of a socially assistive robot on mood, anxiety, and arousal in children. Prof Psychol Res Pract 49(1):48–56. https://doi.org/10.1037/pro0000177

Dinç L, Gastmans C (2011) Trust and trustworthiness in nursing: an argument-based literature review. Nurs Inq 19(3):223–237. https://doi.org/10.1111/j.1440-1800.2011.00582.x

Dinç L, Gastmans C (2013) Trust in nurse–patient relationships: A literature review. Nurs Ethics 20(5):501–516. https://doi.org/10.1177/0969733012468463

Duran JM, Jongsma KR (2021) Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. J Med Ethics 47:329–335. https://doi.org/10.1136/medethics-2020-106820

Erikson E (1950) Childhood and society. W. W. Norton, New York

Feil-Seifer D, Mataric M (2005) Defining socially assistive robots. 9th International Conference on rehabilitation robotics, 2005. In: ICORR 2005, pp 465–468 https://doi.org/10.1109/ICORR.2005.1501143

Ferrario A, Loi M, Viganò E (2021) Trust does not need to be human: it is possible to trust medical AI. J Med Ethics 47:437–438. https://doi.org/10.1177/0018720814547570

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

Fuchs T (2015) Vertrautheit und Vertrauen als Grundlagen der Lebenswelt. In: Bermes C, Hand A (eds) Phänomenologische Forschungen. Felix Meiner, Hamburg, pp 101–117

Germani F, Kellmeyer P, Wäscher S, Biller-Andorno N (2021) Engineering minds? Ethical considerations on biotechnological approaches to mental health, well-being, and human flourishing. Trends Biotechnol 39:1111–1113

Gervais D (2019) The machine as author. Iowa Law Rev 105: 2053–2106. Vanderbilt Law Research Paper No. 19–35. https://ssrn.com/abstract=3359524. Accessed 25 May 2022

Geva N, Hermoni N, Levy-Tzedek S (2022) Interaction matters: the effect of touching the social robot PARO on pain and stress is stronger when turned ON vs. OFF. Front Robot AI 9:926185. https://doi.org/10.3389/frobt.2022.926185

Gibson J (1979) The ecological approach to visual perception: classic edition. Houghton Mifflin, Boston

Gille F, Jobin A, Ienca M (2020) What we talk about when we talk about trust: Theory of trust for AI in healthcare. Intell Based Med. https://doi.org/10.1016/j.ibmed.2020.100001

Gilligan C (1982) In a different voice. Psychological theory and women's development. Harvard UP, Cambridge

Goldberg S (2020) Trust and reliance 1. In: Simon J (ed) The Routledge handbook of trust and philosophy, 1st edn. Routledge, London, pp 97–108

Goldstein K (1934) Der Aufbau des Organismus. Einführung in die Biologie unter besonderer Berücksichtigung der Erfahrungen am kranken Menschen. Nijhoff, Den Haag (Neuausgabe: Hoffmann T, Stahnisch F (Hrsg) Fink, Paderborn 2014)

Greene J, Ramos C (2021) A mixed methods examination of health care provider behaviors that build patients' trust. Patient Educ Couns 104:1222–1228

Hancock P, Billings D, Schaefer K, Chen J, de Visser E, Parasuraman R (2011) A meta-analysis of factors affecting trust in human-robot interaction. Hum Factors 53(5):517–527. https://doi.org/10.1177/0018720811417254

Haraway D (2003) The companion species manifesto. Dogs, people, and significant otherness. Prickly Paradigm, Chicago

Hardin R (2002) Trust and trustworthiness. SAGE, New York

Henschel A, Laban G, Cross E (2021) What makes a robot social? A review of social robots from science fiction to a home or hospital near you. Curr Robot Rep 2:9–19. https://doi.org/10.1007/s43154-020-00035-0

Herzog L, Kellmeyer P, Wild V (2022) Digital behavioral technology, vulnerability and justice: towards an integrated approach. Rev Soc Econ 80(1):7–28. https://doi.org/10.1080/00346764.2021.1943755

High-Level Expert Group on Artificial Intelligence (HLEG AI) set up by the European Commission (2019) Ethics guidelines for trustworthy artificial intelligence. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai. Accessed 25 July 2022

Hoff K, Bashir M (2015) Trust in automation: integrating empirical evidence on factors that influence trust. Hum Factors 57(3):407–434. https://doi.org/10.1177/0018720814547570

Hoffman G, Birnbaum G, Vanunu K, Sass O, Reis T (2014) Robot responsiveness to human disclosure affects social impression and appeal. In: Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction (HRI '14), pp 1–8 https://doi.org/10.1145/2559636.2559660

Holton H (1994) Deciding to trust, coming to believe. Australas J Philos 72(1):63–76. https://doi.org/10.1080/00048409412345881

Husserl E (2009) Ideen zu einer reinen Phänomenologie und Phänomenologischen Philosophie. Philosophische Bibliothek, vol 602. Meiner, Hamburg

