Ethics Inf Technol (2012) 14:53–60
DOI 10.1007/s10676-011-9279-1

ORIGINAL PAPER

Can we trust robots?

Mark Coeckelbergh

Published online: 3 September 2011

Abstract Can we trust robots? Responding to the literature on trust and e-trust, this paper asks if the question of trust is applicable to robots, discusses different approaches to trust, and analyses some preconditions for trust. In the course of the paper a phenomenological-social approach to trust is articulated, which provides a way of thinking about trust that puts less emphasis on individual choice and control than the contractarian-individualist approach. In addition, the argument is made that while robots are neither human nor mere tools, we have sufficient functional, agency-based, appearance-based, social-relational, and existential criteria left to evaluate trust in robots. It is also argued that such evaluations must be sensitive to cultural differences, which impact on how we interpret the criteria and how we think of trust in robots. Finally, it is suggested that when it comes to shaping conditions under which humans can trust robots, fine-tuning human expectations and robotic appearances is advisable.

Keywords Trust · Ethics · Robots · Social relations · Phenomenology · Culture

Introduction

To frame the ethical question concerning robotics in terms of trust may easily suggest science-fiction scenarios like the story in the film 'I, robot' in which robots become artificially intelligent to such an extent that humans wonder if they can be trusted—which is usually interpreted as: Will they rise up against us? But there is a broader and certainly more urgent issue about trust in intelligent autonomous technologies that are already available today or will be available soon. Robotics for entertainment, sex, health care, and military applications are fast developing fields of research, autonomous cars are being developed, remote controlled robots are used in medicine and in the military, and we already heavily rely on semi-robots such as auto-pilot-airplanes. And of course we (often heavily) rely on computers and other information technology in our daily lives. The more we rely on these technologies, the more urgent becomes the issue of trust. As Taddeo writes in her introduction to a special issue devoted to trust in technology:

> As the outsourcing to (informational) artefacts becomes more pervasive, the *trust* and the dependence of the users on such artefacts also grow, bringing to the fore [issues] like the nature of trust, the necessary requirements for its occurrence, whether trust can be developed toward an artefact or can only concern human beings (…). (Taddeo 2010b, p. 283)

In other words, we already delegate tasks to machines and apparently we already trust them (I will return to this point below). This seems also the case with robots. But do we have good reasons to trust robots?[1] Do we need to develop what Wallach and Allen call 'moral machines' (Wallach and Allen 2008) in order to make robots trustworthy and avoid misuse of our trust? Does that mean we have to develop rational artificial agents? Is it appropriate at all to talk about trust here?

M. Coeckelbergh (✉)
Department of Philosophy, University of Twente, Postbus 217, 7500 AE Enschede, Netherlands
e-mail: m.coeckelbergh@utwente.nl

[1] We should also ask: Do *they* have reasons to trust *us*? I will discuss this question elsewhere.

In this paper, I postpone a direct analysis of the conditions under which we can trust robots. Rather, I discuss what trust means in relation to robots. First I provide a more general, brief analysis of trust: I discuss different approaches to analysing the conditions of trust (worthiness) in relation to artefacts and to people in general. This gives me some preconditions for trust. In the process I also comment on (other) discussions of trust in the literature on trust and e-trust. Then I explore if we can apply these preconditions to robots and how different approaches shed light on this question. I argue that in so far as robots appear human, trusting them requires that they fulfil demanding criteria concerning the *appearance* of language use, freedom, and social relations. This kind of virtual trust develops in science-fiction scenarios and can even become a question of 'mutual' trust. However, I also show that even if robots do not appear human and/or do not fulfil these criteria, as is the case today, we have sufficient functional, agency-based, social-relational, and existential criteria left to talk about, and evaluate, trust in robots. Finally, I argue that all criteria depend to some degree on the culture in which one lives and that therefore to evaluate trust in robots we have to attend to, and understand, differences in cultural attitude towards technology (and robots in particular) and, more generally, cultural differences in ways of seeing and ways of doing.

By proceeding in this way, I hope not only to contribute to the normative-ethical discussion about trust in robots, but also to discussions about how to approach information ethics and ethics of robotics.

