TRAMES, 2019, **23(73/68)**, 4, 499–522

SHOULD WE TRUST ARTIFICIAL INTELLIGENCE?

Margit Sutrop

University of Tartu

Abstract. Trust is believed to be a foundational cornerstone for artificial intelligence (AI). In April 2019 the European Commission High Level Expert Group on AI adopted the Ethics Guidelines for Trustworthy AI, stressing that human beings will only be able to confidently and fully reap the benefits of AI if they can trust the technology. Trustworthy AI is defined as ethical, lawful and robust AI. Three things strike me about the EC Guidelines. Firstly, though building trust in AI seems to be a shared aim, it is not explicated what trust is, and how it can be built and maintained. Secondly, the Guidelines ignore the widespread distinction made in philosophical literature between trust and reliance. Thirdly, it is not clear how the values have been selected with which AI has to align and what would happen if they came into conflict. In this paper, I shall provide a conceptual analysis of trust in contrast to reliance and ask when it is warranted to talk about trust in AI and trustworthy AI. I shall show how trust and risk are related and what benefits and risks are associated with narrow and general AI. Also, I shall point out that metaphorical talk about ethically aligned AI ignores the real disagreements we have about ethical values.

Keywords: artificial intelligence, general intelligence, superintelligence, AI, AGI, SAI, trust, trustworthiness, reliance, risk, EC guidelines for trustworthy AI

DOI: https//doi.org/10.3176/tr.2019.4.07

1. Introduction

Trust is believed to be a cornerstone for artificial intelligence (AI). In April 2019 the European Commission's High-Level Expert Group on Artificial Intelligence (AI HLEG) published *Ethics Guidelines for Trustworthy AI*, stressing that human beings will be able to confidently and fully reap the benefits of AI only if they have trust in it. The Guidelines call for the development of 'Trustworthy AI', featuring a human-

centric approach and emphasizing two components: (1) respect for fundamental rights, applicable regulations and core principles and values, ensuring an 'ethical purpose' and (2) technical robustness and reliability, to avoid unintentional harm caused by lack of technological mastery (EU Commission 2019b: I). Trustworthy AI is, according to the Guidelines, ethical, lawful, and robust AI.

Formulation of the Guidelines is a pioneering initiative, since for the first time they set forth a normative framework for developing, deploying and using AI in the EU, while also aspiring to offer guidance to discussions taking place outside the EU. Arguing that the EU should follow a human-centric approach, where humans have primacy in civil, political, economic and social fields, the Guidelines employ an individual rights-based approach. They set the foundations of Trustworthy AI in fundamental rights and four principles (respect for human autonomy, prevention of harm, fairness, and explicability), using these to formulate specific requirements in the AI context. The document also describes technical and non-technical methods to achieve Trustworthy AI.

However, there are three things that strike me about the Guidelines. Firstly, although there is much talk about trust, surprisingly little is said about what constitutes trust and what it depends upon. But how we understand the **nature of trust** will make a difference in what we say about the conditions in which trust is justified and how it should be built and maintained.

Trust seems to be understood in terms of trustworthiness, as there is an implicit assumption that being demonstrably worthy of trust will create trust. The Guidelines identify Trustworthy AI as the European Commission's foundational ambition, since trustworthiness is "a prerequisite for people and societies to develop, deploy and use AI systems" (2019b: 4). Indeed, ideally, those whom we trust will be trustworthy, and those who are trustworthy will be trusted. But, in reality, having the property of trustworthiness is not a guarantee for being trusted. We know that sometimes it happens that those whom we trust are not trustworthy, and those who are trustworthy are not actually trusted. Thus, if it is important that people trust AI systems, it is not enough to establish and articulate the purpose of achieving trustworthy AI. It is imperative that we also think about how to build trust in AI.

Secondly, it is surprising that the document talks about trust in AI and not about reliance. Since Annette Baier's study (1986) it has been a commonplace in philosophical literature to distinguish between trust and reliance. Trust is thought to be an interpersonal relationship between two peers, while our attitudes towards inanimate objects, such as cars, computers, or alarm clocks evoke the mental attitude of reliance. An important condition for trust is the potential of betrayal, whereas the condition corresponding to trustworthiness is the power to betray. The standard view in philosophical literature is that people considered to be trustworthy have the power to betray us, whereas people and inanimate objects considered to be merely reliable can only disappoint us (Baier 1986, Holton 1994, Wright 2010, McLeod 2015).

The way in which the European Commission's *Ethics Guidelines for Trustworthy AI* talk about trust in AI and trustworthy AI raises the question whether AI is being treated on par with a human, or whether the document ignores this conceptual

distinction as it is made in the philosophical literature. Also, one should ask whether there are any consequences to talking about the trustworthiness of AI, instead of its reliability or accountability.

Thirdly, it is noteworthy that the Guidelines employ an individual rights-based approach, ignoring the fact that liberal individualism, with its conceptual base of autonomy, dignity, and privacy has, after a long period of dominance in research ethics increasingly come under attack from communitarian ideologies which promote a more salient role for concepts of solidarity, community, and public interest (Sutrop 2011a; 2011b). The Guidelines list four principles: respect for human autonomy, prevention of harm, fairness, and explicability, the source of which seems to lie in existing legal requirements, with the admission that the "adherence to ethical principles goes beyond formal compliance with existing laws" (EU Commission 2019b: 12). There is also some inconsistency, as on the one hand the Guidelines stress that AI systems should respect the plurality of values and choices of individuals (EU Commission 2019b: 11); on the other hand, they claim that certain fundamental rights and principles, such as human dignity, are absolute and cannot be subject to a balancing exercise (EU Commission 2019b: 13). Also, the Guidelines are not helpful in advising what should be done if the principles conflict.

In the following I will first provide a brief overview of the definition of AI and how the benefits and risks of AI are being envisioned in scholarly literature. I will then conduct a philosophical analysis of trust and distinguish between its different forms. On the basis of this conceptual analysis of trust, I will hopefully be able to answer the question, whether we should avoid talking about trust in AI and limit the concept of trust to peer relationships; thus also limiting the concept of trustworthiness for people and institutions who design, deploy, and govern AI. The other alternative, if it is warranted, is to continue to speak of trusting AI and its trustworthiness. In the last part of the article I shall point out that metaphorical talk about trustworthy AI and ethically-aligned AI ignores the real disagreements that we have about ethical values.

2. The definition of AI

Artificial intelligence has been described in various ways. AI can refer to certain systems designed by humans, or to a scientific discipline that includes several approaches and techniques, such as machine learning, machine reasoning, and robotics. In this paper I will draw upon the definition of AI developed by AI HLEG in the document "A Definition of AI: Main Capabilities and Disciplines," made public in April 2019.

AI is defined by AI HLEG as follows: "Artificial intelligence (AI) systems are software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information derived from this data

and deciding the best action(s) to take to achieve the given goal. AI systems can either use symbolic rules or learn a numeric model, and they can also adapt their behaviour by analysing how the environment is affected by their previous actions" (EU Commission 2019a).

Currently, AI systems are narrowly dedicated to specific tasks such as face recognition, spam filters or self-driving cars, and they are not capable of setting their own goals or choosing the best courses of action across domains. An AI which, with human-level ability or beyond is able to find a solution when presented with an unfamiliar task, is called Artificial General Intelligence (AGI). The achievements of AGI are hypothesised to lead to an intelligence explosion, facilitated by recursive self-improvement of the AGI, eventually attaining the level of Superintelligence (SAI) (Kurzweil 2005, Tegmark 2017). SAI is "any intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest" (Bostrom 2014: 26).

Many researchers now take seriously the possibility that within the current century intelligence equal to our own will be created in computers. Freed from biological constraints such as limited memory and slow biochemical processing speeds, machines may eventually become more intelligent than we are – with profound implications for us all. As the famous physicist Stephen Hawking says in his posthumously published book, *Brief Answers to the Big Questions*: "… the advent of super-intelligent AI would be either the best or the worst thing ever to happen to humanity. /…/ Our future is a race between the growing power of our technology and the wisdom with which we use it. Let us make sure that the wisdom wins" (Hawking 2018).

