Philosophy & Technology (2021) 34:1429–1447
https://doi.org/10.1007/s13347-021-00465-4

RESEARCH ARTICLE

Trust and Trust-Engineering in Artificial Intelligence Research: Theory and Praxis

Melvin Chen[1]

Received: 29 March 2021 / Accepted: 2 July 2021 / Published online: 5 August 2021

Abstract
In this paper, I will identify two problems of trust in an AI-relevant context: a *theoretical* problem and a *practical* one. I will identify and address a number of skeptical challenges to an AI-relevant theory of trust. In addition, I will identify what I shall term the 'scope challenge', which I take to hold for any AI-relevant theory (or collection of theories) of trust that purports to be representationally adequate to the multifarious forms of trust and AI. Thereafter, I will suggest how trust-engineering, a position that is intermediate between the modified pure rational-choice account and an account that gives rise to trustworthy AI, might allow us to address the *practical* problem of trust, before identifying and critically evaluating two candidate trust-engineering approaches.

Keywords Trust · Trustworthiness · Problems of trust · Trust-engineering · Cunning of trust

1 Two Problems of Trust

There are two problems of trust in the domain of AI research: a *theoretical* problem and a *practical* one. According to the *theoretical* problem, how are we to theorize about trust in an AI-relevant sense? The *theoretical* problem of trust is about whether an AI-relevant theory of trust is feasible or articulable. According to the *practical* problem, how ought we to design and engineer our AI systems to address any trust-relevant worries and concerns? In Section 2, I will characterize trust as a quaternary relation and identify an accompanying set of conditions to be satisfied before we have sufficient grounds for believing that a relation of trust exists between the trustor and the trustee. In Section 3, I will outline the positions of a number of skeptics vis-à-vis AI-relevant theories of trust: the reliance-without-trust skeptic, the no-intentionality

✉ Melvin Chen
melvinchen@ntu.edu.sg

[1] Philosophy, Nanyang Technological University, Singapore, Singapore

skeptic, the value-neutrality skeptic, and the distrust-as-default-position skeptic. In Section 4, I will address the skeptical challenges that have been outlined in Section 3 and propose an account of intentionality and related conceptual manoeuvres (e.g. the overruling of the fact-value distinction) to defuse the force of various skeptical challenges. In Section 5, I will highlight a fifth skeptical challenge that I shall term the 'scope challenge'. In Section 6, I will propose what I shall term a 'trust-engineering approach' to address the *practical* problem of trust. My attempts to address both the *theoretical* problem of trust and the *practical* problem of trust will culminate in a discussion of how the reward function may be constructed to encourage reinforcement learning-based agents to respond in a trust-responsive fashion, which I take to be a novel contribution to both the philosophy of AI and AI research.

2 Trust as a Quaternary Relation and Conditions for Satisfaction

Before we discuss about the *practical* problem of how we might increase the level of trust in human-AI interactions, we must first address the extant skepticism about AI-relevant theories of trust. According to Baier (1986), trust is a phenomenon with which we are so familiar that we scarcely notice its presence and its variety, whether in the form of our putting our bodily safety into the hands of pilots, drivers, or doctors, refraining from suspecting that the food we purchase may be deliberately poisoned, or trusting that our children will be fine in their day-care centres. Trust is also ubiquitous in civil society: we trust both individuals who are demonstrably trustworthy (loyal, virtuous, and prudent) and complete strangers (Pettit, 1995). AI is a similarly ubiquitous phenomenon: alongside such other emerging technologies as nanotechnology, quantum computing, and biotechnology, AI is the poster child for the Fourth Industrial Revolution and the use of AI systems across a multiplicity of domains is becoming more rather than less widespread (Schwab, 2017). Given the ubiquity of both trust and AI, we should have good reason to expect the untrammelled development of AI-relevant theories of trust. What plausible grounds for skepticism might exist with respect to AI-relevant theories of trust?