Trusting artefacts

Although the notion of trust is usually applied to human–human relations, we sometimes say of an artefact that we (do not) trust it. What we mean then is that we expect the artefact to function, that is, to do what it is meant to do as an instrument to attain goals set by humans. Although we do not have full epistemic certainty that the instrument *will* actually function, we expect it do so. For example, one may trust a cleaning robot to do what it's supposed to do—cleaning. This kind of trust is what sometimes is called 'trust as reliance'.

Related to this type of direct trust in artefacts is indirect trust in the humans related to the technology: we trust that the designer has done a good job to avoid bad outcomes and—if we are not ourselves the users—we trust that the users will make *good* use of it, that is, that they will use the artefact for morally justifiable purposes. For example, military use of robots and artificially intelligent systems may be controversial because the (human) aims may be controversial (military action in general or particular uses, actions and targets).

But does this mean that trusting artefacts is *only* about reliance? Or does it mean that, as Pitt argues (Pitt 2010), the question about trust in technology *comes down to* trusting people (to do this or that)?[2]

Trusting people

In relation to people, the question of trust is different and more complicated than reliance. Trust is generally regarded as something that develops within (or *establishes*) a relation between humans (usually called a *trustor* and a *trustee*) that has ethical aspects (or *creates* an ethical dimension).

Trust

Before I discuss the relation between trust and ethics, let me first say something about trust and how it is usually approached, which influences the interpretation of the definition just given.

I propose that we distinguish between a contractarian-individualist and a phenomenological-social approach to trust and its relation to ethics. In the former approach, there are 'first' individuals who 'then' create a social relation (in particular social expectations) and hence trust between them. In the latter approach, the social or the community is prior to the individual, which means that when we talk about trust in the context of a given relation between humans, it is presupposed rather than created. Here trust cannot be captured in a formula and is something given, not entirely within anyone's control.

I take most standard accounts of trust to belong to the contractarian-individualist category. For example, in philosophy of information technology Taddeo's overview of the debate on trust and e-trust, and indeed her own work on e-trust, is formulated within a contractarian-individualist framework (Taddeo 2009, 2010a, b).

In her overview (Taddeo 2009) she discusses Luhmann (1979) and Gambetta (1988). Luhmann's analysis starts from the problem of co-ordination. Trust is seen as a response to that problem: it is needed to reduce complexity and uncertainty. Thus, the analysis starts from individuals and then it is questioned how the social can come into existence (if at all). Similarly, when Gambetta defines trust as 'a particular level of the subjective probability with which an agent assesses that another agent or group of agents will perform a particular action' (Gambetta 1988, p. 216), we find ourselves in a kind of game-theoretical

[2] In a later section of this paper, I will question these instrumentalist and one-sided views of trust in technology (and trust in robots in particular).

setting typical of contractarian views of the social. Whether or not to trust someone else is a matter of decision and choice. The agent calculates (at least in Gambetta's view) and cooperative relations are the result of these choices and calculations.

Taddeo's account of e-trust differs from these views but does not question the contractarian-individualist paradigm itself. E-trust may arise between agents under certain conditions (Taddeo 2009; see also Turili et al. 2010 and Nissenbaum 2001), but the starting point of the analysis remains individual agents, who take a risk when they trust someone else, form beliefs about the other agent(s) before they decide to trust, assesses the other's trustworthiness, etc. The starting point is the (Kantian regulative ideal of the) 'rational agent, able to choose the best option for itself, given a specific scenario and a goal to achieve' (Taddeo 2010a, p. 244). What matters is individual advantage and achievement of the individual goal, and then an 'objective assessment' is made in order to decide whether or not to enter into the relation.[3] Taddeo's view belongs to what Ess calls 'cognitive' or 'rational' accounts of trust: trust is the product of a rational process; 'we have *reasons* that justify our trusting others' (Ess 2010, pp. 289–290). While she acknowledges that less rational agents may follow different, more emotional criteria and that the criteria must be specified case by case (Taddeo 2010a, p. 254), her main analysis takes place at a higher level of abstraction and remains centred around the goals agents want to achieve, the choices they make, and the gains and costs their choices incur.