The impact of AI will permeate all spheres of our lives, from commercial and social interactions to relationships with the state, including dramatic structural transformations in the public sphere (cf. Habermas [1962]1991, Beck 1992, 2016). Let us look at some specific examples. (1) If AI gets increasingly better at performing tasks towards a given goal (e.g. good governance, better healthcare), then we ought to let AI work on that goal. This requires figuring out the proper distribution of control between humans and AI which is, in turn, premised on humans trusting AI. (2) The power and effect of bots to influence social media discussion has been well documented and researched (e.g. Suarez-Serrato et al. 2016, Hwang and Rosen 2017, Bessi and Ferrara 2016). The more advanced AI becomes, the easier it will be to influence public discussion. Thus, AI has the potential to erode trust in publicly available information. (3) In the context of healthcare, AI is already playing a significant role in triaging and diagnosing patients (Varun et al. 2018). Given that we already have several diagnostic devices that outperform human specialists (Bien et al. 2018), it is likely that in some fields AI will technically produce better results. However, the possible consequences of these changes to the doctor-patient relationship and healthcare systems more generally are far from obvious and require investigation. Arguably AI can "depersonalize medical interactions" (Terrasse et al. 2019: 26), although we have the tendency to anthropomorphize entities with which we interact (Banja 2019: 34). Initial studies with conversational AI demonstrated

that in some contexts (e.g. mental health) participants trusted a non-human chatbot more than a human-controlled agent – they were less fearful and more forthright with their own expressions (Lucas et al. 2014). Medical knowledge has cultural and regional aspects (endemic diseases, culturally stigmatized conditions) that AI might miss. Also, the data on which medical AI is trained, can be biased, leading to unfairness in its applications.

Issues of risk and trust also arise in connection with potential developments in the human-machine relationship. In the future humans could merge with machines; an extreme example of this is the mind-upload possibility (Tegmark 2017). Chalmers (2010) has suggested that the most beneficial option for humans in a post-singularity world is to become integrated with SAI, and recommends that one should upload one's mind onto a computer. From there, the uploaded mind could be downloaded into an artificial body. Apart from several practical and ethical issues (Strout 2014), mind-upload gives rise to foundational philosophical issues concerning the nature of uploaded minds, their identity, and questions about their consciousness. Could one trust technology enough to undergo a mind-upload process oneself? Although the uploaded minds are human minds with human ethical principles, these may change substantially when uploaded, thus calling their trustability into question.

Although there is no consensus among experts as to when/if AGI will be achieved (Müller and Bostrom 2016, Grace et al. 2018), two things are undeniable. First, the desire to develop AGI exists, and second, the number of AI systems that may have a functional impact on our society is rising. At the same time, public concerns are becoming more urgent and alarming (Russell, Dewey et al. 2015). AI can potentially threaten our autonomy and social relationships (e.g. Złotowski et al. 2017, Gladden 2014, Seibt et al. 2014, Sharkey and Sharkey 2010, Bryson 2010) and influence our understanding of what it means to be human (Al-Rodhan 2015). AI or AGI could be a threat to humanity either through its malicious use by humans (Brundage et al. 2018), unintended consequences (Amodei et al. 2016), or autonomous acts by the artificial system (Farquhar et al. 2017). In sum, the development of AI continues to pose questions of trust and risk (see Dafoe 2018). In order to find out whether it is justified to speak about trust in AI we should know what we are doing when we trust.

3. The nature of trust and trustworthiness

There is no unanimous agreement on what trust is. A usual way of explaining trust is to start from the question of what it is for humans to live socially (Luhmann 1979, Simpson 2012). The nature of trust has been explained by describing the requirements that have to be fulfilled. "Trusting requires that we can, 1) be **vulnerable** to others (vulnerable to betrayal in particular); 2) think well of others, at least in certain domains; and 3) be optimistic that they are, or at least will be, competent in certain respects" (McLeod 2015).

As trust is valuable when placed in trustworthy agents and activities, but damaging or costly when misplaced, we have to place or refuse trust intelligently

(O'Neill 2018). What it means to be trustworthy has been variously understood. Russell Hardin (1996, 2002) presents trustworthiness as compliance – if people are trustworthy, they will do what we want them to do. Karen Jones (2012) thinks that the trustee must consider the expectation of the trustor and use this as their primary motivation for being trustworthy. Stephen Wright (2010) takes a similar line of argumentation to that of Jones, but stresses that it is not important that the trustee considers the specific expectation of the trustor; instead, it is sufficient for trustworthiness that the trustee consider the value of the trusting relationship, which will motivate him to fulfil the wishes of the trustor. I tend to agree with yet a different view which stresses that to be trustworthy means to be competent and committed to do what the trustor expects the trustee to do. Here trustworthiness is a **property** or a virtue possessed by a trusted person (McLeod 2015). Trustworthiness may be specific (to a certain relationship) or full in extent (one has a disposition to be trustworthy to everyone) (Potter 2002: 8). These two conceptions, 'specific trustworthiness' and 'full trustworthiness' have also been described as thin and thick conceptions of trustworthiness. Philosophical accounts of trustworthiness diverge on the question of what could motivate a trustworthy person. The virtue account, which argues for a thick conception of trustworthiness, explains such motivation by the possession of the virtue – trustworthiness has been ingrained in the trustee's character. Others look for the source of motivation in good will or self-interest (either in the sense of avoiding social constraints or in the interest of holding a good relationship with the trustor).

Robert C. Solomon and Fernando Flores believe that "trust is a matter of reciprocal relationships", not of prediction, confidence, or reliance, and that it therefore makes sense to speak about trust only in relation to human agents and institutions (Solomon and Flores 2001: 14). Although some philosophers claim that trust is only interpersonal, I agree with Solomon and Flores that it is also justified to talk about institutional trust, since institutions consist of humans, and when we trust institutions we actually trust individuals who represent the institution. Also, we often trust individuals because they represent some profession, and we know what kind of values and standards the professional community is committed to follow. This commitment is often articulated in professional codes of ethics, which inform the public about what can be expected from a representative of this profession. Representatives of all professions must fulfil roles they have adopted and perform according to standards, as well as in view of certain values.

For example, a medical doctor must have knowledge of the physiology of the human being and about illnesses; diagnostic skills and the competence to heal; in addition they should have the attributes of goodwill, benevolence, non-maleficence, carefulness, empathy etc. We expect a bus driver to have an appropriate driver's licence and the requisite skills; that he is sober, well rested and friendly. We expect a baker not to put poison in the cake batter; a chef should not drop hairs in the soup. These are our expectations, but we do not visit the kitchen to confirm that this is so. If possible, we choose a better, more expensive restaurant, or an airline company with a recognized brand in order to be more certain that we are receiving better-quality

services: better-quality food or a safer trip. We assume that the firms involved are concerned about their reputation and afraid of losing clients. Similarly, we choose a school where we think our child will receive a better education (even though we cannot directly choose who will teach them). We choose an institution, not its staff, the people concerned; we trust the institution that has chosen the people.

4. Trust and reliance

The conceptual analysis of trust usually begins with making a distinction between trust and reliance. The most influential account of trust has been provided by Annette Baier (1986) who describes the difference between trust and reliance by means of a difference in our **reactions**. Breaches of trust by another person make us **feel betrayed**, whereas if we have simply relied on someone or something (e.g. a car to start) and our expectations are not fulfilled, we simply feel **disappointed** or angry. This view, that trust and reliance can be differentiated by our reactions, is shared by most of the philosophers who have written on trust. Where they diverge is in answering the question, what does trust involve?