I shall assume from the outset and without further argument that trust is a quaternary relation R(A, B, ϕ, G) consisting of four relata: the trustor A, a trustee B, some action ϕ to be performed, and a goal G that makes the performance of ϕ desirable (see Fig. 1).[1] More specifically, trust is a mental state that A holds toward B with respect to the performance of ϕ relevant for the goal G (Castelfranchi & Falcone, 1998). A theory of trust, taking this quaternary relation as basic for conceptual analysis, typically identifies the conditions that must be satisfied before we have sufficient grounds for believing that there is a relation of trust between A and B with respect to a G-relevant ϕ. Traditional approaches in the philosophy of trust have taken both A and B to be persons and concerned themselves with an analysis of interpersonal trust.

[1] This idea extends the trinary relation of Horsburgh (1961), Holton (1994), and Hardin (1992): A trusts B to perform ϕ. Not all relations of trust fit the trinary or quaternary relation schemata: A might trust B generally and in a way that does not involve *any* particular action or task (Carter & Simion, 2020).

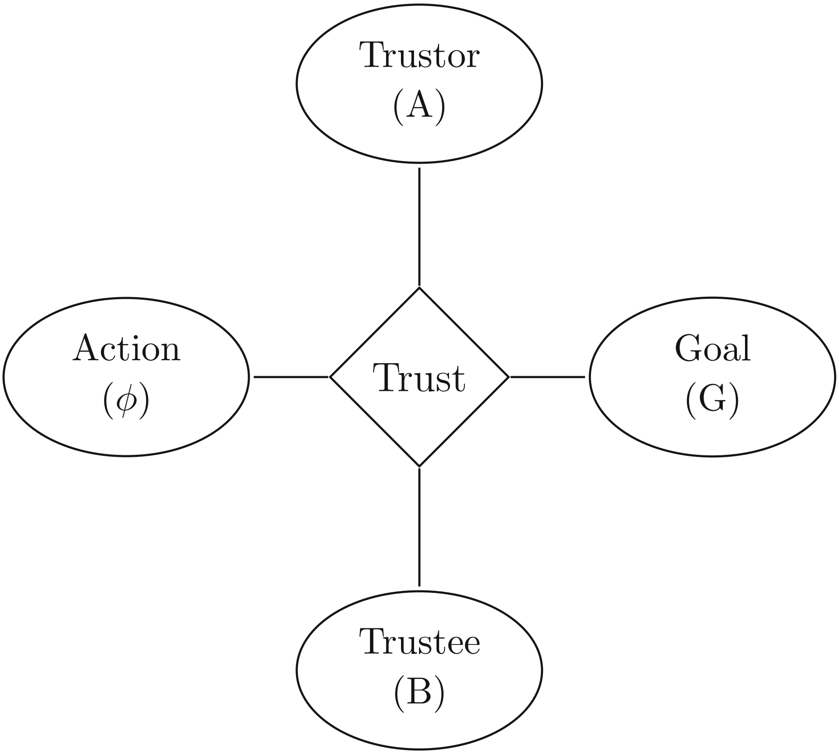

Fig. 1 Trust as a quaternary relation (degree = 4)

The AI-relevant theory of trust with which I am concerned identifies B as an artificial agent and ϕ as an anticipated action, event, or decision-outcome that is brought about by B.

Some theories of trust maintain that only one condition has to be satisfied before a relation of trust obtains between A and B: a probability threshold condition. Trust is grounded in probabilities that A attributes to her own beliefs about the behaviour and competences of B with respect to the G-relevant ϕ. On a probability distribution with the parameter p $\in$ [0, 1], 1 denotes complete trust, 0 denotes complete distrust, and 0.50 denotes uncertainty. Let n denote the probability threshold value and let m denote the probability value that A attributes to her trust-relevant beliefs. If $m \geq n$, then the sole condition will be satisfied and there will be a trust relation between A and B (Gambetta, 1998). These theories of trust characterize trust in terms of mere reliance. Reliability (regularly designated by R) has been defined by engineers as probability that the item can perform a required function under given conditions for a stated time interval (Birolini, 2013, p. 334).[2] If both A and B are persons and our object of conceptual analysis is interpersonal trust, then B figures more or less as an operator who is expected to perform a G-relevant action ϕ. B is said to be reliable if

[2] Where the duration T is considered as a variable t, the reliability function is given by R(t).

she can perform ϕ under given conditions for a stated time interval. Alternatively, it will be urged by these theorists that A can trust B with respect to the G-relevant ϕ for the stated time interval. Furthermore, it may be suggested that an AI-relevant theory of trust that is consistent with these theoretical foundations will claim the following: A trusts B, where B is an artificial agent, to a degree m iff A is willing to risk the use of B on the basis that it will perform the G-relevant ϕ with probability m, where $m \geq n$.[3] Following Nickel et al. (2010), we may term this candidate AI-relevant theory of trust a 'modified pure rational-choice account'.

