Techné: Research in Philosophy and Technology ISSN: 2691-5928
26:1 (2022): 1–30 DOI: 10.5840/techne2022523158

What Does It Mean for a Robot to Be Respectful?

Dina Babushkina

Abstract: Intelligent systems are increasingly incorporated into relationships that had, until recently, been reserved solely for humans, and are delegated the role of a partner, which, if human, would presuppose a system of normatively regulated interactivity. This includes expectations of reciprocity and certain attitudes/actions towards human actors, such as respect. Even though a robot cannot respect, I argue that it can be respectful. A robot can be attributed respectfulness (in the direct sense) iff its interactions with persons reflect the respectful attitude of the humans involved in its design and operation. Robot respectfulness is a compound of (a) robotic actions governed by principles that (b) reflect the attitude of respect for persons by humans involved in its design, implementation, and professional use. I define respect for persons as a commitment to core values that make someone a person (i.e., intellect, rationality of reactive attitudes, autonomy, personal integrity, and trust in expertise).

Key words: respect, robot respectfulness, artificial intelligence, disrespect, responsibility, trust

1. The Problem with Robot Respect

Despite the insufficient discussion of robot respect in the literature, it is an issue that is gradually emerging (and should emerge), especially in the field of human-robot interaction (see, e.g., Van Kleek et al. 2018; Seymour and Van Kleek 2019).

Dina Babushkina, Assistant Professor, Faculty of Behavioural, Management and Social Sciences (BMS), Section of Philosophy (WIJSB), University of Twente, P.O. Box 217, 7500 AE Enschede, Netherlands; d.babushkina@utwente.nl; Visiting Scholar, Faculty of Social Sciences, Practical Philosophy (the RADAR group), University of Helsinki, P.O. Box 24, 00014, Helsinki, Finland.

The widespread tendency to anthropomorphise technology, on the one hand, and the trend of fine-tuning social robots and "smart" devices to the emotional needs of human users,[1] on the other, create a demand for respectful technology (and, as a result, the need to conceptualize respectfulness of artifacts and non-human agents), in the same manner that they create the need for trustworthy and responsible technology (and by extension, the need for the concepts of trustworthiness of and responsibility for artifacts and non-human agents). Each of these is a proper response to the expectations that we have from the parties involved in *interpersonal relationships*, that is, relationships that, thus far, have been reserved for humans, such as friendship, companionship[2] (e.g., Wilks 2010), love and sexual relationships (e.g., Levy 2007; Danaher and McArthur 2017; Jason 2017), care and therapy (e.g., Wada and Shibata 2007; Turkle 2004; Turkle et al. 2006; Richardson et al. 2018), and pedagogical relationships (e.g., Kim 2005; Veletsianos and Miller 2008). Robots and AI are quickly being introduced and absorbed into these relationships. Normally, we expect those with whom we engage in these kinds of relationships to be trustworthy in some sense: we expect them to be responsible for their actions toward us, and we expect them to respect us. The problem with robotics and AI is that, as they enter these relationships, they become players in these relationships (playing the role of partner, caregiver, therapist, etc.), without being able to respond to these demands and expectations in the way humans can. This leads many to reflect on the way the introduction of computer technology, including robotics, is transforming our relationships with each other (see, e.g., Turkle 1985, 2011; Richardson 2015, 2019; Seibt, Nørskov, and Andersen 2016). In the end, we have to rethink the very nature of the relationships that we can have with machines.

In general, when it comes to respect in the interaction between a human and a robot, there are two general lines of inquiry open for discussion:[3]

(a) When respect is viewed as an active stance (i.e., when a robot is thought of as an entity giving respect to others): Are we justified in attributing respectful attitudes to a robot? In want sense can a robot be said to be respectful? When demanding that robots are respectful, what are the rational limitations of such a demand? How can the respectfulness of robots be achieved?

(b) When respect is understood as a passive stance (i.e., when a robot is seen as a recipient of respect): In what sense can a robot be said to be worthy of respect? Is there anything more to the respectful treatment of a robot

> than a proper maintenance of a tool? What are the rational constraints on a respectful attitude of a human towards a robot?

The second line of inquiry has less urgent practical relevance and is most interesting to those working on the question of rights for artificial agents. This article is concerned solely with the first line and aims at outlining a general account of what can be called *artificial respect*. Here, this concept is meant to capture a *mode* of acting by an artificial agent (such as a robot) towards a person. This mode implies that, were the person treated this way by another human agent, it would justifiably evoke in its recipient the feeling of having been treated in a due manner.[4] The problems with artificial respect start with the very attempt to attribute respect to an artificial agent. In the literature, there is often an implicit assumption that a person is the only possible subject of respect. Robin S. Dillon, for instance, sums up the assumption this way:

> While a very wide variety of things can be appropriate objects of one kind of respect or another, the subject of respect (the respecter) is always a person, that is, a conscious rational being capable of recognizing and acknowledging things, of self-consciously and intentionally responding to them, of having and expressing values with regard to them, and of being accountable for disrespecting or failing to respect them. (Dillon 2018)

This restriction on the subject is implicit in the very concept of respect, which seems to involve a bundle of various cognitive attitudes (at least, a belief about respect-worthiness of someone/something) and affective attitudes (emotive states produced by such belief). Within this framework, robots and AI, no matter how complex they are, are excluded from the domain of respecting beings. The appeal of this position is understandable. At least in the current state of technological development, robots cannot feel, take a stance, or have beliefs. We still can talk about "respectful" robots in the same way that we talk about "loving" or "caring" robots. But this is meant in an indirect sense, as a figure of speech or metaphor. The phrase "a respectful robot" (when this is used as a metaphor) *necessarily* comes with the implicit clause saying that respectfulness cannot be attributed to the robot (*in the direct sense*). So, the meaning of "a respectful robot" (in the inverted commas) is something like this: a robot, that is described with the adjective respectful, which adjective cannot be applied to it. Simply said, the expression in the inverted commas is an attribution mistake. This mistake contributes to the ambiguity of popular and marketing narratives about robots, but it is still frequently overlooked.

The question this article aims to tackle is how we can talk about respect in its application to robots without the inverted commas. This question requires a closer look at the concept of respectfulness, and the ways in which it can be manifested. This boils down to asking: *What does it mean for a robot to be respectful*? How can we conceptualize respectfulness as an attribute of an inanimate artifact in such a way that we (a) will not lose anything substantial from the phenomenon of respect itself, but at the same time (b) acknowledge the unique ontological status of AI? Framing the problem this way, I insist on the importance of maintaining the distinction between humans and artificial agents. The goal is to give the concept of artificial respect its own meaning and thus to avoid feeding ambiguity in the narratives about robots. Removing the inverted commas from "a respectful robot" does not consist in borrowing the adjective from the phrase *respectful human*; it does not, therefore, amount to blurring the distinction between the two types of agents.[5] *In such an account*, no jump can be made from calling a robot respectful (without inverted commas) to any anthropomorphic claim.

If they are not strong AI[6] (and it is at least dubious if and when strong AI can be achieved), robots are not capable of taking any stance, let alone *feeling respect*. But this does not mean that robots cannot *be respectful*.[7] I further suggest that they can be respectful in a way that is different from the way humans can be respectful. For a robot to be respectful, it means that its algorithms and hardware are designed according to the principles of respect, which subsequently manifest in its activities and interactions with its users. This requires re-engineering the concept of respectfulness within the domain of robotics. The task is, then, to understand:

a) To what extent are certain cognitive capacities (such as beliefs, acknowledgements, judgements, commitments) and emotive capacities (emotions and feelings) necessary for someone to be considered respectful? Such capacities will have to be reserved for people responsible for the crucial stages of the process that leads to the robot's functioning in an interpersonal relationship, such as the design and development of robots (e.g., programmers and engineers), its integration into society, and its professional use (e.g., medical staff or care-providers employing a robot in their practice).

b) Which part of being respectful can and should be played by robots, and in which way?

The way we can ascribe respectfulness to robots is different from the way we can ascribe it to persons, and there is nothing wrong with that—as long as we know what it consists in and can form our expectations appropriately, that is.