According to Annette Baier, trust is reliance on another's **good will** or, in other words, trust is "accepted vulnerability to another's possible but not expected ill will (or lack of good will) toward one" (Baier 1986: 234). However, to my mind this is not a comprehensive account of what trust depends upon. Since in trusting one expects another to do what s/he has promised, it is not enough to believe that s/he has good will; one also has to believe that the other will be competent or capable of doing as s/he has promised. For example, we should not trust a doctor to treat an illness only on the basis of her/his good will alone. A patient's trust in a doctor also relates to the latter's **competence**. One trusts that s/he is up to date on medical information and that s/he has the necessary skills. Thus, it is evident that trust relates both to good will and competence, and that this is not only relevant to doctors, but also to other trustees.

An alternative suggestion has been made by Richard Holton (1994), who considers the **participant stance** to be the distinction between trust and reliance. In his view trust is a specific kind of reliance where something more is involved than just expecting that somebody will do something. According to Holton the participant stance differs from a simple attitude of expectation as follows: when we decide to place trust in somebody, taking the participant stance means either being ready to feel gratified, should s/he do as we expected, or else feeling betrayed by her/him in the event s/he fails to meet our expectations (Holton 1994: 67).

There is no guarantee that our expectations will not be betrayed. Although our decision to place trust is based on our beliefs and feelings about those trusted, these may be wrong. Granted, trust is earned by previous behaviour, but a record of previous trustworthiness (possessing a property or virtue that the trustor expects the trustee to possess) only shows the likelihood that the person can be trusted. Suppose there is a man who has always had a temptation to steal something but has been

too afraid of being caught. When a situation arises where it is likely that nobody will learn of his theft, he may follow his non-moral desire. Therefore his previous behaviour does not give us any guarantees. But the situation may also be reversed – a man who has acted badly in the past might sincerely want to improve his behaviour, but now nobody trusts him. Thus, our trust can be based on a false belief that can render either trust or mistrust inappropriate. But does trust always involve belief?

5. What does trust involve?

Philosophical accounts of trust differ on whether they argue that trust involves a belief that the trustee is trustworthy, that trust is an affective attitude, or both. Arnon Keren (2014) has recently explained the difference between doxastic and non-doxastic accounts of trust: doxastic accounts of trust hold that "trusting a person to (do, perform, enact) *Φ* involves, among other things, holding a belief about the trusted person; either that this person is trustworthy or that she actually will do (perform, enact) *Φ*" (Keren 2014: 2593). Non-doxastic accounts rely on the phenomenology of trust, and state that in trust the trustor holds some affective attitude towards the trusted person.

Richard Holton (1994: 63) has argued that in order to trust one need not believe. He gives an example of a shopkeeper who decides to trust his employee, although the latter has been convicted of petty theft. Holton argues that the shopkeeper can decide to trust the man without believing that he will not steal. He may trust him because he wants to give him moral support, a new chance to earn trust. This sort of trust has been called "therapeutic trust". This is certainly not an unlikely case. The way we treat former criminals or fellow humans who have done something bad, or how we treat our own children to help them grow in their sense of responsibility, show that we can trust without the belief that they are trustworthy. However, contrary to Holton, I think that when we do decide to trust, some **belief** must still be involved. This is not a belief about the likelihood of the other's behaviour, but simply a belief in the trustee's ability to change or grow. We cannot decide to trust when we do not believe that the other person can live up to our expectations (although certainly our expectations may be higher or lower in different cases).

Indeed, to suggest that trust involves belief (even if sometimes a false one), seems to indicate that it meets the requirements of rationality. In reality, however, our trust often goes beyond or against the available evidence. This can only be explained by the fact that trust also involves **an emotional element** which need not be rational. Just as we can be afraid of a dog, even if we know that it does not bite, we sometimes cannot help mistrusting someone despite evidence to the contrary. Although many critics have tried to limit trust by keeping it within the bounds of rationality, these attempts are not totally convincing. In my view, mistrust, as well as trust, may be both rational and irrational, and we should take such feelings and emotions seriously. Imagine a person who knows that Swissair is one of the most trustworthy companies, but is still afraid of flying it, since he remembers the recent plane crash that may have

been the only one during the last hundred years of the company's history. In this case mistrust is based more on feelings than a judgement of the likelihood that the plane might crash again. Such situations highlight the problem that trust is difficult to earn, but very easy to lose (because people seem to be willing to extend their negative experiences to the brand as a whole). A similar phenomenon also pertains to human relationships. For example, a woman might not trust her new male partner because she has had generally negative experiences with men (all men betray!).

Sometimes we place trust without any previous judgement and rely mostly on our feelings, whereas at other times our trust is based on a rational belief that the other will do something. Expectations can be based on either belief or emotion or a combination of both. Nevertheless, trust involves more than a belief and an affective mental state. In this respect I agree with Baier (1986) and Holton (1994) that there is something more going on in trust than in reliance. As distinct from them, however, I think that it is not enough to describe this 'more' in terms of the participant stance or the readiness to develop certain feelings. There has to be some explanation why we feel betrayed when the trustee does not fulfil our expectations.

My source of inspiration for developing my own account of trust has been the Wittgensteinian approach of Olli Lagerspetz, who, in his monograph *Trust: The Tacit Demand* (1998) asks what we are doing when we speak of trust. In his words, "to see an action as an expression of trust is to see it as involving a demand – **a tacit demand** – not to betray the expectations of those who trust us" (1998: 5). I think that Lagerspetz has put his finger on something very important by suggesting that trust is an act of communication in which the trustor sends out the message to the trustee that s/he is expected to fulfil the trustor's expectations. However, we have to go further and ask what kind of act the trustee is performing.

If we follow Lagerspetz's suggestion, we could say that the act of trusting involves not only an expectation that the trustee will act in a certain way, but also a tacit demand that the trustee not betray the expectations of the trustor; in his or her reply to the offer of trust, the trustee makes a tacit promise to do so. Although it may indeed sometimes happen that we decide to place our trust without having perceived the trustee's commitment to act in a certain way, in most cases trusting is a reaction to a certain disposition of the trustee. Usually, the trustee has previously communicated his competence and good will, and in the form of a tacit promise made clear his commitment to do what is expected. On the basis of this indication, the trustor is shaping the desire that the trustee will act as he trusted him to. This is the reason why in the event that the trustor's expectations are not met, the trustor will feel betrayed, not simply disappointed.

We are now better equipped to explain the difference between reliance and trust. Let us consider an example: a shopkeeper leaves a shop employee alone in the shop with a significant amount of money in the till, while she goes out on some important errand. Both reliance and trust can be possible in this situation. The shopkeeper can rely on the employee without trusting him, because he knows that the employee is afraid of sanctions in case he is caught taking the money. If he does not do as expected, the shopkeeper will feel disappointed. In the case of trust, something else

is involved besides the expectation that the employee will have such a fear and act in accordance with it, that is, not take the money. In addition, the shopkeeper has to have some positive feelings about his employee, and he has to desire that the man will prove to be trustworthy; this is accompanied by the disposition to feel betrayed if the employee does not succeed.

Thus my analysis of trust has identified a fourth element in the attitude of trust, alongside belief, affective mental state, and participant stance. This fourth element is **a desire** that the trustee turn out to be trustworthy either in a specific relationship or in general (we can then say that the person has moral integrity). What kind of mental state is desire? In contemporary philosophy of mind mental states are distinguished on the basis of "directions of fit": beliefs are like declarative sentences, which are satisfied (made true) by whether or not the world as such conforms to them. Alternatively, desires can be likened to imperative sentences, which are satisfied (fulfilled) by changes in the world, bringing the world into conformity with them (Gregory 2012). Thus we could say that in trusting the trustor desires the world to conform to his desire that the trustee will be trustworthy and do what we expect him to be or do.

Trust turns out to be a **reciprocal activity**, whereas reliance can be described as **one-way traffic**. We can rely both on other people or inanimate objects. For example, we rely on a computer to start, and an alarm clock to show the right time, or the sun to rise. Reliance involves one's expectation that somebody or something will do what he or it is expected to do on the ground of one's assessment of his reliability or its accountability. *Reliance* has to do with predictability and law-like regularities: one simply makes a prediction about what the other person or inanimate object is going to do, taking their previous behaviour and all relevant conditions into account.