Weckert's view (2005) is different since it puts less emphasis on calculation and rational evaluation of beliefs and risks: he has criticized rational accounts for neglecting experiences of trust such as those of children in relation to their parents and those of friends. He has argued that we have the tendency to see other agents *as if* they were trustworthy and hence choose to act *as if* we trust and postpone evaluation. But regardless of (other) problems with his more 'affective' view (Ess 2010, p. 291), it stays within the lines of the contractarian-individualist approach: it is concerned with the agent's attitude and a similar contractarian evaluation is made—only afterwards.

The social-phenomenological view I attempt to articulate here, by contrast, defines trust not as something that needs to be 'produced' but that is already there, in the social. It is close to Myskja's phenomenological view (2008), which is based on a development of the pragmatic rather than the rational dimension of Kant's philosophy, and which locates trust in the centre of our embodied and social human condition. But is this still about ethics? Is this still about trust?

This leads me to the question: How do the different approaches conceptualize the relation between trust and ethics?

Trust and ethics

In the contractarian-individualist approach, responsibility is the flip side of trust. When we say that we trust someone, we also ascribe responsibility to that person: the person must be answerable to us for what he or she does. For example, if someone asks me to keep an eye on her bag and I walk away, I must answer to that person why I did not do what she expected. Thus, trust ascription creates a deontic field: if someone trusts me, I feel under an obligation not to misuse that trust. The person relies on me. The expectation becomes *normative* (as opposed to merely predictive, as in the case of artefacts—although indirect trust also involves normative expectations since then we deal with people). In response to the expectation, I may make a promise Whether or not I explicitly communicate this promise,[4] trust-giving and trust-receiving involves the employment of a kind of 'moral language' (I say I trust someone, I make promises, I communicate my expectations, etc.,) which creates the trust relation and its deontic implications.[5]

In the phenomenological-social approach, by contrast, responsibility is already built-in in the social, communitarian relation, which crucially has non-verbal and implicit aspects. Here morality is not something that is created but that is *already* embedded in the social. There are moral-social relations. There is a kind of basic confidence, in which reliance and reasoning are rooted and which they must presuppose.[6]

[3] Note that Taddeo also expands this rationalist and individualist view to artificial agents, see for instance Taddeo (2010a, p. 248); I will comment on 'artificial agents' below.

[4] Note, however, that even if there is no explicit communication about expectations and trust, this account presupposes that both parties are aware of the trust relation. If the trustee is not aware of being trusted, then it seems to me that the trustee is under no obligation and no deontic field is created. To use an example: if I were to 'trust' my neighbor to switch on his radio every morning in order for me to wake up (without him knowing that I 'trust' him in this way), this creates no obligation on the part of my neighbor. Perhaps this is more about expectations (alone) than about trust between people. In the same sense of 'trust' I may expect the sun to rise tomorrow morning. But no deontic field is created.

[5] My articulation of the contractarian-individualist view of the relation between trust and ethics is inspired by Searle's social ontology (Searle 1995 and subsequent developments of this theory).

[6] Note that these relations are also *felt*: feelings of friendship, love, etc. accompany the trust relation: among other things they make possible trust and—looking at it from a contractarian-individualist point of view—such feelings may be created and intensified by the construction of trust.

Preconditions for trust

Several conditions for trust (or, more precise, *trustworthiness*) have been discussed in the literature on trust and e-trust, such as direct interactions and shared values (see for example Taddeo 2009). But I wish to make a different kind of claim. The previous discussion of trust in human–human relations gives us at least three conditions we must *presuppose* about persons that might trust one another, regardless of (other) conditions under which they are justified to trust each other:

1. *Ability to use language*, in particular the moral language of giving trust, promise making, expressing expectations etc. In the contractarian-constructivist view of trust, this is absolutely necessary in order to establish trust. But accounts of trustworthiness 'assessment' based on calculation miss this moral-linguistic dimension. A social-phenomenological account, by contrast, has the conceptual resources to point to both linguistic and non-linguistic preconditions of trust. Talking about trust presupposes a subject-talker, who does not talk and act as an abstract 'rational agent' but as an embodied and social being.
2. *Freedom and uncertainty*: the giver of trust (the *trustor*) must be free, since the trustor cannot be forced to trust someone. A true gift cannot be forced. Moreover, the receiver of trust (*trustee*) must be free, since we must suppose that the receiver has the possibility to misuse the given trust—if there is absolute certainty about what will happen then there is no point in trusting someone. This means that there has to be freedom in the sense of proper delegation and no (direct) supervision. As Turilli et al. write when summarizing Taddeo's view: 'The trustor does not supervise the trustee's behaviour (…). Delegation and absence of supervision are then the defining characteristics of the occurrence of trust.' (Turili et al. 2010, p. 340). In the contractarian-constructivist view, individual freedom is crucial. The social-phenomenological view, however, can point out that the game of giving and receiving trust is already part of a social context in which trust is less under the control of individuals than assumed by the individualist view, and is more an emergent and/or embedded property. This also allows us to take a different perspective on the uncertainty related to trust: it is not so much that I am uncertain about whether or not I (as a rational agent) will reach my goal by delegating a particular task to someone else; rather, if there is a problem of trust I am uncertain about the social relation itself. When trust is an issue, the social relation, and therefore *I*, am at stake as a vulnerable and embodied social being.
3. *Social relations*. From a contractarian-individualist point of view, social relations are constructed or produced by individuals and any concept or institution that is related to the social, such as trust, also has this status: it is a construction or product. From a phenomenological-social point of view, trust-talk and talk about individual freedom presupposes social relations (and embodiment). In other words, there is trust because there are already social relations. Trust is something basic that must be presupposed; it is not created but emerges from social relations.[7] Therefore, we must presuppose of persons that might trust one another that there is already a social relation, which the persons experience as embodied and vulnerable beings that stand-in-relation.

I take the latter perspective to be in line with the phenomenological and virtue ethics views of trust summarized by Ess, which pay more attention to embodied, affective and social dimensions of trust as opposed to the rational and the individual dimensions (Ess 2010).

Can the two approaches be reconciled?

One may object now that I over-emphasize the differences between the two approaches. Contractarian-individualism, or at least one version of it, could respond to the phenomenological social challenge by claiming that both approaches are not far apart since they could agree that trust is the 'default'; only when there is a problem we switch to trust assessment. This objection rests on a particular version of the contractarian-individualist approach: the thesis is changed from 'One trusts only when there is a good reason to' to the different, modified thesis: 'One trusts unless there is a good reason not to trust'. If this modification is made, then it seems that the gap between the two approaches is not as wide as I suggest, since it seems that both approaches could agree that we trust by default and that in the default mode no thought is given to trusting. However, I believe the two approaches still differ in how they describe this default trust and in how they understand mistrust.

For contractarian-individualists, default trust is a matter of individual attitude or stance, whereas the phenomenological-social approach understands trust as something that

[7] One may object that there can be trust without social relations: I may trust myself. (e.g. to do something) However, self-trust is at least a special case of trust, if it is about trust at all. Perhaps self-trust depends on the kind of relation I have with myself, which is 'social' in the sense that it supposes that I talk with myself as if there are two persons having a conversation. By itself, talking with yourself is a form of thinking and is very common. But saying that you trust yourself remains a philosophically (and probably also psychologically) problematic use of the word 'trust'.

arises or emerges from the social relations in which we *find* ourselves. The very term 'default' still belongs to the contractarian-individualist vocabulary since it presupposes that there is always a choice situation. As in electronics and computing, 'default' refers to a pre-selected option that is followed except when changed. Used as a metaphor by contractarian-individualists, it means that the social 'system' may well pre-select the option 'trust' but that we individuals can change this if there is a good reason to do so (the thesis is that 'one trusts unless there is a good reason not to trust'), which implies that we can always assess and re-assess, and then adapt our relation to others. But this presupposes that we have always a choice with regard to social relations and their form(ation). And if we are aware of this, then the seed of mistrust has already been planted. The phenomenological-social approach, by contrast, attends us to the possibility that we do not always and perhaps not usually have full control over giving trust or not giving trust (and indeed over our social relations and their form). Sometimes we trust in spite of good reasons not to trust, or sometimes we mistrust in spite of good reasons to trust. The phenomenological-social approach concedes that sometimes we live in the mode of 'trust assessment', but it stresses that the contractarian-individualist thesis 'One trusts unless there is good reason not to trust' describes only *one* way of shaping our social relations, *one* possibility of how we can look at human relations. Of course the new formula is already more 'trustful' than the initial 'One trusts only when there is a good reason to', but still gives the last word to human reasoning. The phenomenological-social approach accounts for the experience that sometimes we are *drawn into* trust or mistrust, that sometimes—and perhaps more often than we like—we cannot help (mis) trusting.