2. On Methodology: Conceptualizing Artificial Respectfulness

One possible strategy would be—following the general Human-Robot Interaction (HRI) approach—to look at the user's preferences. This would require an empirical study of people's beliefs about what constitutes a respectful treatment, followed by a search for a way to align the design of a robot with these preferences. This approach can succeed making robots more likeable and accepted, lowering the threshold of resistance to their use. The problem is, however, that this does not make robots ethical. The empirical approach is only able to identify what people in a given cultural context count as respectful behavior (descriptive level). As a result, robots are made compliant with the existing social requirements, which, even though commonly confused with ethical norms, neither equal nor warrant them (normative level).[8] The truth is that social norms may encode morally reprehensible ideals and requirements. Allowing the design process to be guided solely by social norms runs the risk of moral transgressions, harm, and injustice. Incorporating user preferences into the design of a robot should always be guided by ethical principles. The goal of this article is to identify ethical constraints on the concept of respectfulness, that is to say, to lay out an account of respect that will not depend on any specific cultural or social framework but will instead be rationally justified.[9] This account can further be used as a guiding principle for future policies regarding robotics.

Given the interdisciplinary audience of this journal, a further note on the methodology used in this article is needed.[10] Like most works in philosophy, this article engages in elements of the clarification of concepts and evaluation of the soundness of arguments. Even though most philosophers agree that the methodological basis for these tasks is logic and principles of rationality (cf. Rescher 2001), they can be carried out in a variety of ways (cf. Tugendhat 1976, 3–4). This is so, as James Chase and Jack Reynolds argue, because different philosophical problems dictate different "methodological commitments" and necessitate "methodological flexibility" (2014, 56).[11] This article belongs to the Analytic tradition[12] and draws from several of its most common methods:[13] (a) conceptual analysis (which Roy Sorensen describes as consisting of "definition, question delegation, drawing distinctions, crafting adequacy conditions, teasing out entailments, advancing possibility proofs, mapping inference patterns" (1992, 15)); (b) linguistic analysis

(e.g., work on semantic, truth-conditional analyses); (c) defining constraints based on various normative systems (such as scientific principles, principles of rationality); and d) evaluation of the relevance of arguments and their premises.

One possible worry concerning the analysis of the term "respect" in this article is that it may appear to over-rely on the everyday use of the term. Let me first remark that even though it is true that ordinary language analysis (Chappell 1964)—as opposed to ideal language analysis (e.g., early L. Wittgenstein, Russell)—has its shortcomings, it is a widely used philosophical method (e.g., by later Wittgenstein, A. A. Ambrose, G. Ryle, J. L. Austin, H. L. A. Hart, P. F. Strawson, J. Searle) practiced after the "linguistic turn" in the early 20th century (cf. Baldwin 1998; Hochberg 2003; Glock and Kalhat 2018), which signified the beginning of the analytic tradition itself. However, this article *is not* engaged in the ordinary language analysis. The goal is to lay down a foundation of an account of artificial respect, and the method used in this article comes the closest to what in currently described as *conceptual engineering*.[14] Conceptual engineering consists in identifying what a concept should mean given certain constraints (Justus 2012; Shepherd and Justus 2015; Isaac 2020; Brun 2020; Burgess, Cappelen, and Plunkett 2020). The idea here is not to develop a definition—as is done in the Aristotelian tradition, in terms of necessary and sufficient conditions—but to *explicate* the meaning of a term (the explicandum) by providing another concept (the explicatum) which is more exact and illuminating, given the specific theoretical purpose (cf. Brun 2016). A part of this work is distinguishing the explicandum from similar concepts often confused with it and positioning it in the system of related concepts. I will show that, when developing a concept of respect suitable for the ethics of technology, respect should not be confused with subjection to authority, accommodation of the wishes of others, admiration, or care. After that, I will move on to a discussion of disrespect, in order to lay out my own account of an ethically justifiable concept of robot respect.[15]

3. What Types of Respect Are We Not Ethically Justified to Expect from a Robot?

There are a number of ways to explicate the concept of respect, depending on what type of attitude it is taken to represent. Most often, perhaps, it is used in the sense of *respect for authority*, that is, a respect for someone/something due to their superior position, for example, with respect for power. Respect for one's parents (citing the reason that they gave birth to you), for the elderly (because they are more experienced), and for the police (as a law-enforcing institution) are all constitute

examples of the respect for authority (Stephen Hudson's (1980) directive and institutional forms of respect come close to this meaning). This type of respect can further be analyzed in terms of the concept of compliance with authority, which is characterized by the acceptance of coercion and implies a submissive stance. This type of respect is often motivated by fear (as highlighted by Joel Feinberg (1975) with his concept of Respekt). When William Seymour and Max Van Kleek (2019) talk about *directive respect* as "the process of adhering to explicit rules and directives," they come closest to this concept of respect-as-compliance. However, when the same concept is used by authors to mean that systems should "satisf[y] safety and regulatory requirements," it is a matter of moral responsibility and not of respect. When, again, the authors equate their directive respect with the imperative that devices should "[follow] preferences and commands expressed by users during the lifetime of the device," this type of respect turns out to be an ethically problematic concept, primarily because, when applied to robots, it takes us in the direction of "robots are slaves" paradigm.[16] It is not necessary for robots to be respectful of humans in this way, at least for the reason that it is morally wrong to aim at creating slaves. The concept of artificial respect should be more ethically sound than this: being respectful in a moral sense cannot entail servitude or fear.

Another frequent use of respect is in the sense of *respect for someone's opinion, point of view, or wishes*. In a trivial sense, when respect is taken as implying that a piece of advice should be unquestionably followed, that a wish should be satisfied no matter what, or that a point of view should be given weight despite the lack of good reasons, this form of respect collapses into respect for authority. In a less trivial sense, to respect one's point of view means to take it into account in the decision-making process. This does not mean that, were you to express your opinion or wish, we would always do what you want. It does mean, however, that we will do our best to properly consider your point of view and/or to accommodate your wish. Your opinion or wish will be taken into account in what will constitute an all-things-considered outcome. This meaning of respect has an advantage because it is a way of being fair. This understanding of respect is reflected in the concept of *obstacle respect* in Seymour and Van Kleek (2019). With obstacle respect, the authors aim to grasp the imperative for a device to give up its goal if it interferes with the preferences of the user. This aspect is especially relevant to persuasive technology. Demanding that a robot be respectful in this sense would amount to requiring it to prioritize the user's wishes and preferences. The advantage to this approach is that it limits paternalistic treatment of the user and helps to prevent some forms of abuse. This type of respect is, however, too

narrow when it comes to understanding the type of behavior we should expect from robots as players in interpersonal relationships. Moreover, when it comes to accommodating users' wishes, desires, and preferences—all of which are subjective attitudes, having notoriously little ground for objectivity—this should be a part of the design of robot's functionality and/or preference settings, rather than a question of a respectful attitude. A respectful robot should be able to exhibit a much stronger sense of respectfulness, such that it is not a matter of changing a setting or opting out.

Yet another concept of respect is referred to as appraisal by Stephen Darwall, evaluative respect by Hudson (1980), and consideration respect by William Frankena (1986) and Carl Cranor (1982, 1983). This concept aims to reflect the aesthetic attitude of *admiration* (a pleasurable contemplation) of something or someone, for example, for her achievements or character traits (cf. Zagzebski 2006, 2015).[17] An extreme expression of such an attitude is "owe,"[18] which is close to Feinberg's (1975) Reverentia.[19] With the development of robotic companions (incl. robotic pets), it is easy to foresee a higher demand for a variety of emotional attitudes similar to those expressed by persons and animals, including admiration. The motivation for this is the feeling of satisfaction and acknowledgement that being admired gives (say, by a friend or a caregiver, due to strength in the face of misfortunes or for personal achievements, or even by a dog, in the form of devotion to its master). With robots substituting for those who give us the satisfaction of being admired, it is natural that this sort of respectful attitude will be expected from robots as well. However, it is hard to imagine how to implement this sort of attitude in the design of a robot. One option would be some sort of reward function activated each time the robot is in the presence of a human, or when its algorithm is operating with variables representing the human as an object. But even if the designers succeed in recreating a state of admiration in a robot, the question is whether the demand for such an attitude from a respectful robot can be well justified. Admiration is not necessary for respect: one could respect (i.e., highly esteem) Napoleon Bonaparte for his military skills, but derive no pleasure from contemplating him. The act of admiration idolizes its object and, therefore, does more than treat the object in the way it deserves—in other words, admiration overreaches the attitudes characteristic of respect (cf. van der Rijt 2018). As a result, to demand admiration is supererogatory.