Trusting is a communication process that includes acts of communication from both the side of the trustor and the trustee. On the trustee's side it involves expression of the trustee's competence and good will, as well as her/his commitment to do what is expected. On the trustor's side, trust involves a desire that the trustee will prove to be trustworthy, a positive feeling towards the trustee, and a participant stance to act in a certain way (either feeling betrayed or gratified).

6. Trust and risks

If no risks were involved, talking about trust would not make sense. Both trust and reliance entail **risk**. Both are directed toward the future functioning of someone or something without actually knowing their future. If we rely on the alarm clock to show the right time, we do this on the basis of our previous experience and knowledge of the quality of Swiss clocks. If we rely on the sun to rise tomorrow, we rely on our experience and our knowledge of natural laws, but we actually do not have complete certainty that the sun will indeed rise. We have even less certainty about people, whose behaviour is influenced by so many factors, such as different motivational forces and weakness of will.

What are our reasons for trusting if there is so much uncertainty involved? The answer is that we have to rely on others, as we are social beings and we need to cooperate. At the same time we have to accept some level of risk or vulnerability, or else we couldn't fulfil our aims. The trustor is vulnerable to the trustee's failure to do what s/he depends on that person to do. Consequently, the trustor might try to reduce this risk by monitoring or imposing certain constraints on the trustee's behaviour; however, perhaps after a certain threshold, the more monitoring and constraining the trustor does, the less s/he trusts that person (McLeod 2015).

Aron Keren (2014) has recently argued that reasons for trust are **"preemptive reasons"** (a term coined by J. Raz) – reasons that justify refraining from taking precautions, or reasons that oppose weighing the available evidence of somebody's trustworthiness. Keren insists that unless we are responding to preemptive reasons, we are not engaged in trusting. In Keren's view, trust is not compatible with excessive precautions. For example, if a shop-owner leaves her employee alone in the shop with a significant amount of money in the till, but before leaving she turns on the CCTV camera to monitor the employee's movements, she is not really trusting her employee. "The extent of our trust is at least partially determined by the degree to which we see ourselves as having a preemptive reason against taking precautions and the degree to which we act accordingly" (Keren 2014: 2605). Keren's account helps us to understand why trust involves being vulnerable to others: if one cannot take any precautionary measures, one will be dependent on the trustee's good will or some other motives to fulfil the trustor's expectations. However, if there is not enough evidence about the trustworthiness of the trustee, one should not relinquish all precautions.

Trust is a matter of **extent**. We have a scale, on the one end of which is full trust and on the other, complete mistrust. Full trust is confidence in someone, relying on them without doubting them; confidence that the trustor will fulfil our expectations, without any need for control or precautionary mechanisms. At the other end of the scale from trust is mistrust, where there is no trust at all, and there is a concomitant desire for control. Both trust and mistrust may be **rational** or **irrational**. Perhaps we have no reason to trust, but we trust anyway, non-rationally. Alternatively, there may be reasons to trust, but we refuse them. Irrational trust or irrational mistrust mean that we are not fully cognizant of risks; that we do not have sufficient knowledge of the trustee; that we do not fully take evidence or previous experience into account. Whereas irrational trust ignores the risks that something could betray trust, irrational mistrust overestimates the possibilities of betrayal. What we need is rational trust which is "open to evidence and the possibilities of betrayal" (Solomon and Flores 2001: 65). But reflection on possible risks and betrayal should not lead to cynicism, which is a refusal of trust. Both rational trust and rational mistrust are reflective; both entail considering risks, and both involve rational choice.

Public trust in new technologies depends on various aspects and elements (Sutrop 2007, 2010, Sutrop and Laas-Mikko 2012). It is equally important to have trustworthy institutions, to choose trusted persons to run them; thirdly, it is important to provide reliable systems in relation to data protection and securing privacy of data subjects,

as well as the safety and security for the users. All this will bring us to the issue of how to maintain trust in AI.

Creating trustworthy institutions does not in itself secure trust. As Solomon and Flores have stated, "One can be perfectly trustworthy but, because of circumstances or the paranoia of everyone involved, not be trusted" (Solomon and Flores 2001: 77). For example, people may express mistrust in innovative firms or even governments because they suspect that commercial interests are in play. A potential source of tension is the controversy between the public interest in new services on the one hand, and the potential commercial benefits of private companies on the other.

Trust is typically construed as risk-taking – one has to place trust without guarantees. Since trust involves vulnerability and risk, it is important to discuss such possible risks. New technologies are often intrusive, disruptive and potentially dangerous. Society will accept such risks only if scientists are able to provide assurance that the benefits of technological advances can be achieved without undue moral risk. Public trust in new technologies also depends on the behaviour of scientists and developers, as well as the public understanding of science and acceptance of the applications of new scientific developments. Public mistrust is often a response to prior untruthfulness and the abuse of trust, but it may instead be caused by an uneasiness that accompanies rapid scientific and technological progress. Trust can be destroyed if some scientists and developers do not follow the rules of good practice and are caught in dishonesty or conflict of interest. More broadly, trust also depends on whether people trust scientists and designers to dedicate themselves to responsible research and innovation, and whether they believe that society will be able to control and maintain the risks which new technologies supposedly introduce.

If we are unaware of the potential risks of new technologies, we cannot protect ourselves. In speaking about risk, we are discussing and arguing about something which is *not* the case but *could* happen if we do not to change a course of action. Ulrich Beck, who coined the term 'risk society', has pointed out that the discourse of risk begins where trust in our security and belief in progress end. It ceases to apply if a potential catastrophe actually occurs. Thus the concept of risk points to a "peculiar, intermediate state between security and destruction, where the perception of threatening risks determines thought and action" (Beck 2000: 213).

Risk assessment should be an essential part of the public discussion of each new technology, including AI. Risk assessment is based both on our ***imagination*** of what might happen and on our ***empirical knowledge*** of what is actually happening. The task of ethicists is to analyse and weigh the risks, as well as to disclose and justify the values endangered by perceived risks. Awareness of risks provides protection against the breach and potential violation of trust. Restoration of breached trust takes a long time, and it is therefore important to engage in continuous reflection about how to create and maintain trust.

7. When is it justified to talk about trusting AI and trustworthy AI?

Having now completed the conceptual analysis of trust, let us now come to the question, when is it justified to speak about trusting AI and the trustworthiness of AI? Trust is a social relationship (Solomon and Flores 2001), and as such it involves social cognition – thinking about the motives, reasons and beliefs of another person. The issue of trusting AI is bound up with the question of whether or not we treat AI systems as agents with mental capacities (Taddeo 2010a). At the moment AI fulfils limited tasks: it is used for making health diagnoses, scheduling appointments, recognizing people's voices and faces in photos, etc. However, AI is getting smarter and smarter, and soon intelligent robots may participate in our society, as self-driving cars, as caregivers for elderly people or children, and in many other ways. A common view is that AI has to align with our values and social norms to earn our trust. "Since society depends on cooperation, which depends on trust, if robots are to participate in society, they must be designed to be trustworthy" (Kuipers 2018: 90).

Several authors have written about trust in AI (e.g. Coeckelbergh 2012, Taddeo and Floridi 2011, Sethumadhavan 2018, Kuipers 2018, Hengstler et al. 2016, Taddeo 2010a, 2010b), but little attention has been paid to the difference between trusting AI as a machine which merely fulfils some human functions, and AGI, or SAI which possesses decision-making competency. It has been pointed out that questions about AI should specify to which type of AI they correspond (Bostrom and Yudkowsky 2014). However, so far there has been no analysis of trust with respect to different kinds of AI (narrow AI, AGI, SAI). Mark Coeckelbergh (2012) argues that in science-fiction scenarios where robots fulfil the criteria of language use, free will and social relations, even mutual trust can be an issue. However, even if today robots do not appear human and do not fulfil these criteria, he believes that "we have sufficient functional, agency-based, social-relational, and existential criteria left to talk about, and evaluate, trust in robots" (Coeckelbergh, 2012: 53). He thinks that our trust depends on our cultural attitude towards technology and robots in particular. Mariarosaria Taddeo argues that trust-based interactions are possible even with the current artificial intelligence since the anthropocentric criteria (freedom and language) can be replaced by operational autonomy and interactivity (Taddeo 2010a: 12).