3 Four Skeptical Challenges to AI-Relevant Theories of Trust

The modified pure rational-choice account makes an important connection between trust and contexts where risk is present. In addition, there is no fundamental difference between a person being reliable and an AI system being reliable. According to the first skeptical challenge ('reliance-without-trust'), however, a conceptual distinction could be still made between trust and mere reliance. The probability threshold condition turns out to be insufficient: AI systems could satisfy the probability threshold condition (i.e. $m \geq n$), yet we merely rely on rather than trust these systems. Whereas misplaced reliance results in disappointment, misplaced trust tends to lead to feelings of resentment and betrayal toward the trustee when let down (Hieronymi, 2008; Wanderer & Townsend, 2013). Trusting, it may be conjectured, is not an attitude that I can adopt toward machines: I may feel disappointed when my computer fails to save important documents as it should, although it would seem strange for me to feel betrayed by or resentful of this malfunctioning computer (Jones, 1996). A necessary condition of trust that has been omitted by the modified pure rational-choice account is the normative expectation condition. A normative expectation is in place when A relies on B to do what B should do, and A does not only expect *that* B will do it but also expects it *of* B (Walker, 2006, p. 79). Insofar that A has a normative expectation that B will perform ϕ, A's normative expectation brings with it a requirement of responsibility on the part of B. If the normative expectation condition is fulfilled and B is responsible for the G-relevant ϕ that A normatively expects her to perform, then A's feelings of resentment and betrayal toward B when B fails to perform ϕ may be explained in terms of B's dereliction of responsibility.

In addition, some theories of trust maintain of the trustee B that it should possess the relevant goodwill. Since only things that have wills can have goodwill, it follows that we can only trust agents that have wills. According to the second skeptical challenge ('no-intentionality'), given that many open questions remain about the concepts of volition, autonomy, rationality, moral responsibility, intentionality, and consciousness and whether AI systems possess them, we should suspend judgment about whether an artificial agent can qualify as a trustee B (Johnson, 2006; Himma, 2009). The no-intentionality skeptic might rely on Ullmann-Margalit (2004), for whom the trust relation is characterized as follows: A trusts B iff A believes that B has

[3] See Nickel et al. (2010).

the appropriate intentions toward her (primary component) and B has the appropriate competence with respect to the G-relevant ϕ (secondary component).[4] The primacy of the intentionality requirement allows the no-intentionality skeptic to distinguish between instances of trust and instances of reliance and confidence. As AI systems lack the relevant intentionality, they can neither take into account interests that they might have, act in favour of the interests of the trustor A because they care about A or possess goodwill toward A, nor experience the possibility of a conflict of A's interests and their own. We can neither normatively expect it *of* AI systems that they will do such-and-such nor predicate of AI systems that they possess wills. Therefore, the reliance-without-trust skeptic and the no-intentionality skeptic will agree that no AI-relevant theory of trust can be developed.

According to the third skeptical challenge ('value-neutrality'), AI systems, as stand-alone implementations, are normatively neutral. This skeptical approach is undergirded by a thesis (viz. the value-neutrality thesis) that applies to AI systems in particular and technologies in general. According to this thesis, a new piece of technology T can have bad consequences only if people have vicious T-relevant preferences or if users with minimally decent preferences act out of ignorance. Conversely, T can have good consequences only if people have minimally decent T-relevant preferences or if users with vicious T-relevant preferences act out of ignorance (Morrow, 2014). AI research is concerned with understanding and building intelligent entities (Russell & Norvig, 2010). AI research, it may be argued, typically relies on mathematical facts (e.g. in linear algebra, differential calculus, mathematical logic, probability), cognitive facts (e.g. Hebb's learning rule of synaptic reinforcement, the McCulloch-Pitts model of the neuron), and physical facts (e.g. I am at my desk, my car is at home). If we accept this characterization of AI research and the standard epistemological distinction between statements of fact (descriptive) and statements of value (normative), then we appear to have good reason to believe that AI systems are normatively neutral as stand-alone implementations. If we are to theorize about trust in an AI context, then the trustee B must be a human user of an AI system rather than the system itself and we are back in the traditional milieu of interpersonal trust.