It is equally hard to ethically justify artificial respect in terms of *care*. *Care respect* (for more on the concept, see Dillon 1992) is used by Seymour and Van

Kleek (2019) to refer to the simulation of love and of concern for well-being of the user by a device:

> [T]his also includes situations where someone or something takes an action that goes against the short term wishes or interests of someone else in order to promote their long term welfare… Examples include providing clear stopping points when users spend large amounts of time using services such as video streaming, or limiting how much money can be spent via micro-transactions. (Seymour and Van Kleek 2019)

Putting aside the question whether these are adequate expressions of care based on love, as far as the ethically sound concept of respect is concerned, care is not a part of respect. It is possible to care for someone without respecting that person, and it is possible to respect someone without having any contribution to the person's wellbeing. Thus, it is not necessary for a respectful robot to care for its user, neither is it sufficient for a robot to provide care to the user in order to be considered respectful. Whether or not a robot is to provide care services to the user is a matter of its purpose and depends on its target audience and the goals for which it has been designed. Respectfulness should not depend on specific design goals; it has to be present in any robot which, in one way or another, is a player in an interpersonal relationship.

There is another concept of respect, which comes closest to the account this article is developing, and that is *recognition respect* (Darwall 1977). The idea here is that to respect someone is to give appropriate consideration to some property or characteristic of that person. One example of this type of respect is *the respect for someone's decision*. Even when I disagree with you, I still may respect your decision. To say that I respect your decision is to say that I recognize the weight of your reasons; that is, I recognize you as a fully rational agent who is capable of reasoning, making decisions, and taking responsibility. This shows that a respectful attitude does not imply agreement.[20] Quite the opposite, to disagree and argue otherwise might be seen as an act of respect for the intelligence of another person, for example, in the case when she is mistaken.[21] Entering a disagreement in such a case would presuppose a belief that the other person is capable of changing her opinion based on good reasons and evidence. This signals the recognition of the intelligence of the conversation partner. A respectful robot would behave in such a way that, when playing its part in an interpersonal relationship, in one way or another, it is also playing the part in this type of recognition.

Seymour and Van Kleek (2019) acknowledged this type of respect when suggesting that a system's ability to recognize that the user deserves a certain treatment due to certain of her aspects should play a crucial role in the design of a respectful device. The problem starts when we try to define what sort of user characteristics the system must take into consideration in order to be seen as respectful in the ethical sense (cf. Hill 1997). For example, incorporating such functions as helping the user to comply with her religious beliefs (Woodruff, Augustin, and Foucault 2007) is not a matter of ethics, but rather of religious and cultural norms. By contrast, avoiding the exploitation of a user's personal information for data mining and marketing is a clear case of an ethical matter. In what follows, I will suggest which characteristics of the user are ethically relevant for the concept of the recognition respect as it can be ascribed to robots.

4. Towards an Account of Artificial Respect: Disrespectfulness

One way to understand what being respectful entails is by conceptualizing the breaches that an act of disrespect commits. Disrespect is not the same as the absence of respect, which by itself may be a neutral attitude to a person or thing. Disrespect is a negatively charged attitude, opposite to that of a respectful one. This attitude aims to deprive its object of what it is due or of some sort of privilege that a respectful attitude aims to recognize.

What is common of most responses towards a disrespectful attitude is the feeling of de-valuation, in one form or another. This may include a feeling that one's professional qualifications were disregarded, or that one has been treated as a child, or as an object, or that one's worth has been diminished due to gender, nationality, or race. Paternalistic attitudes are based on a form of disrespect, i.e., a de-valuation of a person's rationality and autonomy. Bullshitting[22] someone is a form of disrespect, because it de-values a person's intelligence and relation to truth. There is, furthermore, *a specific sense of disrespect* when a person is justified in feeling insulted for an idiosyncratic reason, such as when she was deprived of a certain treatment which she deserves due to her achievement, status, or social role. And then there is *a general sense* in which any person can be disrespected by virtue of her having dignity. In this case, she has been subjected to a treatment that diminishes her worth as a person. This is an important distinction because, while the first is about a loss of preferential treatment, the second is about the loss of a fundamental right. This gives the *general* sense of respect its special weight.

Not all felt disrespect is justified; that is to say, it is not necessarily true that, if one feels disrespected by A, that A has been disrespectful. People may feel disre-

spected due to various subjective considerations which are not necessarily true or well-grounded. This is not to deny that it matters how actions or attitudes of others make us feel, but rather to claim that these feelings *alone* are not good grounds for considering someone or something disrespectful. For example, parents may feel disrespected because their daughter married against their wishes. But that feeling in itself is not sufficient to render the daughter's actions disrespectful from a rational point of view. The parent's feelings are motivated by frustrated behavioral expectations which they deemed appropriate under a paternalistic power-paradigm. Her actions seem not to be appropriate towards idealized authority figures of the elderly, and therefore they are seen as a threat to that authority, as ignoring and undermining their superiority. By itself, the woman's choice of a spouse cannot de-value the humanity, worth, rationality, or intelligence of her parents.[23] This distinction is fruitful because it highlights the importance of *objective criteria* for respectful robotics, which lie outside the subjective feedback of a user.

Feelings of disrespect can also be misplaced. Take the example of an advertisement that reduces an image of a woman to that of a sexual object. While feelings of disrespect are fully warranted in this case, it does not make sense to direct blame for these feelings to the ad itself. The ad may be disrespectful, but it does not disrespect because it is not capable of intentional states. The blame is correctly placed at the agent whose motivational states resulted in the disrespectful act. In this case, the guilty party is to be found among humans, among those who designed, approved, and capitalized on the ad. This latter consideration brings us back to the starting point: if a robot is an artifact, does that not imply that any attempt to attribute disrespectfulness to it will just miss the target? I do not think that it necessarily does. There is a way to think about robot respectfulness in a non-contradictory way.

5. Two Sides of Respectfulness

The solution lies in the fact that respectfulness has two aspects: attitudinal and behavioral (cf. Dillon 2018). The former is respect as a stance or a standing attitude to someone/something in response to her/it having a certain value. It is characterized, furthermore, by the presence of certain psychological states. Allen Wood calls respect "the appropriate attitude toward any objective value making a valid claim on us" (2010, 562). The latter are actions through which this standing attitude is manifested. Respectful behavior is about (a) performing (or being predisposed to) actions that are in some way fitting for or deserved by the object of respect and (b) refraining from actions that are unfitting for or un-deserved by the object.

When it comes to persons, it is the stance or the standing attitude that, arguably, has the primary importance because it shapes a person's intentions and motivations; actions are then seen as the expression of these intentions and the resolution of motivations. When it comes to robots, from the perspective of a user, it is the action component that has the primary importance because of the nature of her experience: it is the robot that she experiences through interaction, not the human who had designed and programmed it. This is not to say that intentions should not be considered. They are an integral part of respectful technology. But the user does not have access to those intentions; they are removed from her experience and mediated through technology. To be sure, we are not concerned here with the relationship between the two components of respect as such; some philosophers claim that these components are two different types of respect and that the behavioral component it not necessary. Be that as it may, when it comes to robotic respectfulness, there cannot be one component without the other. Here is a preliminary definition of robot respectfulness:

> Robot respectfulness is a case of *mediated respectfulness* or *respectfulness by design*. By this I mean that a robot can only be attributed respectfulness in a proper sense if its interactions with persons reflect the respectful attitude of the humans involved in its design and operation.

Respectfulness as an attribute of robotic actions is thus not an isolated phenomenon. It cannot be separated from the intentions of humans involved in its design, production, and professional use. At the same time, this attitude is not experienced by the user in an immediate way, as happens in the case of a direct human-to-human interaction (human intention → human action). It is mediated through the design of the systems that determine the robotic choice of actions: human/robot-to-human interaction (human intention → robot action).[24]

Robot respectfulness consists of two elements:

a) robotic actions (or other forms of operations) that are determined by principles that

b) reflect the corresponding standing attitude of respect by humans who are involved in its design, implementation, and professional use.