In any case, we now have an overview of different approaches to human trust and we have some preconditions for trust in humans. But what about trust in robots?

Trusting robots

In so far as robots can be considered as 'mere' artefacts, our trust in them must be based on functional criteria. If they are means to an end, then whether or not they attain the end (success or no success or a certain degree of success) must be the criterion for trust. But is this all that can be said about trust between humans and robots? It seems that robots are 'more' than mere tools.

I believe that there are at least two ways to go beyond the instrumentalist view of robots, the approach of which roughly fall within the categories I called 'contractarian-individualist' and 'phenomenological-social'.

First, within the (individualist) analytical tradition, one may discuss trust between humans and robots in terms of trust between human agents and 'artificial agents'. Taddeo has even discussed trust *between* artificial agents. In the latter case, the human-style preconditions such as freedom and language do not seem relevant; rather, non-anthropocentric criteria such as (operational) autonomy and interactivity are proposed (Taddeo mentions for example the criteria for artificial agency proposed by Floridi and Sanders 2004). Taddeo has argued that trust-based interactions are possible 'even when social and moral norms are not present' (Taddeo 2009, p. 19). Thus, here the 'problem' is solved by conceptualizing both humans and robots as *agents*. Considering both the human and the robot as an agent, that is, an 'it' (Taddeo 2010a, p. 244), this approach allows one to employ the contractarian-individualist apparatus across the human-robot distinction.

Second, contemporary philosophy of technology in the phenomenological tradition has shown in a different way that the instrumentalist view of technology is inadequate. For example, influenced by (mainly) Heideggerian phenomenology, the insight has emerged that technological artefacts 'do' more than is intended by humans: they co-shape how we understand and act in the world (Ihde 1990; Verbeek 2005). Hence, robots do not just do what they are made for and their meaning is what Ihde calls 'multistable': we might see them as machines but also as *more than machines*. In particular, we might treat them as if they were animals or as if they are (human) persons—a type of entity we can relate to as social beings. In the latter case, they become 'quasi-others'.

In recent work I have argued that from a phenomenological point of view, robots may *appear* as more than machines and that this has consequences for ethics of robotics: a relational, social, and phenomenological approach helps us to better understand human-robot relations and the ethical problems they raise (Coeckelbergh 2009a, 2009b, 2010a, 2010b, 2010c, 2010d). Thus, here the question is not whether or not robots *are* agents (individual-ontological approach) but how they appear and how that appearance is shaped by, and shapes, the social (social-relational approach). Appearance-making, sometimes named 'deception', then is part of 'the social game' (see also Myskja 2008, p. 217) and it does not undermine trust but supports it.

For trust in robots, the latter approach implies that although robots *are* not human and do not meet the two constructivist-individualist preconditions for trust in human–human relations (ability to use language and freedom), they may nevertheless contribute to the establishment of 'virtual trust' or 'quasi-trust' in so far that they *appear* as quasi-others or social others, as language users, and as free agents. We trust robots if they appear trustworthy and they appear trustworthy if they are good

players in the social game.[8] (This might even create the feeling of 'mutual' trust on the part of the human.)

One may object that such robots are still science-fiction. However, one should not underestimate the potential implications of progress in the field of social robotics and the capacity for humans to anthropomorphize. No perfect 'copy' of the human is necessary to trigger quasi-social responses.

But let us suppose for the sake of argument that this reply is unsatisfactory and that we cannot possibly build robots that appear to meet the mentioned preconditions for trust in human–human relations. Can we still trust (current) robots? How can we evaluate existing robots in terms of trust?

First, we can still use the functionalist, performance criterion: can the robot do what it is supposed—that is, expected—to do? Is it a means to the end set by us?

Second, we may consider the robot as an 'artificial agent' and apply Taddeo's conceptual framework to discuss trust between humans and robots-as-artificial agents.