According to the fourth skeptical challenge ('distrust-as-default-position'), the default position with respect to AI systems ought to be one of distrust rather than trust. This account is an especially powerful one in regimes that exclude marginalized groups from trust networks, cast oppressed individuals as untrustworthy, and prevent the oppressed from having their knowledge claims trusted by both themselves and others (Fricker, 2007; Daukas, 2011). Consider how algorithms, employed by judges and parole officers to assess a criminal defendant's likelihood of recidivism, have demonstrated a bias against certain marginalized groups (Angwin et al., 2016). Where epistemic injustices of this sort exist and may be exacerbated by the use of AI systems and technologies, we have good reason to maintain a skeptical default posi-

[4] According to Ullmann-Margalit (2004), if A trusts in B's competence but not her intentions, then we would probably say that A distrusts B. This explains the primacy of B's intentions over B's competence with respect to the G-relevant ϕ in Ullmann-Margalit's account of trust.

tion of distrust rather than trust. I am sympathetic to the view that distrust, betrayal of trust, and even strategies of resistance could be more appropriate when AI systems result in the maintenance or exacerbation of epistemic injustices. The productive and protective value of distrust in the face of tyranny and injustice and the ability of distrust to temper tyranny cannot be underestimated (Krishnamurthy, 2015). When AI systems function as tools of tyrannical and unjust regimes, there is democratic value in observing a default position of distrust.[5]

The reliance-without-trust skeptic argues that the probability threshold condition, while it may be satisfied by certain reliable AI systems, is insufficient: we do not trust these systems in the way that we might other human beings. The no-intentionality skeptic argues that AI systems lack the relevant intentionality to qualify as appropriate trustees. The value-neutrality skeptic argues that AI systems are normatively neutral as stand-alone implementations and cannot qualify as trustees, although their human users can. The distrust-as-default-position skeptic argues that an AI-relevant theory of distrust may be more useful than an AI-relevant theory of trust. Both the reliance-without-trust skeptic and the no-intentionality skeptic will incline toward the impossibility of developing an AI-relevant theory of trust. The value-neutrality skeptic, in addition, falls back on traditional accounts of trust that have interpersonal trust as their object of analysis. The distrust-as-default-position skeptic, aware of how AI systems might result in the maintenance or exacerbation of epistemic injustices, recommends distrust rather than trust as the default position.

4 Addressing the Skeptics

Skeptics about the possibility of developing an AI-relevant theory of trust must recognize our growing vulnerability to AI systems and the effects that they bring about. Consider our vulnerability to a GPS system misfiring, self-driving cars malfunctioning, medical diagnostic programs making erroneous diagnoses, and machine learning-based lending algorithms making flawed credit decisions against us. The more we work under the supposition that AI systems will perform requisite tasks into our plans, the more we count on them to do something.[6] While information processing systems may still have been relatively simple in the 1990s, we are in an era when these systems are increasingly complex, powerful, indispensable, and automated in their computational capacities. We increasingly count on AI systems, in the sense that we embed in our plans assumptions about what these AI systems will do in ways that leave us vulnerable if they do not.[7] Consequently, the idea (as defended by the

[5] Although my argument does not require it, one could also consider how a betrayal of trust, when it expands trust networks to involve the oppressed, could be construed as epistemically virtuous (Frost-Arnold, 2014).

[6] This point becomes especially salient once we realize that tasks may be nested within other tasks. The more tasks there are to consider (viz. tasks nested within tasks, tasks nested within tasks-within-tasks, etc.), the higher the probability that AI systems will be involved and therefore counted on at some point (typically where the competences of AI systems match or exceed the competences of humans and the case for automation, whether full or partial, is strongest).