By applying stance-respect to robots, I understand a commitment or set of commitments on the part of designers, implementers, and users *to apply well-justified principles of treating others based on certain value assumptions*. At this point, let us call these commitments *the considerations of respect*. These consid-

erations or principles could, for example, take the form of a professional code. Action-respect can then be understood as the application of these principles to the actual design, implementation, and use of robots as service-providers or as players in interpersonal relationships. For a robot, this implies that it is respectful when its functions (actions, elements of design, narrative that they carry, etc.) in its interaction with humans are determined by the considerations of respect. This also implies that these respectful robotic functions reflect a commitment to such treatment by the humans involved in all stages of robots' development and integration into society. I call robot-respectfulness a mediated respectfulness or respectfulness-by-design because it reflects and instantiates *attitudes of humans toward humans*, that is, the attitudes of professionals involved in robotics toward the users-benefactors of robotic services. In other words, the design must acknowledge that which the robot must implement in its actions.

6. Respectfulness as an Attitude

How a robot, as an actor in interpersonal relationships, will act depends on the consideration of respect as the principles guiding the design, implementation, and use of the robot. This stance-respect, as a standing attitude of the humans responsible for the design, implementation, and professional use of robots, has three elements: (a) recognition, (b) approval, and (c) commitment.

Respect signals *recognition* of special needs, rights, and treatments that someone or something deserves, due to a certain property or feature that gives her value.[25] In this context, recognition is understood an act of acknowledgment.[26] There is an important sense in which a certain phenomenon is only possible because we acknowledge it; the phenomenon is created and sustained by this act. Even though recognition is important for respect, it is not the same thing as respect. One cannot have a respectful attitude toward someone/something without acknowledging her/its special value and her/its right to due treatment. But recognition does not necessarily entail respect. For instance, one could recognize a church, in the sense of acknowledging that it has a right to operate in a given society. But one may have no respect for it as an institution. On the other hand, disrespect does not imply the withdrawal or denial of recognition; that is, it does not imply a failure to acknowledge that one has a certain value or status or right. Rather, it can take the form of pure humiliation, sending a message of the sort: "yes you are entitled, but I have the power to take it away from you, I can deprive you of this due treatment, preventing you from having access to it."

What this tells us is that recognition is not enough to constitute respect. One thing that is missing is *a commitment* to providing due treatment. Moreover, one may obey the law but not respect it (for example, thinking it to be unjust). What this shows is that respect presupposes a certain degree of personal evaluative involvement; that is, one must also *accept* the value of a person or thing in order to behave respectfully towards it. In the case of human-to-human interaction, acceptance and commitment presuppose one's readiness to act (a disposition to act) in a way that the object of respect deserves (and an abstention from acting in an undue way). However, in the case of human/robot-to-human interaction this must be understood as the robot's *preparedness to act* in that way, in other words, the robot's decision-making system being pre-determined by such considerations. When it comes to social robotics, stance-respect does not count for much unless it has been transformed from the motivation of a human to the action-guiding principles of machine behavior. In the end, we do not expect of a robot companion to merely store principles of respect in its memory or user-manual; we expect that our interaction with it happens in a certain modality.

Preparedness to act in accordance with the considerations of respect is the guarantee that the user will receive due treatment which is, in turn, a way to highlight and venerate a certain value that she has or represents, due to her actions or cause.

7. The Question of Value

It is important to keep in mind that respect is a normative *but not an ethical* concept. It is normative for two reasons. First, it is about the way we should treat other people, but it is not necessarily based on ethical norms. Respect as an acceptance of coercion is one example of non-ethical respect. Second, what you respect tells what you value. If you respect someone for her cause, that means that you approve of or value that cause; if you respect a murderer for daring to kill, you most likely value courage over life; if you are respectful of authoritative figures, this implies that you are committed to the value of power; and so on.

This is why it is not enough to find out how to make robots respectful. We need to ask: respectful of what? We need to decide what types of respect are worth being (and ought to be) implemented in robot design. Not all types of respect are agreeable and desirable. We definitely do not want to multiply cases of ethically questionable respect, such as the respect for criminals and mass murderers. We should exclude any *uncritical* inclusion of culturally justified forms of respect, such as respect for the elderly. Such forms of respect may require, for example, a

robot to be quieter in the presence of the elderly, or to overrule other considerations in favor of their opinion. It is also questionable whether it is worth predisposing a robot to nonsensical respect, such as respect for being able to drink large quantities of alcohol or being able to withstand extreme sauna heat.

To put it shortly, respect presupposes value-veneration, but not all values are ethical. So the question is: which values do we have solid ethical reasons to venerate? This requires a revision of my working definition of respectfulness and opens up a question about the nature of commitments that constitute attitude-respect. The very fact that we have to ask this question shows that positions such as "everything goes" or "it depends on perspective" are not satisfactory. What value ought we to respect so that respect becomes *an ethical category*, i.e., so that it can be applied to everyone in the same manner?[27]

The answer by itself is not new—it is about the respect for humanity;[28] in other words, it is the respect for persons.[29] That is, it concerns recognizing the value of humanity and committing to treating human beings in the way that they deserve and that does not diminish their humanity. This is the only way to maximize the universal applicability of the attitude of respect. But how are we to interpret this principle in its application to robot respectfulness?

8. Considerations of Respect for Humanity

In this approach, a robot can be said to be respectful if *its functionality in the interactions with a human is determined by the considerations of respect for humanity*. In very general terms, when it comes to the development, implementation, and use of technology, to occupy this perspective means to shift priorities from reducing the potential user merely to a consumer towards a view of the user as a person. When the value of the user is reduced to her value as a consumer, it is her readiness to buy (keep buying) a certain product that guides the design, functionality, and goals of a robot as a product. The robot is expected to be appealing, useful, amusing, easy, and pleasant to deal with. Its qualities and functions are expected to answer desires and wishes of the demand-supplier.

When the value of the user is primarily identified as that of a person, different considerations become decisive in the design decisions. (This does not mean that the consumer's perspective should be excluded from the design. What is meant here is rather that, if there is a conflict between the two considerations, the priority must be given to the latter.) Given that X is what makes one a person, robot actions must not violate or undermine X when interacting with humans or otherwise via its actions. This article does not aim at developing a complete account of person-

hood. The goal is, instead, to attract attention to the minimum requirements that are the most practically relevant today. A respectful robot must not undermine, at least, these attributes of a person:

- Intellect
- Rationality of emotional reactions and reactive attitudes
- Autonomy (when applicable)
- Personal integrity
- Trust in expertise

Each of these are characteristics that a person has by virtue of being a person, and each of them is a valuable feature, on the basis of which each person is due a certain treatment by another person. The concept of intellect in this context does not refer to any unique skill that one has to train for. It simply means the ability to see facts from illusions and truth from falsity on the basis of evidence and reasons as well as the ability, when properly informed, to evaluate these reasons based on logic. Actions that deprive a person of the value of her intellect are those that deny her the right to use her intellect, create circumstances in which her appeal to reason is disregarded, ignored, or unjustifiably contested, or in which she is unable to use it altogether. Such actions would manipulate a person to live in a world with blurred distinctions between reality and fantasy, de-valued truth, and blurred boundaries between various phenomena and ideas.

As principles of rationality guide one's cognitive attitude toward the world, they also guide our emotional responses, including our responses towards the actions and attitudes of others towards us. These determine when such reactive attitudes as blame, anger, joy, etc. are appropriate, correctly directed, and expressed. The significance of emotional life is often unjustly overlooked in places other than therapeutic practice and discourse. It is a fact, however, that attitudes and treatment that inhibit one's emotional function and awareness—i.e., deprive one of the ability to appropriately express emotions or distort one's ability to appropriately place them and receive feedback on them—devalue one's affective experience. Such attitudes and treatment, in short, deprive a person of her productive and natural relationship with her own body.

Autonomy is taken here to mean the ability of a person to decide, based on reasons, upon a course of action and upon things that constitute his/her own good. Of course, there are cases when persons do not have such an ability, have lost it, or have not yet developed it. These are special cases and have to be analyzed

separately. Here we are interested in such attitudes and treatment that ignore or diminish autonomy in those who are capable of it. To deny one's autonomy is to deny one's subjectivity and ownership of one's action.

De-valuing intellect, emotional rationality, and autonomy threaten a person's integrity. Here, integrity is taken as a physiological term:[30] that is, as the experienced unity of one's self. Inhibited conditions for exercising one's ability to reason lead to cognitive dissonance and a confused relationship between one's self, the world (facts), and states of mind (illusions, desires, goals, emotional states). Depriving the person of the ability to express and appropriately place her reactive attitudes leads to a distorted relationship between her self, her body, and the actions of others. This results in emotional confusion, a feeling of absurdness, pathological emotional development, and psychosomatic disorders. Denying a person autonomy and ownership of her actions creates a feeling of helplessness, meaninglessness, and alienation from one's life. It also makes one vulnerable to manipulation. Preservation of integrity is our fundamental right and a question of mental health.