Third, whether or not the robot and the human-robot relation meet the constructivist-individualist criteria or the criteria for artificial (moral) agency, they may fulfil the following phenomenological-social requirement: if a human-robot relation grows as a *social* relation, then trust is already there as a 'default' in the social relation—albeit in an implicit, affective way—regardless of how people construct the relation (e.g. as human–*machine* interaction) and how they talk about the robot and about themselves (e.g. as individuals). In contrast to my description of 'virtual trust', there is no requirement here that the robot appears as a quasi-other; the emphasis is not so much on perception but on the relational bond, which is more 'felt' and experienced than seen or ac-*knowledge*d. Most of the time, no deliberation is needed about who or what to trust. We live with technology and with others; we are engaged in social-technological activity.

[8] By 'social game' I mean to emphasize that social relations should not be considered as a 'state' but as an ongoing process, which has its own rules and dynamics (and contractarian game theory is only one possible way of describing this process). I also hint at Myskja's use of the term in his discussion of Kant's argument that pretending to be better than we are, actually makes us better: pretending is acceptable as part of the 'social game' and therefore even required. Thus, playing the 'social game' (for robots and for humans) means producing the right kind of appearance and pretending that you are better than you are. Note that next to this 'moral' deception the robot also has to perform an 'ontological' deception in the sense that it has to pretend to be an entity that it is not. Note also that not only (humanoid or social) robots can appear in a personal and social way. For example, some people talk to their navigation device in their car. People also talk to computers and other things. Whether or not something appears as quasi-other is a matter of degree, and robots that resemble humans seem to promote a high degree of what Ihde calls an 'alterity' experience: the technology is not perceived as a tool, as something in the background, or as something that has become part of ourselves (e.g. glasses), but as an other.

Third, robots can and should also be evaluated not only in terms of what they do in the world in relation to the goal set (functionalist, performance criterion) or in terms of how they shape the human-robot relation (social criterion) but also in terms of how they help us to understand and shape ourselves. For example, I have argued in other conference contributions that robots are hermeneutical tools that help us to understand ourselves. And recently Kiran and Verbeek have argued that technology puts at stake what it is to be a human being: humans and technologies have an 'intimate' and 'internal' relation and as we trust ourselves to technology we shape our existence. Rather than reliance, trust then takes on the character of confidence (Kiran and Verbeek 2010).[9]

Thus, from a social-phenomenological (and from an existential-phenomenological) perspective I conclude that we *already* trust ourselves to technologies and depend on them. The individualist-constructivist approach (and to some extent the existentialist-individualist approach) presupposes that there is first the human individual, who asks himself the question if he or she will use the technology and trust it. But our lives are already interwoven with technologies. They are not just tools we use to attain our goals (including goals concerning self-care or self-constitution); they are part of the social and existential fabric, from which we emerge as individuals and selves in the first place. In this sense, evaluating whether or not we can trust robots means to evaluate the social and ourselves as social beings.

Cultural differences

This analysis of preconditions for trust and their application to human-robot relations need not be interpreted in an entirely universalist way. The application of all criteria for trusting robots (and, more generally, technology) proposed here depends to a significant degree on the culture in which one lives. Consider again the criteria, from which I derive the following hypotheses concerning the role of cultural difference in trusting robots:

1. Functionality/performance: in so far as cultures differ with respect to their ends (values) and with respect to the ways they see fit to reach those ends (using technology or not, how technology is used, for which end), whether or not a person has reasons to trust a robot, depends on the culture he or she lives in. For example, in some parts of the world using robots in

[9] Note, however, that by emphasizing the intimate relation between technology and individual subjectivity as self-constitution and self-care, Kiran and Verbeek do not pay much attention to the social aspect of trust. Moreover, they suggest that trusting ourselves to technology—developing a 'free' relation to technology—requires a *deliberate* effort on our part, whereas I have paid attention to the non-deliberative aspect.

elderly care may be seen as less problematic than in other parts of the world.
2. The term 'agent' (and hence 'artificial agent' and the type of analysis that can be conducted in the wake of this term) may be less intelligible in societies that have a less individualistic culture than ours. Hence references to criteria for agency, the emphasis on rational agency, etc. may be problematic in some cultures.
3. Language use: whether or not a robot appears to use moral language depends on whether or not the robot can imitate the moral language of a particular culture. Cultures differ with respect to the way they express expectations, promises, etc. related to trust.
4. Freedom: if a culture places more value on individual freedom, it will be more difficult for a robot to produce the appearance of freedom and hence to meet this condition.
5. Social relations: cultures differ in the extent and the way the embed robots into the social fabric. If robots are already perceived and *lived* as part of society and culture, one can expect a higher degree of trust in robots.
6. Human-technological existence: if a culture puts less emphasis on the subject-object distinction or even rejects such a distinction, we can expect that trust in robots will be higher.