Ihde D (1990) Technology and the lifeworld. From garden to earth. Indiana University Press, Bloomington

Jackson RB, Williams T (2021) A theory of social agency for human-robot interaction. Front Robot AI 8:1–15. https://doi.org/10.3389/frobt.2021.687726

Jones K (1996) Trust as an affective attitude. Ethics 107(1):4–25. https://doi.org/10.1086/233694

Kasprowicz D, Rieger S (eds) (2020) Handbuch Virtualität. Springer, Wiesbaden https://doi.org/10.1007/978-3-658-16342-6

Kellmeyer P (2019a) Artificial intelligence in basic and clinical neuroscience: opportunities and ethical challenges. Neuroforum 25:241–250

Kellmeyer P (2019b) Digital vulnerability: a new challenge in the age of super-convergent technologies. Bioeth Forum 12(1/2):60–62

Kellmeyer P, Mueller O, Feingold-Polak R, Levy-Tzedek S (2018) Social robots in rehabilitation: A question of trust. Sci Robot 3(21):1–2. https://doi.org/10.1126/scirobotics.aat1587

Keymolen E (2016) Trust on the line. A philosophical exploration of trust in the networked era. Wolf, Oisterwijk

Koepke S, Denissen JJA (2012) Dynamics of identity development and separation–individuation in parent–child relationships during adolescence and emerging adulthood—A conceptual integration. Dev Rev 32:67–88

Koren Y, Feingold Polak R, Levy-Tzedek S (2022) Extended interviews with stroke patients over a long-term rehabilitation using human–robot or human–computer interactions. Int J of Soc Robotics 14(8):1893–1911. https://doi.org/10.1007/s12369-022-00909-7

Kuipers B (2022) Trust and cooperation front. Robot AI 8:130–147. https://doi.org/10.3389/frobt.2022.676767

Langer A, Feingold-Polak R, Mueller O, Kellmeyer P, Levy-Tzedek S (2019) Trust in socially assistive robots: Considerations for use in rehabilitation. Neurosci Biobehav Rev 104:231–239. https://doi.org/10.1016/j.neubiorev.2019.07.014

Latour B (2006) Über den Rückruf der ANT. In: Belliger A, Krieger D (eds) ANThology. Ein einführendes Handbuch zur Akteur-Netzwerk-Theorie. transcript, Bielefeld, pp 561–572 (engl. 1999)

Levinas E (1969) Totality and Infinity. An essay about exteriority. Duquesne University Press (Totalité et Infini: essai sur l'extériorité, 1961)

Levinas E (1992) Jenseits des Seins oder Anders als Sein geschieht. Karl Alber, Freiburg (Autrement qu'être ou au-delà de l'essence, 1974)

Lewis M, Sycara K, Walker P (2018) The role of trust in human-robot-interaction. In: Abbass H, Scholz J, Reid D (eds) Foundations of trusted autonomy. Studies in systems, decision and control, 117th edn. Springer, Cham, pp 135–159 https://doi.org/10.1007/978-3-319-64816-3_8

Loh J (2019) Roboterethik. Eine Einführung. Suhrkamp, Berlin

Loh J, Coeckelbergh M (2019) Feminist philosophy of technology. J.B. Metzler, Stuttgart

Luhmann N (1968) Vertrauen. Ein Mechanismus der Reduktion sozialer Komplexität. Enke, Stuttgart

Maisel N, Gable S, Strachman L (2008) Responsive behaviors in good times and in bad. Pers Relatsh 15:317–338

McCormack J (2008) Facing the future: evolutionary possibilities for human-machine creativity. In: Romero J, Machado P (eds) The art of artificial evolution. Springer, Berlin, pp 417–451

McLeod C, Ryman E (2020) Trust, autonomy, and the fiduciary relationship. In: Miller P, Harding M (eds) Fiduciaries and trust: ethics, politics, economics, and law. Cambridge University Press, Cambridge, pp 74–86

Misselhorn C (2021) Künstliche Intelligenz und Empathie. Vom Leben mit Emotionserkennung, Sexrobotern & Co. Reclam, Stuttgart

Möllering G (2006) Trust: reason, routine, reflexivity. Elsevier, Oxford

Montgomery T, Berns JS, Braddock CH III (2020) Transparency as a trust-building practice in physician relationships with patients. JAMA 324:2365–2366

Müller O (2022) Maschinelle Alterität. Philosophische Perspektiven auf Begegnungen mit künstlicher Intelligenz. In: Schnell M, Nehlsen L (eds) Begegnungen mit künstlicher Intelligenz, 1st edn. Velbrück Wissenschaft, Weilerswist, pp 23–47 https://doi.org/10.5771/9783748934493-23

Müller E, Zill JM, Dirmaier J, Härter M, Scholl I (2014) Assessment of trust in physician: a systematic review of measures. PLoS ONE 9:e106844

Nickel P (2022) Trust in medical artificial intelligence: a discretionary account. Ethics Inf Technol 24:7. https://doi.org/10.1007/s10676-022-09630-5