And, finally, there is trust. Trust has no necessary connection to integrity. It is, however, fundamental for mutual dependence, which lies at the core of our social relationships. We depend on each other as representatives of various professions. Despite the fact that this phenomenon may sometimes be wrongly associated with privileged access to certain knowledge, it is connected to the division of labor-intensive knowledge, which allows each of us to enjoy the benefits of many skills (learned by others) without the need to devote time and effort to learning them ourselves. Trust in professions is about reliance on the good will of others towards you. In our case, it is a reliance on the competence of professionals producing and employing robots, and on their good will towards the user-consumer.

All of these considerations are important for our ability to form realistic and reasonable expectations of oneself, others, and the world, based on facts, reason, and norms governing interpersonal relationships.

9. Putting Respect into Robotic Practice

How can a robot be respectful of persons in a way that acknowledges and venerates these values? The answer is: through (a) its designer/implementor/professional user recognizing and being bound by rules protecting these values and (b) their sharing of this commitment with robots by provisioning a deterministic component in their algorithms that will ensure that robots act in a way that does not insult

or degrade the humanity of its user. The human component of robot respectfulness must, thus:

- Eliminate from the design of robots, from the conditions of their implementation and professional use, assumptions that violate the considerations of respect for persons
- Take precautions such that no deception or misinformation about a robot's capabilities, functionality, or role take place, so that technology does not create false expectations in consumers. This implies striving for transparency, explainability, and openness
- Never count on users not being able to know or understand certain processes due to their not being experts in the field
- Ensure that the technology will not exploit people's weaknesses (such as emotional, e.g., loneliness, idiosyncratic sexuality, and intellectual, such as pre-disposition to rely on authoritative opinions).

The robotic component of robot respectfulness:

- The way a robot is presented to the users, the way it functions and is used by professionals must not inhibit a person's ability to reason and see facts, to judge the truth, and/or to evaluate the robot's decisions and choices, if she chooses to. For example, when a robot or an algorithm is used for diagnosis, the patient must have an opportunity to make her judgement about the diagnosis, evaluate the way the diagnosis was achieved, and have the freedom to obtain knowledge sufficient for forming such an opinion. In other words, robots must function in such a way that there is room for negotiation and dialogue, if the user deems it necessary or if there is a reason to doubt and verify its calculations. No latent appeals to the "machine knows better so you don't need to understand yourself" or "you cannot understand this because you are not a professional" are acceptable. Anyone can understand any piece of knowledge, given an opportunity, access to the information, and proper exercise of their cognitive abilities.
- When used as a substitute for a human in interpersonal relationships, the robot must not deceive the user about its ability to reciprocate, especially given the known tendency to anthropomorphize non-human beings.[31] There is a niche for robots in society even without the illusion of reciprocity. The robot must communicate to its user its limited role in their relationship.

- The way a robot is presented to the users, the way it functions and is used by professionals must not inhibit autonomy in the individuals capable of it (cf. Düwell (2017), who argues that technology must be developed in a way that protects one's authority to govern one's own life which, for the author, amounts to dignity). This means a person must have a right to refuse to be treated, aided, or otherwise interact with the robot and conditions must be met for her to be able to exercise informed consent. The latter implies that she must have access to a description of the robot's functionality and its limitations. This is important for the user to be able to form realistic expectations from her interaction with the robot. Furthermore, the user must have access to parameters and considerations relevant to decisions that a robot makes on her behalf and be able to disagree or refuse treatment.

10. Why Should Robots Be Respectful of Personhood?

So far it has been argued that it is not necessary to make all robots respectful in any other way than following the considerations of the respect for humanity.[32] We have good reason[33] to insist on making our robots respectful in this specific sense *if they continue to be introduced as players in interpersonal relationships*. If we do not make them respectful of personhood, we run the risk that the integration of robots into the sphere of relationships which have always been reserved for humans will undermine our practical identity. Their use will contribute to the pathological distortion of personalities.[34] This is not meant in a general way, as regards the human species. This is meant very specifically in terms of the mental and psychological health and well-being of individuals.

The nature of interpersonal relationships is such that they create conditions for (a) building (via recognition[35]) and (b) maintaining (re-assurance, various social scripts, norms governing these relationships) one's identity and relationship to oneself. Practical identity is not easy to define, but it includes a variety of cognitive and emotive states, such as: beliefs about one's social roles, status, expectations that people have from each other as players in interpersonal relationships, considerations of action-guiding restrictions and requirements, conclusions about certain mental states, choices, and actions related to the ideal vision of one's self, and the effect that the approval of others has on this ideal. The entity who occupies the other end of an interpersonal relationship (be it a human or a robot) holds a certain power over my practical identity. When the attitude of this entity is motivated by the considerations of respect for humanity, this relationship helps me to maintain

a positive and constructive relationship with myself. When it is disrespectful, that is, in some sense depriving me of the recognition of certain claims to identity or treatment that I deserve by virtue of such claims, it has a potential to "culturally downgrade" my "form of living" and collapse my entire person, leading to personal degradation (Honneth 1992, 192). Apart from the metaphysical and moral significance of these effects on a person, one has to consider the psychological costs to the individual.

The consequences of disrespect are profound. They harm one's image and reputation, but most importantly they have the power to harm one's psychological well-being. Disrespect at work causes doubt about one's abilities and competence and may lead to depression. Disrespect in personal relationships may severely undermine one's integrity and self-worth. Cultural disrespect leads to stereotyping, objectivization, and abuse. Disrespectful services may cause frustration, absurd experiences, and may also undermine trust in a profession, company, or an institute. The same is true for technology. When an interaction with a robot—no matter how complex and potentially beneficial it is—is justifiably felt to disregard the very essence of what makes us human, the consequences can be devastating. For instance, it is not uncommon that a machine's calculations outweigh the patient's reasoning when it comes to algorithm-aided diagnosis. This may happen even despite an obvious mistake in the interpretation of the suggestions made by the program. In such a case, an algorithm-based suggestion, in the eyes of the medical professional, rules out the possibility of the patient being right to doubt the diagnosis. Even if we leave aside the fact of misdiagnosis and the potential harm to the patient's health, the sheer invalidation of the patient's intelligence, especially coming from a figure of institutional authority, is enough to evoke indignation and create psychological damage (such as doubting one's ability to judge a situation based on facts). Another problem is that of *a blame vacuum*. A blame vacuum arises when a robot is, by design, assigned a certain function in an interpersonal relationship (such as care), which warrants the responsibility attribution but fails to fulfil the role of the appropriate object of blame (due to it being an object), while no human agent can be held accountable. When things go wrong, this creates an absurd situation for the victim that causes frustration and further traumatization.

The consequences of robot disrespect are not only about the loss of trust in the device and the lessening of the beneficial effects that it was intended for. Because some interaction with robots purports to establish an interpersonal relationship, disrespect entails psychological costs for the user, similar to the disrespect of a human. Technology, in itself, is indifferent to such considerations, so it is up to

us to ensure that—along with its benefits—the use of robotics will not also bring with it the bitterness of personal degradation.

11. Insufficiency of Considerations of Artificial Respectfulness for Ethics

The moral-neutrality of respect is the reason why establishing whether a robot is respectful cannot be sufficient when it comes to the ethical guidance of robotics. The respectfulness of a device must be seen as complimentary to the question of responsibility for the harm it may cause and the trustworthiness (reliability) of the device, to which it cannot be reduced.

The question about respectfulness should not be confused with that of responsibility[36] (incl. accountability[37] and liability). Responsibility discourse is motivated by the imperative of harm prevention. The attitude of respect does not, however, obey such a requirement. In fact, harm might be prescribed by a norm governing a certain value, as long as this is a non-ethical value. Killing a knight who has lost a battle may be carried out as an act of respect for the knight. In a robot governed by no more than a principle of respect, is it not necessary that its decision-making system will be determined not to harm the user (robot-assisted suicide, for example). Thus, if we want to make our robots safe, the work on respectfulness should go hand in hand with the work on responsibility for the harm that may result from robotic actions, and harm prevention.