These hypotheses could guide empirical research but can also be used for the purpose of further philosophical reflection on, and understanding of, what trusting robots means. Furthermore, apart from pointing us to differences in interpretation and application of trust criteria with regard to human-robot relations, attention to cultural differences has also the philosophical merit of re-opening the discussion about overall approaches to trust. To what extent can we 'choose' between the contractarian-individualist and the phenomenological-social approach, given (and assuming) that the first is more prominent in the West? If we take a phenomenological-social approach, the meta-discussion about approaches is not seen as being isolated from cultural context. The way we think about trust, even at this more abstract level, also depends on the culture we live in. The cultural context of this inquiry is both enabling and limiting. The contractarian-individualist approach gives us powerful conceptual tools; perhaps it is difficult to fully explore and articulate a more social approach to trust in a social-cultural environment that is impregnated with a different way of seeing and doing.

Conclusion

Based on a brief, preliminary analysis of what it means to trust artefacts and to trust people, I analysed two opposing approaches to trust(worthiness) and suggested some preconditions for trusting robots. Since some of the criteria were based on human–human relations, they first seemed not applicable to human-robot relations, but I have discussed two solutions for this problem, based on (1) a contractarian-individualist approach focused on agency and (2) a phenomenological and social approach. I have mainly supported the latter approach, which often receives little consideration in information ethics and ethics of robotics.

What remains is the insight that if we wish to fully understand what it takes for us to trust robots, we should not take for granted the instrumental view of human-technology relations and the individualist-constructivist view of trust and social relations. Thus, trusting technology need not only concern the human attempt to be in control and in power, as Luhmann suggested (1979) and as contractarian-individualist echo when they define trust in terms of a decision. Adaptation to environments (e.g., techno-social environments) does not necessarily require the exercise of agency. Often we cannot help trusting technology and trusting others, and luckily we often do so without having a reason and without calculation (not even afterwards). In so far as robots are already part of the social and part of us, we trust them as we are already related to them. And if they are new, then we trust them as we are *beginning* to relate to them. As we are learning and as we are developing skills for dealing with these new entities, trust grows. In this sense, trusting robots is not science-fiction but is already happening, and rational calculation is only one interpretation of, and way of how one's relation to the technology takes shape. (I do not deny, of course, that rational assessment of one's relations is possible; however, it is important to see that it is only one way of seeing and doing, which presupposes a more basic social-moral ground.)

Moreover, I should be careful with using the 'we' here. I have highlighted how the application of any of these criteria is dependent on culture. This does not imply that we should refrain from attempting to offer general ethical guidelines when it comes to trusting robots and ethics of robotics—the discussion in this essay could be viewed as an exploration of what kind of framework could possibly justify such guidelines—but rather that when we have such general principles or criteria, they should not be understood as standing outside the cultural-hermeneutical process; they will always require interpretation as we move on. The same is true for the approaches to trust presented here.

To conclude, in this discussion I have mainly discussed the question if it is appropriate at all to talk about trusting robots, which has allowed me to distinguish between different approaches to trust and which has given me some preconditions for trust. I have also indicated the relevance of cultural differences for answering the question regarding

trust in robots. However, I have not provided a more straightforward discussion of the question under what conditions we can trust robots (the answer to which must be informed by one of the approaches). Such an inquiry might for example involve the question if 'affective' robots are deceptive—a question I discussed elsewhere. But whether or not affective robots are deceptive, one condition for trust seems to concern fine-tuning human expectations about robots, perhaps by fine-tuning robotic appearances. Picard, who initiated the term and field of 'affective computing' (Picard 1995, 1997, 2003), writes in her book the following, which seems applicable to 'affective' robots:

> A danger with personified characters (…) is that people may consequently expect human-like intelligence, understanding, and actions from them. In some cases, a machine may need to explain what it can and cannot do. In any case, it will be important to help people accurately set expectations of the computer's abilities. (Picard 1997, p. 114)

As this quote suggests, perhaps there is also a sense in which we may put *too much* trust in some kinds of robots—especially if there *is* a kind of basic trust and confidence that permeates social relations and if this kind of trust is carried over to human-robot relations. More should be said about robotic emotions, appearance, and (conditions for) trust. But here I pause my reflections on the question regarding (the evaluation of) trust in robots and how we can approach this question.

References

Coeckelbergh, M. (2009a). Personal robots, appearance, and human good: A methodological reflection on roboethics. *International Journal of Social Robotics, 1*(3), 217–221

Coeckelbergh, M. (2009b). Virtual moral agency, virtual moral responsibility. *AI & Society 24*(2), 181–189

Coeckelbergh, M. (2010a). Humans, animals, and robots: A phenomenological approach to human-robot relations. *International Journal of Social Robotics, 3*(2), 197–204

Coeckelbergh, M. (2010b). You, robot: On the linguistic construction of artificial others. *AI & Society, 26*(1), 61–69

Coeckelbergh, M. (2010c). Robot rights? Towards a social-relational justification of moral consideration. *Ethics and Information Technology, 12*(3), 209–221

Coeckelbergh, M. (2010d). Moral appearances: Emotions, robots, and human morality. *Ethics and Information Technology*, *12*(3), 235–241

Ess, C. M. (2010). Trust and new communication technologies: Vicious circles, virtuous circles, possible futures. *Knowledge, Technology and Policy, 23*(3–4), 287–305.

Floridi, L., & Sanders, J. (2004). On the morality of artificial agents. *Minds and Machines, 14*(3), 349–379.

Gambetta, D. (1988). Can we trust trust?. In D. Gambetta (Ed.), *Trust: Making and breaking cooperative relations* (pp. 213–238). Oxford: Basil Blackwell.

Ihde, D. (1990). *Technology and the lifeworld*. Bloomington/Minneapolis: Indiana University Press.

Kiran, A., & Verbeek, P. P. (2010). Trusting our selves to technology. *Knowledge, Technology and Policy, 23*(3–4), 409–427.

Luhmann, N. (1979). *Trust and power*. Chichester: John Wiley.

Myskja, B. K. (2008). The categorical imperative and the ethics of trust. *Ethics and Information Technology, 10*, 213–220.

Nissenbaum, H. (2001). Securing trust online. *Boston University Law Review, 81*(3), 635–664.

Picard, R.W. (1995). *Affective computing*, MIT laboratory perceptual computing section technical report No. 321. Cambridge, MA: MIT. Retrieved from http://affect.media.mit.edu/pdfs/95.picard.pdf.

Picard, R. W. (1997). *Affective computing*. Cambrige, MA: MIT Press.

Picard, R. W. (2003). Affective computing: Challenges. *International Journal of Human-Computer Studies, 59*, 55–64.

Pitt, J. C. (2010). 'It's not about technology. *Knowledge, Technology and Policy, 23*(3–4), 445–454.

Searle, J. R. (1995). *The construction of social reality*. New York: Free Press.

Taddeo, M. (2009). Defining trust and e-trust: Old theories and new problems. *International Journal of Technology and Human Interaction, 5*(2), 23–35.

Taddeo, M. (2010a). Modelling trust in artificial agents, a first step toward the analysis of e-trust. *Minds and Machines, 20*(2), 243–257.

Taddeo, M. (2010b). Trust in technology: A distinctive and a problematic relation. *Knowledge, Technology and Policy, 23*(3–4), 283–286.

Turili, M., Vaccaro, A., & Taddeo, M. (2010). The case of online trust. *Knowledge, Technology and Policy, 23*(3–4), 333–354.

Verbeek, P.-P. (2005). *What things do*. Peen: Penn State University Press.

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

Weckert, J. (2005). Trust in cyberspace. In R. J. Cavelier (Ed.), *The impact of the internet on our moral lives* (pp. 95–120). Albany: University of New York Press.