Recently, Joanna Bryson published a paper with the provocative title, "No One Should Trust Artificial Intelligence" (2018). Her claim is that no one *can* or *should* trust AI. Firstly, in Bryson's terms trust is a relationship between peers, in which the trusting party – while not knowing for certain what the trusted party will do – believes any promises that are being made. As, in Bryson's opinion, AI is simply a set of system development techniques and therefore does not qualify as a 'peer', only other software development techniques can be peers with AI, and since these do not have the capacity to trust, no one actually *can* trust AI.

Secondly, Bryson claims that no human *should need* to trust an AI system. Her view is that AI is not a thing to be trusted. Rather, by means of AI we should be increasing the trustworthiness of our institutions and ourselves. Our aim should be

to engineer AI for accountability: "When a system using AI causes damage, we need to know we can hold the human beings behind that system to account" (Bryson 2018). Bryson points out that there is a real need for transnational coordination of the enforcement of standards for developing safer and more stable (intelligent) software systems. She stresses that AI systems should not be presented as being responsible, nor should they be construed as legal entities: "Like any other manufactured product, either the manufacturer or the owner/operation must be accountable for any damage it causes. Otherwise, malicious actors will attempt to evade liability for the software systems they create by blaming the system's characteristics, such as autonomy or consciousness" (Bryson 2018).

I think that Bryson has convincingly shown that talking about trust in AI instead of reliability, and about trustworthy AI instead of reliable or accountable AI may have serious consequences. She is certainly right that the responsibility for the functioning of AI should rest on human beings and institutions and not with software systems. However, it seems that we can speak of trust in AI in two cases: when we speak about human-like AI or when we mean the individuals and institutions behind AI systems. When the object of our attitude is narrow AI which is designed to fulfil only some specific tasks, we should better describe our attitude towards it as **reliance**. Only if AI is capable of deciding the best action to achieve some complex goal, is it justified to use the notion of **trust**. However, it may well be that when we speak about trust in AI, in reality we are speaking about trust or distrust of individuals and institutions who are responsible for developing, deploying and using AI. In order to avoid confusion, we should make the object of our attitude of trust clearer. As a matter of fact, trust can have different objects: we can trust individuals or institutions who are designing and manufacturing AI, who own AI, or who are responsible for regulating AI, who are overseeing its use, or providing conditions for its development and use. Since there are so many different parties involved in the development, deployment and use of AI, there are various trust relationships, all of which matter for eventual success.

Using the example of nine case studies in the transportation and medical technology industries, Hengstler et al. (2016) shows how trust in applied AI is fostered by firms. They point out that despite the growing use of automation with inherent AI, there is still noticeable scepticism in the society. They claim that the reasons for innovation resistance go beyond the technical characteristics of the products. The acceptance of the new technological product depends on its societal acceptance of the product (social context and cultural values matter!) and on trust in the innovating firm and its communication (which should emphasize how an innovation is compatible with the current lifestyle of users).

Hengstler et al. (2016) claim that trust in automation has three bases: performance, process and purpose (referring to Lee and See 2004). Their multiple cases research showed that in order to initiate performance trust it is important to develop operational safety and to define standards: "prior to use, a technology must be certificated and approved, and policies established to govern it" (Hengstler et al. 2016: 105). Second, data security turned out to be a factor that was considered

important across all cases and regardless of industry sector. In the automotive sector the stress was on security standards that protect the safety paths and sensor systems of vehicles, while in the healthcare industry, privacy protection was considered to be the major issue (how data is used and who has access to what kind of data). Process information describes how the automation operates and refers to understandability which was illustrated in the studies in three ways: first it turned out that the users tend to trust automation if the algorithms are understandable and guide users towards achieving their goals, and, second, if users are invited to participate in trials, thirdly, the usability studies must be designed so as to find out what amount of control the target use group is willing to delegate to the machine in the specific situation. For Hengstler et al. purpose is the third determinant of trust in the technology, requiring that potential users will understand why an innovation has been made and how it is compatible with the current lifestyle of the users (Hengstler et al. 2016: 108–111).

In order to be sure that AI is technically robust and reliable, we need transparency and explanations. Wolter Pieters (2011) has clarified that any explanation related to trust in technology is founded on how a system works, based on revealing details of its internal operations. Explanations related to confidence (corresponding to what we have called 'reliance') are founded on what makes the user feel comfortable in using the system, by means of providing information on its external communications. In trust-related explanations, the black box of the system is opened, which is not the case, however, in explanations relevant to confidence (Pieters 2011: 57). Here confidence is described as self-assurance concerning the safety or security of a system without knowing the risks or considering alternatives, while trust denotes self-assurance achieved by means of assessing risks and alternatives.

Alan F. T. Winfield and Marina Jirotka (2018) have shown that building trust in AI will require a multiplicity of approaches, from building trust at the level of individual systems and application domains to those at an institutional level but the key element is ethical governance. They define **ethical governance** as a set of processes, procedures, cultures, and values designed to ensure the highest standards of behaviour, and they propose five activities as its pillars: publishing of an ethical code of conduct, providing ethics and RI training to everyone, practising responsible innovation, being transparent about ethical governance, and really valuing ethical governance (Winfield and Jirotka 2018: 10). The way in which these scholars describe the implementation of their framework shows that they are fully aware that the most difficult part of it is the creation and maintenance of transparency, and how institutional ethical policies are translated into practice. How to produce realistic and workable ethical codes or regulations in the field of AI has been recently shown by Paula Boddington in her book *Towards a Code of Ethics for Artificial Intelligence* (2017).

8. What is the value of trust?

Trust is regarded as essential by all authors who write on AI (Kuipers 2018, Lee et al. 2015, Li et al. 2007, Winfield and Jirotka 2018) as without trust, the economic and social benefits of AI will not be actualized. We want to be sure that if AI can or should act autonomously, in view of our best interests we can trust it and those who design it and take all the ethically relevant aspects into account; we want to trust that AI will not be easily abused. Although currently AI systems are narrowly dedicated to specific tasks, visionary scenarios consider the possibility that intelligence equal to our own, or even exceeding it, will be created. Various proposals have been made on how ethics could be integrated into such AI systems (Vakkuri and Abrahamsson 2018, Yu et al. 2018), ranging from the suggestion of a general framework for 'ethics by design' (Iphofen and Kritikos 2019) to the discussion of whether or not AI can become a moral agent (Etzioni and Etzioni 2018).

The task of integrating ethics into AI is most often associated with the value-alignment challenge. Value-alignment is a property of an intelligent agent which signifies that it can only pursue goals that are beneficial to humans (Soares and Fallenstein 2014, Soares 2015, Russell, Hauert et al. 2015). AI systems will operate with increasing autonomy and capability in complex domains in the real world. How can we ensure that they will obtain the right behavioural dispositions – the goals or 'values' needed to ensure that things will turn out well, from a human point of view? Can AI learn from life and tracking people's behaviour? It is understood that living entails all kinds of behaviours, including bad behaviour. One example is how AI learns hate speech from the internet. Similarly, in traffic, AI may learn road rage and violation of regulations.

One possible means of learning values is through narratives. Mark O. Riedl and Brent Harrison (2016) argue that if values cannot easily be enumerated by human programmes, they can be learned. They propose that an AI can learn human values and what it means to be human by reading stories, as stories are necessarily reflections on the culture and society in which they were produced. For example, fables and allegorical tales contain examples of good behaviour. However, this can be problematic because fables are liable to multiple meanings and may require additional interpretation.