Yet again, the considerations of respect are not the same as the considerations of trustworthiness. The trustworthiness of a robot cannot be understood in a stronger sense which presupposes an expectation that the trustee will be able to form a genuine motivational response to the normative demands of the trustor's vulnerability. No AI system is currently capable of such a response. Robot trustworthiness, then, must be understood in a weaker sense of reliability and predictability.[38] A trustworthy robot is a reliable tool, which does what it is supposed to do and is technically flawless. But a reliable tool is nothing more than a reliable tool, which can be used for good as well as for bad. Think, for example, about killer robots. It is at least questionable as to how respectful a killer robot is—one that remains true to its design and does not miss its target.

Acknowledgments

The major work on this article was funded by Kone Foundation in 2020 through the project *Towards Responsible Artificial Intelligence* (University of Helsinki, Finland). The article was completed at the University of Twente.

Notes

1. See, for example, Breazeal 2002.
2. Excluding animal companionship.
3. I am thankful to an anonymous reviewer for suggesting this distinction.
4. For the purposes of this article, I will use artificial respect and robot respect interchangeably.
5. In this respect, a reviewer made an interesting comment that, despite all the effort of programming respectful behavior into the robot, the user's experience of the robot's respect might be very different from her experience of a respectful attitude from a human. For example, how would earning the respect of a robot feel like? Would not it feel more like a game? Let me take this last point first. I argue that the only type of respect that we have ethical grounds to demand from robots is the respect of personhood. This type of respect does not need to be earned. One deserves such respect simply by virtue of being a human. As for other types of respect, engaging with algorithms encouraging respectful behaviour from the user (= "earning respect" of the robot) can indeed be seen a sort of a game. The benefit of my approach is that it provisions such differences in the experience of the user. These differences are explained by the fact that artificial respect is a different phenomenon from human respect. So, the difference in practices that constitute respecting/being respected in HRI are expected; they follow from the very nature of the asymmetric relationship between a human and an artificial agent.
6. By strong AI, in this context, I mean a hypothetical artificial system capable of thought, consciousness, and emotions in a genuinely human way. The concept was introduced by John Searle (1980) and is best understood in contrast to Alan Turing's (1950) imitation game.
7. For more on the difference between the two, see Dillon 2018.
8. What is known as Hume's law in ethics: no ought from is.
9. I am not hereby arguing against what appears to be an implicit assumption in HRI that robots should be designed in accordance to the user's preferences. However, I deny that crowd-sourcing is a reliable way to identify the universalizable (=ethical) constraints in robot design. (My thanks go to an anonymous reviewer for encouraging this clarification).
10. I am grateful to an anonymous reviewer for bringing up the topic of methodology and formulating possible objections.
11. For more on philosophical methodology in general, see Eder, Lawler, and van Riel 2020; Cappelen, Gendler, and Hawthorne 2016; Haug 2013.
12. Arguably, the Continental tradition is most methodologically creative, featuring methods like phenomenology (e.g., Husserl), dialectics (e.g., Socrates, Hegel), hermeneutics (Gadamer), structuralism (e.g., Lévi-Strauss, Chomsky), psychoanalysis

(e.g., Freud, Jung, Adorno), the transcendental method (Kant), as well as deconstruction and even metaphor (Derrida).

13. There is a substantial methodological diversity in the Analytic tradition as well. For example, in addition to the methods mentioned, Chase and Reynolds (2014) distinguish the formalization of arguments, i.e., modelling conditions with the rules of logic (symbolic, probabilistic, modal) and "coherence building methods," such as the reflective equilibrium, thought experiments, and identification of invalid argument patterns.

14. This methodology is a new topic in metaphilosophy. Even though I consider myself as doing conceptual engineering, I do not necessary accept all principles discussed in the abovementioned sources.

15. As I do not aim at a comprehensive account of respect, but only to explicate the ethically relevant features within the context of robotics, I cannot venture into a full-fledged literature analysis of the many sources on respect and self-respect. For an excellent overview and literature, I refer the reader to Dillon 2018 and Giorgini and Irrera 2017.

16. E.g., Bryson "Robots should be slaves," in Wilks 2010.

17. Admiration is a concept rarely discussed in contemporary moral philosophy outside virtue ethics, perhaps due to the fact that admiration of virtue ("moral examplars") has became a somewhat outdated concept. The reader might find useful Jan-Willem van der Rijt (2018), who develops a broadly Kantian approach to admiration and argues for its moral in-appropriateness; Zagzebski (2006) on emotional and cognitive elements of admiration, which she understands as a way to identify examplars; Alfred Archer (2019) on the connection between admiration and the motivation to emulate.

18. See more on "owe" in Keltner and Haidt 2010.

19. Cf. Sarah Buss (1999), who argues that Reverentia cannot be seen as a form of appraisal respect, but forms its own category.

20. See Schmidt 2000 for an account of the role of respect in conversation, aimed at truth and understanding.

21. For an interesting application of this principle in medical practice, see Barilan and Weintraub 2001. The authors argue that persuading a patient to accept a treatment is an act of respect for them as agents who can argue, persuade, and be persuaded, especially in matters of utmost importance such as health. For some alternative views that rational persuasion is morally wrong, see Tsai 2014 and Davis 2017.

22. On bullshit as disrespect for truth, see Frankfurt 2005; on data bullshitting by means of technology, see Bergstrom and West 2020.

23. Unless, of course, she carried that choice out in a manner that was directly targeted at the de-appreciation of any of these.

24. An objection may rise at this point—as an anonymous reviewer pointed out—saying that it is sufficient to program a robot in such a way that it will roughly

approximate human behavior that is normally considered respectful. The problem that the imitation approach faces is that of the justification of a certain type of respectful behavior that a robot is programmed to imitate. This problem is far more complex than it may appear because, as I show in this article, respect is not an ethical term, and if one chooses empirical methods to establish the principles of respectful behavior, one risks building a robot that imitates non-ethical behavior and promotes reprehensible values. So, in order to incorporate ethical constraints of respect in robot design, one has to first define these constraints and ensure that they effectively govern the actions of the robot. This article suggests that this goal can only be achieved by (a) explicating the ethical relevance of respect for personhood, and (b) showing that such respect will effectively govern the robotic actions *only if* its design and professional use is based on a commitment to the respect of personhood.

25. See Pettit 1989 on the kind of valuing that is involved in respect.

26. For a phenomenological account of recognition in the attitude of respect, see Drummond 2006.

27. In this article, I do not discuss the constraints on respect in social robotics targeted at aiding groups of people with special needs due to their unequal position, such as the elderly. For the latter, see Laitinen, Niemelä, and Pirhonen 2016.

28. The concept is traced back to Kant (1785/1996, 1788/1996, 1797/1996) and has been developed by, e.g., Alan Donagan (1977), Robert Downie and Elizabeth Telfer (1969), Thomas Hill (1993), William Frankena (1986), Carl Cranor (1975), Allen Wood (1999), and Iris Marion Young (1997).

29. I will use these two interchangeably. Alternative accounts of respect can be found in Brody 1982.

30. There are other accounts of integrity. Moral integrity, for example, is the compliance of one's actions with the coherent set of one's commitments (von Eschenbach 2012).

31. See, e.g., Musiał 2016, Scheutz 2012, and Whitby 2016.

32. This does not mean that we cannot choose to do so, given reasons other than the recognition of personhood.

33. For an alternative explanation in terms of categorical reasons for respect grounded in the concept of humans as valuers, see Raz 2001.

34. On the effects of disrespect on practical identity in general, see Honneth 1992; on responses to failing to respect, see Frankfurt 1999.

35. An interesting account of the role of robots in the phenomenon of recognition (non-social feedback in combination with simulated recognition), see Laitinen 2016.

36. On a factually informed account of responsible AI, see Dignum 2019.

37. See Nissenbaum 1996.

38. See more Ryan 2020, Simon 2020, Sullins 2020, and Section II Trust and Human-Robot Interaction in Lin, Abney, and Jenkins 2017.

References

Archer, Alfred. 2019. "Admiration and Motivation." *Emotion Review* 11(2): 140–50. https://doi.org/10.1177/1754073918787235

Baldwin, Thomas. 1998. "Analytical Philosophy." In *Routledge Encyclopedia of Philosophy Online*, ed. Edward Craig. https://www.rep.routledge.com/articles/thematic/analytical-philosophy/v-1.