[7] This formula for counting on is from Holton (1994).

reliance-without-trust skeptic) that we are merely relying on these AI systems seems less plausible now than it might have been in the 1990s. To be perfectly clear, risk and vulnerability do not seem to be the sole issue that matters in trust. While we are sufficiently vulnerable to tornadoes and other forms of inclement weather, we do not feel resentment toward tornadoes or other forms of inclement weather if our metereological expectations are overturned.[8] Our growing vulnerability to AI systems, combined with an account of intentionality that encompasses both human beings and certain technical artifacts, increases the argumentative burden on skeptics who pooh-pooh theorizing about trust in an AI-relevant sense as being out of place. The more vulnerable we become to instances in which our AI systems fail to act as expected, even as these AI systems have states that are directed at objects or states of affairs in the world, the more likely we are to feel betrayed.

In addition, skepticism about the possibility of theorizing about trust in an AI-relevant sense appears to be counterintuitive. While it has been claimed that there has (at least until recently) been relatively little discussion about trust in the context of AI and human-AI interactions (Ribeiro et al., 2016), an ever-growing corpus of research on trust in the context of AI suggests otherwise. A casual Google Scholar search using the Boolean search string '~trust AND ("artificial agents" OR "AAs") AND ("AI" OR "Artificial Intelligence")' for articles published since 2020 yielded c. 2,540 search results.[9] Groundbreaking work on the concept of e-trust and how it is appropriate for any analysis involving artificial agents (Floridi & Sanders, 2004; Taddeo, 2009; Taddeo & Floridi, 2011) has inspired theories of trust that pertain to AI, artificial agents, robots (Buechner & Tavani, 2011; Grodzinsky et al., 2011), social robots (Coeckelbergh, 2012), and (more recently) both human-human and human-AI interactions (Ferrario et al., 2019).

A proper account of intentionality may help to defuse the force of various skeptical challenges. One need not agree with the no-intentionality skeptic that AI systems lack the relevant intentionality. If one can demonstrate how AI systems might have intentional states, then one may be more justified in having certain normative expectations with respect to these systems. A distinction between original intentionality and derived intentionality could be invoked here. It may be claimed that AI systems have derived intentionality rather than no intentionality whatsoever. When an AI system consists of a symbol manipulation system with formal elements and syntactic rules, it may demonstrate some form of intelligent behaviour (e.g. chess-playing, theorem-proving, natural language processing). However, this intentionality is derived rather than original, since the syntactic rules and formal elements cannot represent beliefs, desires, expectations, and other intentional mental states. Original intentionality, on the other hand, may be thought to refer to the intentionality

[8] After all, as the no-intentionality skeptic will be at pains to remind us, AI systems, tornadoes, and other forms of inclement weather lack the appropriate intentionality. I owe this point to an anonymous reviewer for *Philosophy & Technology*.

[9] This search was conducted on 29 May 2021. If a more generic input string '~trust AND "AI" OR "Artificial Intelligence"' is used and the other filter parameters are held constant, then c. 20,700 search results are obtained. The tilde operator (~) allows Google to expand the search to related keywords and was applied to the term 'trust' in the search string.

of mental states: this intentionality is not derived from some more prior forms of intentionality but is intrinsic to these mental states themselves (Searle, 1983).

Furthermore, it may be argued that these AI systems borrow whatever derived intentionality that they might have from the original intentionality of human beings who design, implement, and use them for various purposes. AI technologies are tools that are designed and used in service of their respective ends, goals, and purposes. As products of human intentional action, they have a prima facie claim to some form of derived intentionality (Ihde, 1990; Latour, 1992; Verbeek, 2008). The use of the distinction between original intentionality and derived intentionality is in line with an increasing tendency to use notions such as autonomy, agency, choice-making, and morality (traditionally for describing and explaining intentional human behaviour) to describe and explain the behaviour of AI systems as products of intentional human behaviour (Nickel et al., 2010).

The no-intentionality skeptic may well concede that AI systems have derived intentionality rather than no intentionality, while maintaining that derived intentionality is irrelevant in the context of trust. After all, the distinction between derived and original intentionality is traditionally predicated on an account of intentionality that defends a conjunction of the following claims (Searle, 1980, 1984):

(C1) No matter how sophisticated or complex it might be, an AI system is never by itself a sufficient condition of intentionality.

(C2) Intentionality is a biological phenomenon that is causally dependent on biochemistry.