Current debate is divided among three main camps: (1) those who believe that AI should be programmed to align with our currently-held values (Russell 2017, Riedl and Harrison 2016), or at most with what we would call value under idealized conditions (the so-called Coherent Extrapolated Volition approach; Yudkowsky 2004); (2) those who believe that AI should be programmed to follow one or more substantive principles or theories (whether or not these are held by the majority), e.g. a utilitarian approach (Bauer 2018), a virtue ethics approach (Howard and Muntean 2017), or the 'love all equally' approach in Prinzing 2017 (see Wallach and Allen 2009 for an overview); (3) those who would delegate the task of finding the correct moral theory to a superintelligence (Bostrom 2014).

Each approach faces serious challenges: (1) preferring currently held values would be arbitrary because values can progress, and the current ones might not be the best; neither is there any guarantee that we would all have the same values, even under ideal conditions; (2) selecting (as well as weighing) substantive principles would require a kind of theoretical agreement that is nowhere near normative ethics; in addition, principles, especially when implemented by a machine (Bostrom 2014: 146) seem to be subject to a 'perverse instantiation'; (3) it is doubtful how a machine would enjoy special access to moral truth. There are additional problems concerning the inscrutability of decisions made by autonomous algorithmic devices (Mittelstadt et al. 2016); even if a superintelligent AI were to find the correct moral theory, we probably wouldn't know what it is; rather, we would only know what AI tells us to do.

Value-alignment is also difficult because there does not seem to be any agreement on universal values, at least not on how they should be ranked. The European Commission *Ethics Guidelines for Trustworthy AI* list four principles: respect for human autonomy, prevention of harm, fairness, and explicability (EU Commission 2019b). However, at least in the context of research ethics, it has been argued that principles of individual autonomy and privacy can seriously hamper scientific research and innovation that aims to further the common good (Sutrop 2011a, 2011b). The need to develop new ethical frameworks focusing on more collective values, such as **reciprocity, mutuality, solidarity, citizenry and universality, has been advocated** since the millennium. In the last two decades there has been talk of a 'communitarian turn' in ethics (Chadwick and Berg 2001, Chadwick 2011).

The basic difference between liberal and communitarian ethical frameworks lies in the fact that liberalism claims that the individual is more important than the society in which s/he lives, whereas communitarianism regards society as more important than the individual. Liberalism stresses rights, communitarianism the common good. In the traditional communitarian model individual rights and interests are subordinate to the common good, however the more recent neo-communitarian approach tries to balance individual rights with community interests. This has not been taken into account in the EU Guidelines.

An additional question remains whether values are the only thing that has to be given to the ethical AI? An autonomous decision-maker must be a moral person, which entails far more criteria than ethical values. We know how difficult it is to teach values to human beings (Sutrop 2015), let alone what we can do for ethical AI. The value conflicts (e.g. privacy *versus* security, autonomy *versus* solidarity) and moral dilemmas make it difficult to programme robots to act ethically by simply giving them the 'right' values. Instead, we should rather provide them with the capacity of **moral deliberation to** decide when a certain principle applies, how to rank values in different contexts and how to balance values. In order to act morally one should also have a proper **moral motivation** (the wish to be a certain kind of person who wishes to live a good life and avoid sanctions); this implies that one wants to be a member of the moral community and that one is open for feedback. Will AI ever become human-like in this respect? How moral can the moral non-human agents be?

Amitai Etzioni and Oren Etzioni (2018) note that to be a moral agent requires a certain set of attributes: "a) Self-consciousness. If the agent is not aware of itself in any given situation, and of the alternative courses that might be followed, then no moral decisions can be rendered. (b) The agent must be aware that she can affect the situation. (c) The agent must be able to understand the moral principles to be employed in arriving at a particular moral choice. (d) The agent must have a motive to act morally. This involves having passions, as otherwise moral preferences are merely intellectual preferences with nothing to fuel the moral action. Some scholars also add that a will or intention is required…" (Etzioni and Etzioni 2018: 239-240). Mark Coeckelbergh (2009) proposes to redirect our attention from the question of what kind of mental states non-humans should have in order to count as moral agents to the question how humans perceive artificial non-humans because it is not their intentional state, but their performance that counts morally. Coeckelbergh argues that "humans are justified in ascribing *virtual* moral agency and moral responsibility to those non-humans that appear similar to themselves – and to the *extent* that they appear" (Coeckelbergh 2009).

There is also a question of control. Will we be able to keep control over AGI? We must develop AGI with the potential to control other AGI. There is a great attendant danger that AGI will extend beyond our control. This is an enormous force that could engender the destruction of humanity. We should think carefully about both potential benefits and harm that may result when AI-based autonomous systems become superintelligent. "When independent of the humans that created them, their true 'intelligence' is tested in terms of their status as 'moral beings'" (Iphofen and Kritikos 2019).

As there are more non-democratic countries than democratic ones in the world, we should be careful not to allow AI to be programmed to do something devastating. Somewhere there will always be efforts underway to create AI that will be destructive, for example, for military purposes. Some experts have questioned the use of robots in military combat, especially if such machines were to be given some degree of autonomous functions, e.g. being able to independently choose targets to attack with weapons. Perhaps we could conclude with what **Stuart Russell stated** in Geneva, at the United Nations Convention on Conventional Weapons meeting, at the presentation of an arms-control advocacy video entitled 'Slaughterbots' (2017): "Artificial intelligence's potential to benefit humanity is enormous, even in defence. …. But allowing machines to choose to kill humans will be devastating to our security and freedom. Thousands of my fellow researchers agree. We have an opportunity to prevent the future you just saw, but the window to act is closing fast" (Sample 2017).

In order to mitigate these risks, we need proper ethics governance of AI. However, whether the risks can be mitigated by aligning the goals of advanced AI with human values, or by designing AI as a fully autonomous moral agent who is motivated to act morally, or whether risks can be minimized by preventing AI from becoming an autonomous agent is an open research question. We have seen that various philosophical problems should be solved before efforts are made to program values into AI. There are plenty of questions about which philosophers have not

reached agreement: Are values universal or relative to culture? Are there absolute values or *prima facie values* which are context dependent? Is there moral progress? Can values be taught? Or how are they learned? Should AGI have a motive to act morally? And would this mean that AGI would be capable of feeling moral emotions (e.g. shame, guilt) in order for principles to have motivating force? Does AGI qualify as a moral agent? More research needs to be done on how to integrate ethics into AI, how to design reliable AI, and how to build trust in individuals and institutions which design, deploy and use artificial intelligence.

Acknowledgements

This work is based on research undertaken for the research project IUT20-5, funded by the Estonian Ministry of Education and Research and supported by the Centre of Excellence in Estonian Studies (European Union, European Regional Development Fund). I would like to thank the following members of the research team for literature review and helpful discussions on the topic of the paper: Bruno Mölder, Francesco Orsi, Mari-Liisa Parder, Kadri Simm, and Mats Volberg. I am also very grateful to Tiina Kirss for her careful editing and help with English expression and to Laura Lilles-Heinsar for her help with proof-reading.

Address:
Margit Sutrop
Department of Philosophy
University of Tartu
Jakobi 2
50090 Tartu, Estonia
E-mail: Margit.Sutrop@ut.ee
Tel.: +37 25207183

References

Al-Rodhan, N. (2015) "*The many ethical implications of emerging technologies*". *Scientific American, March 13.* Available online at <http://www.scientificamerican.com/article/the-many-ethical-implications-of-emerging-technologies/>. Accessed on 10 November 2019.

Amodei, D., C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané (2016) "*Concrete problems in AI safety*". *ArXiv*, 25 July, v2. Available online at <https://arxiv.org/abs/1606.06565>. Accessed on 10 November 2019.

Baier, A. (1986) "Trust and anti-trust". *Ethics* 96, 231–260.