Barilan, Y. Michael, and Moshe Weintraub. 2001. "Persuasion as Respect for Persons." *The Journal of Medicine and Philosophy* 26(1): 13–34. https://doi.org/10.1076/jmep.26.1.13.3033

Bergstrom, Carl T., and Jevin D. West. 2020. *Calling Bullshit: The Art of Skepticism in a Data-Driven World*. London: Allen Lane.

Breazeal, Cynthia L. 2002. *Designing Sociable Robots*. Cambridge, MA: MIT Press. https://doi.org/10.1007/0-306-47373-9_18

Brody, Baruch A. 1982. "Towards a Theory of Respect for Persons." *Tulane Studies in Philosophy* 31: 61–76. https://doi.org/10.5840/tulane1982313

Brun, Georg. 2016. "Explication as a Method of Conceptual Re-Engineering." *Erkenntnis* 81(6): 1211–41. https://doi.org/10.1007/s10670-015-9791-5

Brun, Georg. 2020. "Conceptual Re-Engineering." *Synthese* 197(3): 925–54. https://doi.org/10.1007/s11229-017-1596-4

Burgess, Alexis, Herman Cappelen, and David Plunkett, eds. 2020. *Conceptual Engineering and Conceptual Ethics*. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780198801856.001.0001

Buss, Sarah. 1999. "Respect for Persons." *Canadian Journal of Philosophy* 29(4): 517–50. https://doi.org/10.1080/00455091.1999.10715990

Cappelen, Herman, Tamar Gendler, and John Hawthorne, eds. 2016. *The Oxford Handbook of Philosophical Methodology*. Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199668779.001.0001

Chappell, Vere Claiborne, ed. 1964. *Ordinary Language: Essays in Philosophical Method*. Englewood Cliffs, NJ: Prentice-Hall.

Chase, James, and Jack Reynolds. 2014. *Analytic Versus Continental: Arguments on the Methods and Value of Philosophy*. Abingdon: Routledge.

Cranor, Carl F. 1975. "Toward a Theory of Respect for Persons." *American Philosophical Quarterly* 12(4): 309–20.

Cranor, Carl F. 1982. "Limitations on Respect-for-Persons Theories." *Tulane Studies in Philosophy* 31: 45–60. https://doi.org/10.5840/tulane1982314

Cranor, Carl F. 1983. "On Respecting Human Beings as Persons." *Journal of Value Inquiry* 17: 103–17. https://doi.org/10.1007/BF00158555

Danaher, John, and Neil McArthur. 2017. *Robot Sex*. Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/9780262036689.001.0001

Darwall, Stephen L. 1977. "Two Kinds of Respect." *Ethics* 88(1): 36–49. https://doi.org/10.1086/292054

Davis, Ryan W. 2017. "Rational Persuasion, Paternalism, and Respect." *Res Publica* 23(4): 513–22. https://doi.org/10.1007/s11158-016-9338-x

Dignum, Virginia. 2019. *Responsible Artificial Intelligence*. Cham: Springer. https://doi.org/10.1007/978-3-030-30371-6

Dillon, Robin S. 1992. "Respect and Care." *Canadian Journal of Philosophy* 22(1): 105–31. https://doi.org/10.1080/00455091.1992.10717273

Dillon, Robin S. 2018. "Respect." In *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta. https://plato.stanford.edu/archives/spr2018/entries/respect/.

Donagan, Alan. 1977. *The Theory of Morality*. Chicago: University of Chicago Press. https://doi.org/10.7208/chicago/9780226228419.001.0001

Downie, Robert, and Elizabeth Telfer. 1969. *Respect for Persons*. London: George Allen and Unwin.

Drummond, John J. 2006. "Respect as a Moral Emotion." *Husserl Studies* 22(1): 1–27. https://doi.org/10.1007/s10743-006-9001-z

Düwell, Marcus. 2017. "Human Dignity and the Ethics and Regulation of Technology." In *The Oxford Handbook of Law, Regulation and Technology*, ed. Roger Brownsword, Eloise Scotford, and Karen Yeung. Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199680832.013.8

Eder, Anna-Maria A., Insa Lawler, and Raphael van Riel. 2020. "Philosophical Methods under Scrutiny: Introduction to the Special Issue *Philosophical Methods*." *Synthese* 197: 915–23. https://doi.org/10.1007/s11229-018-02051-2

Feinberg, Joel. 1975. "Some Conjectures on the Concept of Respect." *Journal of Social Philosophy* 4: 1–3. https://doi.org/10.1111/j.1467-9833.1973.tb00163.x

Frankena, William K. 1986. "The Ethics of Respect for Persons." *Philosophical Topics* 14(2): 149–67. https://doi.org/10.5840/philtopics19861428

Frankfurt, Harry G. 1999. "Equality and Respect." In *Necessity, Volition, and Love*, ed. Harry G. Frankfurt, 146–54. Cambridge: Cambridge University Press.

Frankfurt, Harry G. 2005. *On Bullshit*. Princeton, NJ: Princeton University Press. https://doi.org/10.1017/CBO9780511624643.014

Giorgini, Giovanni, and Elena Irrera. 2017. *The Roots of Respect*. Berlin: De Gruyter. https://doi.org/10.1515/9783110526288

Glock, Hans-Johann, and Javier Kalhat. 2018. "Linguistic Turn." In *Routledge Encyclopedia of Philosophy Online*, ed. Edward Craig. https://www.rep.routledge.com/articles/thematic/linguistic-turn/v-1.

Haug, Matthew C., ed. 2013. *Philosophical Methodology*. Abingdon: Routledge.

Hill, Thomas E., Jr. 1993. "Donagan's Kant." *Ethics* 104(1): 22–52. https://doi.org/10.1086/293574

Hill, Thomas E., Jr. 1997. "Respect for Humanity." In *The Tanner Lectures on Human Values* XVIII, ed. Grether Peterson, 1–76. Salt-Lake City: University of Utah Press.

Hochberg, Herbert. 2003. *Introducing Analytic Philosophy*. Frankfurt: Ontos Verlag. https://doi.org/10.1515/9783110320763

Honneth, Axel. 1992. "Integrity and Disrespect." *Political Theory* 20(2): 187–201. https://doi.org/10.1177/0090591792020002001

Hudson, Stephen D. 1980. "The Nature of Respect." *Social Theory and Practice* 6(1): 69–90. https://doi.org/10.5840/soctheorpract19806112

Isaac, Manuel Gustavo. 2020. "How to Conceptually Engineer Conceptual Engineering?" *Inquiry*. https://doi.org/10.1080/0020174X.2020.1719881

Justus, James. 2012. "Carnap on Concept Determination." *European Journal for the Philosophy of Science* 2(2): 161–79. https://doi.org/10.1007/s13194-011-0027-5

Kant, Immanuel. (1785) 1996. "Groundwork of the Metaphysics of Morals." In *Immanuel Kant: Practical Philosophy*, ed. and trans. Mary Gregor, 37–108. New York: Cambridge University Press. https://doi.org/10.1017/CBO9780511813306.007

Kant, Immanuel. (1788) 1996. "Critique of Practical Reason." In *Immanuel Kant: Practical Philosophy*, ed. and trans. Mary Gregor, 133–272. New York: Cambridge University Press. https://doi.org/10.1017/CBO9780511813306.010

Kant, Immanuel. (1797) 1996. "The Metaphysics of Morals." In *Immanuel Kant: Practical Philosophy*, ed. and trans. Mary Gregor, 353–604. New York: Cambridge University Press. https://doi.org/10.1017/CBO9780511813306.013

Keltner, Dacher, and Jonathan Haidt. 2010. "Approaching Awe, a Moral, Spiritual, and Aesthetic Emotion." *Cognition and Motion* 17(2): 297–314. https://doi.org/10.1080/02699930302297

Laitinen, Arto. 2016. "Robots and Human Sociality." In *What Social Robots Can and Should Do: Proceedings of Robophilosophy 2016/TRANSOR 2016*, ed. Johanna Seibt, Marco Nørskov, and S. Schack Andersen, 313–22. Amsterdam: IOS.

Laitinen, Arto, Marketta Niemelä, and Jari Pirhonen. 2016. "Social Robotics, Elderly Care, and Human Dignity: A Recognition-Theoretical Approach." In *What Social Robots Can and Should Do: Proceedings of Robophilosophy 2016/TRANSOR 2016*, ed. Johanna Seibt, Marco Nørskov, and S. Schack Andersen, 155–63. Amsterdam: IOS.