Nussbaum M (2000) Women and human development. The capabilities approach. Cambridge University Press, Cambridge

Nussbaum M (2006) Frontiers of justice: disability, nationality, species membership. Harvard University Press, Cambridge MA

Nussbaum M (2011) Creating capabilities. Harvard University Press, Cambridge MA

O'Neill O (2018) Linking trust to trustworthiness. Int J Philos Stud 26(2):293–300. https://doi.org/10.1080/09672559.2018.1454637

Pellegrini CA (2017) Trust: the keystone of the patient-physician relationship. J Am Coll Surg 224:95

Peter E, Morgan K (2001) Explorations of a trust approach for nursing ethics. Nurs Inq 8(1):3–10

Plessner H (2016) Die Stufen des Organischen und der Mensch. Einleitung in die philosophische Anthropologie. Suhrkamp, Frankfurt a.M. (Erstausgabe 1928)

Reis H (2014) Responsiveness: affective interdependence in close relationships. In: Mikulincer M, Shaver P (eds) Mechanisms of social connection: from brain to group. American Psychological Association, Washington, pp 255–271 https://doi.org/10.1037/14250-015

Reis H, Clark M (2013) Responsiveness. In: Simpson J, Campbell L (eds) The Oxford handbook of close relationships. Oxford University Press, Oxford, pp 400–423

Ridd M, Shaw A, Lewis G, Salisbury C (2009) The patient–doctor relationship: a synthesis of the qualitative literature on patients' perspectives. Br J Gen Pract 59:e116–e133

Ryan M (2020) In AI we trust: ethics, artificial intelligence, and reliability. Sci Eng Ethics 26:2749–2767. https://doi.org/10.1007/s11948-020-00228-y

Schönau A (2019) Schnittstellenprobleme in Neurowissenschaften und Philosophie. Willensfreiheit aus handlungstheoretischer Perspektive. J.B. Metzler, Stuttgart

Schütz A, Luckmann T (2003) Strukturen der Lebenswelt. UTB, Stuttgart

Siau K, Wang W (2018) Building trust in artificial intelligence, machine learning and robotics. Cut Bus Technol J 31:47–53

Starke G, Van den Brule R, Elger B, Haselager P (2021) Intentional machines: A defense of trust in medical artificial intelligence. Bioethics 36(2):154–161. https://doi.org/10.1111/bioe.12891

Sullins J (2006) When is a robot a moral agent? Mach Ethics 6:23–30

Taddeo M, Floridi L (2011) The case for e-trust. Ethics Inf Technol 13:1–3. https://doi.org/10.1007/s10676-010-9263-1

Tanaka M, Ishii A, Yamano E, Ogikubo H, Okazaki M, Kamimura K et al (2012) Effect of a human-type communication robot on cognitive function in elderly women living alone. Med Sci Monit 18(9):CR550–CR557. https://doi.org/10.12659/msm.883350

Thom D, Ribisl K, Stewart A, Luke D, The Stanford Trust Study Physicians (1999) Further validation and reliability testing of the trust in physician scale. Med Care 37(5):510–517

Vaesen K et al (2013) Artefactual norms. In: De Vries M (ed) Norms in technology. Philosophy of engineering and technology, vol 9. Springer, Dordrecht

Von Weizsäcker V (1987) Der Arzt und der Kranke. Stücke einer medizinischen Anthropologie. Gesammelte Schriften in zehn Bänden, vol 5. Suhrkamp, Berlin

Waldenfels B (1987) Ordnung im Zwielicht. Suhrkamp, Frankfurt a.M.

Waldenfels B (1994) Response und Responsivität in der Psychologie. J Psychol 2(2):71–80

Waldenfels B (1997) Topographie des Fremden. Studien zur Topographie des Fremden, vol 1. Suhrkamp, Berlin

Waldenfels B (2007) Antwortregister. Suhrkamp, Berlin

Waldenfels B (2010) Provost lecture: response and trust: some aspects of responsive ethics. Stony Brook University. https://www.youtube.com/watch?v=t6iOsQ_ho94. Accessed 22 May 2022

Waldenfels B (2011) Phenomenology of the alien (Studies in phenomenology and existential philosophy). Northwestern University Press, Evanston

Waldenfels B (2019) Erfahrung, die zur Sprache drängt. Studien zur Psychoanalyse und Psychotherapie aus phänomenologischer Sicht. Suhrkamp, Berlin

Waldenfels B (2020) Care of the self and care of the other. In: Voman F, Nortvedt P (eds) Care ethics and phenomenology. A contested kinship. Peeters, Leuven

Waldenfels B (2021) Das leibliche Selbst. Vorlesungen zur Phänomenologie des Leibes. Suhrkamp, Frankfurt a.M.

Whelan S, Murphy K, Barrett E et al (2018) Factors affecting the acceptability of social robots by older adults including people with dementia or cognitive impairment: a literature review. Int J Soc Robotics 10:643–668. https://doi.org/10.1007/s12369-018-0471-x