While a chess-playing program is capable of demonstrating intelligent behaviour through the rule-governed manipulation of formally specified elements, C2 implies that intentionality must still be traced to the designers, programmers, and even users of this chess-playing program and imputed to the causal powers of their brains.[10] We may have plausible reservations about the account of intentionality that is given by the conjunction (C1 $\wedge$ C2). While an AI system might consist solely of a symbol manipulation system with formal elements and syntactic rules, it could also comprise additionally of a set of transducers (tying the computational system in some sense to the outside world) and an etiology or context (the environment in which the computation-cum-transducer system finds itself). Given the right input and the right history and context, it has been argued that an embodied computational system could have (rudimentary) intentional states, insofar that this system's states are about or directed at objects and states of affairs in the world (Bynum, 1985). Even in the absence of human beings, artificial agents may continue interacting with each other

[10] See Sayre (1986) for an articulation of this view. After C1 has been affirmed, the no-intentionality skeptic might argue that there is no case to answer. Nonetheless, an appeal could still be made to the more conservative reading of my position (see Section 7) and the role and relevance of trustworthy AI ecosystems.

and their environment and maintaining these (rudimentary) intentional states, giving the lie to C1 and C2.[11]

A few more things could be said in response to the value-neutrality skeptic and the distrust-as-default-position skeptic. The value-neutrality skeptic relies from the outset on the standard epistemological distinction between fact and value. However, this fact-value distinction is an artificial one and ignores the leaky nature of the fact/value divide: facts appear to seep into values even as we observe a leakage from values to facts. In addition, AI technologies constitute more than merely applied science. The goal-directed and purposive nature of AI systems, as aforementioned in the claim in favour of their derived intentionality, puts paid to the idea that they are normatively neutral. It may also be advanced against the distrust-as-default-position skeptic that she is relying on certain assumptions about the nature of the regime in which AI systems are implemented. However, not all regimes are characterized by tyranny and injustice, AI systems could be designed as tools to aid marginalized groups, and even regimes given to tyranny and injustice could well use AI systems for other more mundane purposes than the maintenance and exacerbation of epistemic injustices. There are at least some instances in which we ought to avoid excessive trust with respect to regimes that are supported by oppressive and exclusive norms. Where these regimes are supported by AI systems (e.g. speech recognition technologies, mass surveillance technologies), distrust of these systems would be favourable as a default position and even healthy. However, for other more ideal moral climates in which people enjoy their freedoms, trust flourishes, and AI systems are used as tools for promoting societal ends, distrust may be less relevant as a default position.[12]

5 The Scope Challenge

Notwithstanding the skeptical challenges from the reliance-without-trust skeptic, the no-intentionality skeptic, the value-neutrality skeptic, and the distrust-as-default-position skeptic (Section 3), attempts to address the *theoretical* problem and articulate an AI-relevant theory of trust have not been in short supply. In this section, I will identify a fifth skeptical challenge that lies in wait. I shall term this latter challenge the ‘scope challenge’. In Section 2, I advanced that trust is a quaternary relation R(A,

[11] In a related vein, several studies in human-computer interaction have identified an anthropomorphic phenomenon according to which humans read intentionality (e.g. the possession of beliefs, desires, and expectations) into AI systems. Roomba robot vacuums are described as ‘crazy’, ‘refined’, or ‘dead, sick, or hospitalized’ (Sung et al., 2007) and computational toys are described in terms of their ability to ‘cheat’ (Turkle, 2005). If my analysis is correct, then this human trait of reading intentionality into AI systems might not be a category mistake in at least some instances. See also Dennett (1987, 1996) on the intentional stance and Zhu (2009) for an extended treatment of the intentional stance in the context of AI systems.

[12] All things considered, evidence seems to point to the conclusion that several regimes and states of affairs are characterized by injustice. Given the documented instances of bias being built into seemingly objective AI algorithms, perhaps some modicum of distrust-as-default-position skepticism may be welcome. Too much of this skepticism may, however, lead to paralysis or inaction and function as an impediment to change. Again, I am grateful to the anonymous reviewer from *Philosophy & Technology* for pointing this out to me.