Banja, J. (2019) "Welcoming the 'intel-ethicist'". *Hastings Centre Report* 49, 1, 33–36.

Bauer, W. A. (2018) "Virtuous vs. utilitarian artificial moral agents". *AI & Society,* 2018. doi:10.1007/s00146-018-0871-3.

Beck, U. (1992) "Risk society revisited: theory, politics, and research programmes". In B. Adam, U. Beck, and J. Loon, eds. *Risk society: towards a new modernity*, 211–227. Trans. Mark Ritter. London: Sage.

Beck, U. (2000) *The risk society and beyond: critical issues for social theory*. London: Sage.

Beck, U. (2016) *The metamorphosis of the world*. London: Polity Press.

Bessi, A. and E. Ferrara (2016) "Social bots distort the 2016 U.S. Presidential election online discussion". *First Monday*, 21. Available online at <https://firstmonday.org/ojs/index.php/fm/article/view/7090/5653>. Accessed on 10 November 2019.

Bien, N., P. Rajpurkar, R. L. Ball, J. Irvin, A. Park, E. Jones, et al. (2018) "Deep-learning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of MRNet". *PLoS Med* 15, 11, e1002699. doi:10.1371/journal.pmed.1002699.

Boddington. P. (2017) *Towards a code of ethics for artificial intelligence*. Cham: Springer.

Bostrom, N. (2014) *Superintelligence: paths, dangers, strategies*. Oxford: Oxford University Press.

Bostrom, N. and E. Yudkowsky (2014) "The ethics of artificial intelligence". In K. Frankish and W. M. Ramsey, eds. *Cambridge handbook of artificial intelligence*, 316–334. Cambridge: Cambridge University Press.

Brundage, M., S. Avin et al. (2018) *The malicious use of artificial intelligence: forecasting, prevention, and mitigation.* Available online at <https://maliciousaireport.com/>. Accessed on 10 November 2019.

Bryson, J. J. (2010) "Why robot nannies probably won't do much psychological damage". *Interaction Studies* 11, 2, 196–200. doi:10.1075/is.11.2.03bry.

Bryson, J. (2018) "*No one should trust artificial intelligence". Science & Technology: Innovation, Governance, Technology* 11, 14. Available online at <http://ourworld.unu.edu/en/no-one-should-trust-artificial-intelligence>. Accessed on 10 November 2019.

Coeckelbergh, M. (2009) "Virtual moral agency, virtual moral responsibility: on the moral significance of the appearance, perception, and performance of artificial agents". *AI & Society* 24, 2, 181–189.

Coeckelbergh, M. (2012) "Can we trust robots?". *Ethics and Information Technology* 14, 1, 53–60.

Chadwick R. and K. Berg (2001) "Solidarity and equity: new ethical frameworks for genetic databases". *Nature Reviews. Genetics,* 2, 318–321.

Chadwick R. (2011) "The communitarian turn: myth or reality?" *Cambridge Quarterly of Healthcare Ethics* 20, 4, 546–553

Chalmers, D. (2010) "The singularity: a philosophical analysis". *Journal of Consciousness Studies* 17, 9–10, 7–65.

Dafoe, A. (2018) *AI governance: a research agenda*. University of Oxford. Available online at <http://www.fhi.ox.ac.uk/govaiagenda>. Accessed on 10 November 2019.

Etzioni, A. and O. Etzioni (2018) "Incorporating ethics into artificial intelligence". In A. Etzioni. *Happiness is the wrong metric: a liberal communitarian response to populism*, 235–252. (Library of Public Policy and Public Administration, 11.) Cham: Springer. Available online at <https://www.springer.com/gp/book/9783319696225>. Accessed on 10 November 2019.

EU Commission (2019a) *A definition of AI: main capabilities and disciplines.* Available online at <https://www.aepd.es/media/docs/ai-definition.pdf>. Accessed on 10 November 2019.

EU Commission (2019b) *Ethics guidelines for trustworthy AI*. Available online at <https://ec.europa.eu/futurium/en/ai-alliance-consultation>. Accessed 10 November 2019.

Farquhar, S., J. Halstead, O. Cotton-Barratt, S. Schubert, H. Belfield, and A. Snyder-Beattie (2017) *Existential risk diplomacy and governance*. Global Priorities Project. Available online at <https://www.fhi.ox.ac.uk/wp-content/uploads/Existential-Risks-2017-01-23.pdf>. Accessed on 10 November 2019.

Gladden, M. E. (2014) "The social robot as 'charismatic leader': a phenomenology of human submission to nonhuman power". *Frontiers in Artificial Intelligence and Applications* 273, 329–339. doi:10.3233/978-1-61499-480-0-329.

Grace, K., J. Salvatier, A. Dafoe, B. Zhang, and O. Evans (2018) "*When will AI exceed human performance? Evidence from AI experts". ArXiv*. Available online at <https://arxiv.org/abs/1705.08807>. Accessed on 10 November 2019.

Gregory, A. (2012) "Changing direction on direction of fit". *Ethical Theory and Moral Practice* 15, 603–14.

Habermas, J. [1962] (1991) *The structural transformation of the public realm*. Thomas Burger, trans. Cambridge, MA: MIT Press.

Hardin, R. (1996) "Trustworthiness". *Ethics* 107, 26–42.

Hardin, R. (2002) *Trust and trustworthiness*. New York: Russell Sage Foundation.

Hawking, S. (2018) *Brief answers to big questions*. New York: Bantam Books.

Hengstler, M., E. Enkel, and S. Duelli (2016) "Applied artificial intelligence and trust – the case of autonomous vehicles and medical assistance devices". *Technological Forecasting & Social Change* 105, 105–120. doi:10.1016/j.techfore.2015.12.014.

Holton, R. (1994) "Deciding to trust, coming to believe". *Australasian Journal of Philosophy* 72, 63–76.

Howard, D. and I. Muntean (2017) "Artificial moral cognition: moral functionalism and autonomous moral agency". In T. M. Powers, ed. *Philosophy and computing*, 121–160. (Philosophical studies series, 128.) New York: Springer.

Hwang, T. and L. Rosen (2017) *Harder, better, faster, stronger: international law and the future of online PsyOps.* (ComProp Working Paper, 1.) Available online at <http://blogs.oii.ox.ac.uk/politicalbots/wp-content/uploads/sites/89/2017/02/Comprop-Working-Paper-Hwang-and-Rosen.pdf>. Accessed on 10 November 2019.

Iphofen, R. and M. Kritikos (2019) "Regulating artificial intelligence and robotics: ethics by design in a digital society". *Contemporary Social Science* 2041, 1–15. doi:10.1080/21582041.2018.1563803.

Jones, K. (2012) "Trustworthiness". *Ethics*, 123, 1, 61–85.

Keren, A. (2014) "Trust and belief: a preemptive reasons account". *Synthese* 191, 2593–2615. doi:10.1007/s11229-014-0416-3.

Kuipers, B. (2018) "How can we trust a robot?" *Communication* of the *ACM* 61, 3, 86–95. doi:10.1145/3173087.

Kurzweil, R. (2005) *The singularity is near*. New York: Viking.

Lee, J. D., and K. A. See (2004) "Trust in automation: designing for appropriate reliance". *Hum. Factors* 46 1, 50-80.

Lee, J.-G., K. J. Kim, S. Lee, and D.-H. Shin (2015) "Can autonomous vehicles be safe and trustworthy? Effects of appearance and autonomy of unmanned driving systems". *International Journal of Human-Computer Interaction* 31, 682–691. doi:10.1080/10447318.2015.1070547.

Lagerspetz, O. (1998) *Trust: the tacit demand*. Dordrecht: Kluwer Academic Publishers.

Li, X., T. J. Hess, and J. S. Valacich (2008) "Why do we trust new technology? A study of initial trust formation with organizational information systems". *Journal of Strategic Information Systems* 17, 39–71.