Lee, Jason. 2017. *Sex Robots*. Cham: Springer International Publishing AG. https://doi.org/10.1007/978-3-319-49322-0

Levy, David. 2007. *Love and Sex with Robots*. New York: Harper.

Lin, Patrick, Keith Abney, George A. Bekey, Colin Allen, Anthony Beavers, Paul Bello, Jason Borenstein, Selmer Bringsjord, Marcello Guarini, and James J. Hughes, eds. 2012. *Robot Ethics*. Cambridge, MA: MIT Press. https://doi.org/10.1093/oso/9780190652951.001.0001

Lin, Patrick, Keith Abney, and Ryan Jenkins, eds. 2017. *Robot Ethics 2.0*. New York: Oxford University Press.

Musiał, Maciej. 2016. "Magical Thinking and Sympathy towards Robots." In *What Social Robots Can and Should Do: Proceedings of Robophilosophy 2016/ TRANSOR 2016*, ed. Johanna Seibt, Marco Nørskov, and S. Schack Andersen, 347–55. Amsterdam: IOS.

Nissenbaum, Helen. 1996. "Accountability in a Computerized Society." *Science and Engineering Ethics* 2(1): 25–42. https://doi.org/10.1007/BF02639315

Pettit, Philip. 1989. "Consequentialism and Respect for Persons." *Ethics* 100(1): 116–26. https://doi.org/10.1086/293149

Raz, Joseph. 2001. *Value, Respect, and Attachment*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511612732

Rescher, Nicholas. 2001. *Philosophical Reasoning: A Study in the Methodology of Philosophizing*. Malden, MA: Blackwell Publishers.

Richardson, Kathleen. 2015. *An Anthropology of Robots and AI*. New York: Routledge, Taylor & Francis. https://doi.org/10.4324/9781315736426

Richardson, Kathleen. 2019. "The Human Relationship in the Ethics of Robotics." *AI & Society* 34(1): 75–82. https://doi.org/10.1007/s00146-017-0699-2

Richardson, Kathleen, Mark Coeckelbergh, Kutoma Wakunuma, Erik Billing, Tom Ziemke, Pablo Gomez, Bram Vanderborght, and Tony Belpaeme. 2018. "Robot Enhanced Therapy for Children with Autism (DREAM)." *IEEE Technology and Society Magazine* 37(1): 30–39. https://doi.org/10.1109/MTS.2018.2795096

Ryan, Mark. 2020. "In AI We Trust: Ethics, Artificial Intelligence, and Reliability." *Science and Engineering Ethics* 26: 2749–67. https://doi.org/10.1007/s11948-020-00228-y

Scheutz, Matthias. 2012. "The Inherent Dangers of Unidirectional Emotional Bonds between Humans and Social Robots." In *Robot Ethics*, ed. Patrick Lin, Abney Keith, George A. Bekey, Colin Allen, Anthony Beavers, Paul Bello, Jason Borenstein, Selmer Bringsjord, Marcello Guarini, and James J. Hughes, 206–21. Cambridge, MA: MIT Press.

Schmidt, Lawrence. 2000. "Respecting Others." *Continental Philosophy Review* 33: 359–79. https://doi.org/10.1023/A:1010002504835

Searle, John. 1980. "Minds, Brains and Programs." *Behavioral and Brain Sciences* 3(3): 417–57. https://doi.org/10.1017/S0140525X00005756

Seibt, Johanna, Marco Nørskov, and S. Schack Andersen, eds. 2016. *What Social Robots Can and Should Do: Proceedings of Robophilosophy 2016/TRANSOR 2016*. Amsterdam: IOS.

Seymour, William, and Max Van Kleek. 2019. "The Internet of Kant: Respect as a Lens for IoT Design." In *Proceedings of CHI'19 Workshop Standing on the Shoulders*

of Giants: Exploring the Intersection of Philosophy and HCI, Glasgow, 2019. http://www.cs.ox.ac.uk/files/11110/workshop-philosophy.pdf.

Shepherd, Joshua, and James Justus. 2015. "X-Phi and Carnapian Explication." *Erkenntnis* 80(2): 381–402. https://doi.org/10.1007/s10670-014-9648-3

Simon, Judith, ed. 2020. *The Routledge Handbook of Trust and Philosophy*. New York: Routledge. https://doi.org/10.4324/9781315542294

Sorensen, Roy A. 1992. *Thought Experiments*. Oxford: Oxford University Press.

Sullins, John. 2020. "Trust in Robots." In *The Routledge Handbook of Trust and Philosophy*, ed. Judith Simon, 313–26. New York: Routledge. https://doi.org/10.4324/9781315542294-24

Tsai, George. 2014. "Rational Persuasion as Paternalism." *Philosophy & Public Affairs* 42(1): 78–112. https://doi.org/10.1111/papa.12026

Tugendhat, Ernst. 1976. *Traditional and Analytical Philosophy*. Cambridge: Cambridge University Press.

Turing, Alan. 1950. "Computing Machinery and Intelligence." *Mind* 59(236): 433–60. https://doi.org/10.1093/mind/LIX.236.433

Turkle, Sherry. 1985. *The Second Self*. Cambridge, MA: MIT Press.

Turkle, Sherry. 2004. "Whither Psychoanalysis in Computer Culture?" *Psychoanalytic Psychology* 21(1): 16–30.

Turkle, Sherry. 2011. *Alone Together*. New York: Basic Books.

Turkle, Sherry, Will Taggart, Cory D. Kidd, and Olivia Dasté. 2006. "Relational Artifacts with Children and Elders." *Connection Science* 18(4): 347–61. https://doi.org/10.1080/09540090600868912

van der Rijt, Jan-Willem. 2018. "The Vice of Admiration." *Philosophy* 93(1): 69–90. https://doi.org/10.1017/S0031819117000353

Van Kleek, Max, William Seymour, Reuben Binns, and Nigel Shadbolt. 2018. "Respectful Things: Adding Social Intelligence to 'Smart' Devices." In *Proceedings of Living in the Internet of Things: Cybersecurity of the IoT—2018*, 1–6, London. https://doi.org/10.1049/cp.2018.0006

Veletsianos, George, and Charles Miller. 2008. "Conversing with Pedagogical Agents." *British Journal of Educational Technology* 39(6): 969–86. https://doi.org/10.1111/j.1467-8535.2007.00797.x

von Eschenbach, Warren J. 2012. "Integrity, Commitment, and a Coherent Self." *Value Inquiry* 46(3): 369–78. https://doi.org/10.1007/s10790-012-9346-9

Wada, Kazuyoshi, and Takanori Shibata. 2007. "Living with Seal Robots: Its Sociopsychological and Physiological Influences on the Elderly at a Care House." *IEEE Transactions on Robotics* 23(5): 972–80. https://doi.org/10.1109/TRO.2007.906261

Whitby, Blay. 2016. "Do You Want a Robot Lover? The Ethics of Caring Technologies." In *Robot Ethics*, ed. Patrick Lin, Abney Keith, George A. Bekey, Colin

Allen, Anthony Beavers, Paul Bello, Jason Borenstein, Selmer Bringsjord, Marcello Guarini, and James J. Hughes, 233–49. Cambridge, MA: MIT Press.

Wilks, Yorick, ed. 2010. *Close Engagements with Artificial Companions*. Amsterdam: John Benjamins Publishing. https://doi.org/10.1075/nlp.8

Wood, Allen W. 1999. *Kant's Ethical Thought*. Cambridge: Cambridge University Press.

Wood, Allen W. 2010. "Respect and Recognition." In *The Routledge Companion to Ethics*, ed. John Skorupski, 562–72. London: Routledge.

Woodruff, Allison, Sally Augustin, and Brooke Foucault. 2007. "Sabbath Day Home Automation." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 527–36. New York. https://doi.org/10.1145/1240624.1240710

Young, Iris Marion. 1997. "Asymmetrical Reciprocity." *Constellations* 3(3): 340–63. https://doi.org/10.1111/j.1467-8675.1997.tb00064.x

Zagzebski, Linda. 2006. "The Admirable and the Desirable Life." In *Values and Virtues*, ed. Timothy Chappell, 53–66. Oxford: Oxford University Press.

Zagzebski, Linda. 2015. "Admiration and the Admirable." *Aristotelian Society Supplementary* 89(1): 205–21. https://doi.org/10.1111/j.1467-8349.2015.00250.x