Lucas, G. M., J. Gratch, A. King, and L.-P. Morency (2014) "It's only a computer: virtual humans increase willingness to disclose". *Computers in Human Behavior* 37, 94–100. doi:10.1016/j.chb.2014.04.043.

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

McLeod, C. (2015) "Trust". In Edward N. Zalta, ed. *The Stanford encyclopedia of philosophy.* Available online at <https://plato.stanford.edu/archives/fall2015/entries/trust/>. Accessed on 10 November 2019.

Mittelstadt, B. D., P. Allo, M. Taddeo, S. Wachter, and L. Floridi (2016) "The ethics of algorithms: mapping the debate". *Big Data & Society* 3, 1–21.

Müller, V. and N. Bostrom (2016) "Future progress in artificial intelligence: a survey of expert opinion". In V. Müller, ed. *fundamental issues of artificial intelligence*, 553–571. (Synthese Library, 376.) Springer.

O'Neill, O. (2018) "Linking trust to trustworthiness". *International Journal of Philosophical Studies* 26, 1, 293–300.

Pieters, W. (2011) "Explanation and trust: what to tell the user in security and AI?" *Ethics and Information Technology* 13, 53–64. doi:10.1007/s10676-010-9253-3.

Potter, N. N. 2002. *How can i be trusted? A virtue theory of trustworthiness*. Lanham, Maryland: Rowman & Littlefield.

Prinzing, M. (2017) "Friendly superintelligent AI: all you need is love". In V. Müller, ed. *The philosophy & theory of artificial intelligence*, 288–301. Berlin: Springer.

Riedl M. and B. Harrison (2016) *Using stories to teach human values to artificial agents*. The Workshops of the Thirtieth AAAI Conference on Artificial Intelligence, AI, Ethics, and Society. February 12–13, 2016. (Technical Report, WS-16-02: AI, Ethics, and Society.) Phoenix, Arizona, USA.

Russell, S. (2017a) "Provably beneficial artificial intelligence". In *The next step: exponential life*. BBVA OpenMind. Available online at <https://people.eecs.berkeley.edu/~russell/papers/russell-bbvabook17-pbai.pdf>. Accessed on 10 November 2019.

Russell, S., D. Dewey, and M. Tegmark (2015) "Research priorities for robust and beneficial artificial intelligence". *AI Magazine* 36, 4, 94–105.

Russell, S., S. Hauert, R. Altman, and M. Veloso (2015) "Robotics: ethics of artificial intelligence". *Nature 521*, 7553, 415–418. doi:10.1038/521415a.

Sample, I. (2017*) "*Ban on killer robots urgently needed, say scientists". *The Guardian* 13 November. Available online at < https://www.theguardian.com/science/2017/nov/13/ban-on-killer-robots-urgently-needed-say-scientists>. Accessed on 10 November 2019.

Seibt, J., R. Hakli, and M. Nørskov, eds. (2014) *Sociable robots and the future of social relations*. (Frontiers in Artificial Intelligence and Applications, 273.) IOS Press. Available online at <https://www.iospress.nl/book/sociable-robots-and-the-future-of-social-relations/>. Accessed on 10 November 2019.

Sethumadhavan, A. (2019) "Trust in artificial intelligence". *Ergonomics in Design* 27, 2, April 1. doi:10.1177/1064804618818592.

Sharkey, N. and A. Sharkey (2010) "The crying shame of robot nannies: an ethical appraisal". *Interaction Studies* 11, 2, 161–190. doi:10.1075/is.11.2.01sha.

Simpson, T. W. (2012) "What is trust?" *Pacific Philosophical Quarterly* 93, 550–569.

Slaughterbots (2017) Arms-control advocacy video. Directed by S. Sugg, produced by M. Nelson, and written by M. Wood. Available online at <https://www.youtube.com/watch?v=9CO6M2HsoIA>. Accessed on 11 November 2019.

Soares N. and B. Fallenstein (2014) *Aligning superintelligence with human interests: a technical research agenda*. (Technical Report, 2014-8.) Machine Intelligence Research Institute.

Soares, N. (2015) *The value learning problem*. (Technical Report, 2015-4.) Machine Intelligence Research Institute.

Solomon, R. and F. Flores (2001) *Building trust in business, politics, relationships, and life*. Oxford: Oxford University Press.

Strout, J. (2014) "Practical implications of mind uploading". In R. Blackford and D. Broderick, eds. *Intelligence unbound: the future of uploaded and machine minds*, 201–211. Wiley.

Suarez-Serrato, P., M. E. Roberts, C. Davis, and F. Menczer (2016) "*On the influence of social bots in online protests: preliminary findings of a Mexican case study"*. *ArXiv*. Available online at <https://arxiv.org/abs/1609.08239>. Accessed on 10 November 2019.

Sutrop, M. (2007) "Trust". In M. Häyry, R. Chadwick, V. Arnason, and G. Arnason, eds. *The ethics and governance of human genetic databases*, 190–198. Cambridge: Cambridge University Press.

Sutrop, M. (2010) "Ethical issues in governing biometric technologies". In A. Kumar and D. Zhang, eds. *Ethics and policy of biometrics*, 102–114. Heidelberg: Springer-Verlag.

Sutrop, M. (2011a) "Changing ethical frameworks: from individual rights to the common good?". *Cambridge Quarterly of Healthcare Ethics* 20, 4, 533–545.

Sutrop M. (2011b) "How to avoid a dichotomy between autonomy and beneficence: from liberalism to communitarianism and beyond". *Journal of Internal Medicine* 269, 4, 375–379.

Sutrop, M. and K. Laas-Mikko (2012) "From identity verification to behaviour prediction: ethical implications of second-generation biometrics". *Review of Policy Research*, 29, 1, 22–36.

Sutrop, M. (2015) "Can values be taught? The myth of value-free education". *Trames* 19, 2, 189–202.

Złotowski, J., K. Yogeeswaran, and C. Bartneck (2017) "Can we control it? Autonomous robots threaten human identity, uniqueness, safety, and resources". *International Journal of Human Computer Studies* 100, 48–54. doi: http://doi.org/10.1016/j.ijhcs.2016.12.008.

Taddeo, M. (2010a) "Modelling trust in artificial agents: a first step towards 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.

Taddeo, M. and L. Floridi (2011) "The case for e-trust". *Ethics and Information Technology 13*, 1, 1–3. doi:10.1007/s10676-010-9263-1.

Tegmark, M. (2017) *Life 3.0: being human in the age of artificial intelligence*. Allen Lane.

Terrasse, M., M. Gorin, and D. Sisti (2019) "Social media, e-health, and medical ethics". *Hastings Centre Report* 49, 1, 24–22.

Vakkuri, V. and P. Abrahamsson (2018) "*The key concepts of ethics of artificial intelligence"*. *IEE International Conference Engineering, Technology and Innovation, 17.06.-19.06.2019*. Sophia Antipolis.

Varun, H. B, A. Irfan, and M. Mahiben (2018) "Artificial intelligence in medicine: current trends and future possibilities". *British Journal of General Practice* 68, 668, 143-144. doi:10.3399/bjgp18X695213.

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

Winfield, A. F. and M. Jirotka (2018) "Ethical governance is essential to building trust in robotics and artificial intelligence systems". *Philosophical Transactions of the Royal Society* A376. 20180085. http://dx.doi.org/10.1098/rsta.2018.0085.

Wright, S. (2010) "Trust and trustworthiness". *Philosophia* 38, 615–627. doi:10.1007/s11406-009-9218-0.

Yu, H., Z. Shen et al. (2018) "*Building ethics into artificial intelligence". Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18), 13-19 July 2018.* Stockholm. Available online at <https://www.ijcai.org/proceedings/2018/0779.pdf>. Accessed on 10 November 2019.

Yudkowsky, E. (2004) *Coherent extrapolated volition*. San Francisco, CA: The Singularity Institute. Available online at <https://intelligence.org/files/CEV.pdf>. Accessed on 10 November 2019.