B, ϕ, G) and further that a set of accompanying conditions must be satisfied if a relation of trust is to exist between A and B with respect to a G-relevant ϕ. Additionally, a number of plausible conditions (viz. the probability threshold condition, the normative expectation condition) were identified in Sections 2 and 3. Suppose that the force of the four skeptical challenges is sufficiently defused and the prospects for an AI-relevant theory of trust are renewed. Suppose that an attempt is made to identify the full suite of conditions that are necessary and sufficient for the relation between the trustor A and the trustee B to count as a trust relation. Last but not least and in accordance with my delineation of an AI-relevant theory of trust, suppose that B is identified as an artificial agent and ϕ is identified as an anticipated action, event, or decision-outcome that is brought about by B. Is the AI-relevant theory of trust or collection of theories representationally adequate to the multifarious forms of trust and AI or not? This is the scope challenge in the interrogative mood.

Trust exists in a spectrum and ranges from full-fledged trust, through substantial trust, therapeutic trust, trust in the face of doubt or entrusting, and agnosticism between trust and distrust to distrust itself.[13] In a similarly multifarious vein, AI systems have been developed from a variety of approaches to AI research: the laws of thought (thinking rationally) approach, the cognitive modelling (thinking humanly) approach, the Turing test (acting humanly) approach, and the reinforcement learning-based (acting rationally) approach (Russell & Norvig, 2010; Bringsjord & Govindarajulu, 2020).[14]

An AI-relevant theory of trust that acknowledges the force of the scope challenge and seeks to address it in a head-on fashion will strive to represent the nature of the relationship between our notions of trust and AI in a manner that is representationally adequate to both notions and consistent across their multifarious realizations.[15] Such a theory or collection of theories will recognize the complex phenomenology of trust and engage seriously with the possibility of an AI-relevant account of trust that extends over its diverse manifestations. The philosophy of trust distinguishes between the disappointment that arises from misplaced reliance and the betrayal and resentment that arise from misplaced trust. Feelings of betrayal and resentment belong, as other feelings such as gratitude and moral anger, to a class of attitudes that

[13] Full-fledged trust is associated with what Hieronymi (2008) has termed ‘trusting belief’. Substantial trust neglects or abjures strategic judgments and renounces the process of evidence-weighing (McGeer, 2008). A liberal and non-purist understanding of trust will allow that a trustor A trusts the trustee B to the extent that she incurs vulnerability to betrayal and concede that trust could exist even in the face of doubt. With therapeutic trust, the trustor relies on the trustee, with the aim of bolstering the latter’s trustworthiness (Horsburgh, 1960). Humble trust is trust that issues from skepticism about the warrant of one’s own felt attitudes of trust and distrust (D’Cruz, 2019).

[14] A fifth approach to AI research, termed ‘hybrid AI’ or ‘neurosymbolic AI’, has recently been proposed by several AI researchers (Marcus, 2020; Mao et al., 2019). It has also been suggested that this fifth approach to AI research, unlike the four current approaches, may get us to AI that we could trust (Marcus & Davis, 2019). From this diverse range of approaches to AI research, we have AI systems of various designs, including automated theorem provers, genetic algorithms, artificial neural networks, expert systems, Bayesian networks, fuzzy logic-based programs, and robots with sensors and effectors.

[15] The bullet-biting alternative would consist of conceding the force of the skeptical challenge and admitting that at least some of the multifarious forms of trust and/or AI lie outside the scope of the AI-relevant theory or collection of theories.

we take toward agents rather than objects. This class of attitudes has been termed the ‘reactive attitudes’ and is intimately bound up with ascriptions of responsibility (Strawson, 1962). These feelings of betrayal and resentment are reactive attitudes that link trust to practices of holding agents responsible for their actions from the participant stance. The explanatory advantages of a theory of trust that relies on the Strawsonian reactive attitudes have been extensively catalogued by philosophers of trust (Holton, 1994; Jones, 2004; Walker, 2006; Hieronymi, 2008; McGeer, 2008). Following Nickel et al. (2010), I shall term a candidate AI-relevant theory of trust that is developed along these lines a ‘modified motivation-attributing account’.[16] The AI-relevant modified motivation-attributing account will provide sufficient room for the Strawsonian reactive attitudes and explain why A is willing, in certain circumstances, to count on B and make herself vulnerable. Where certain essential attributes of garden-variety trust in an interpersonal context (viz. intentionality, normative expectations) will be difficult or impossible to locate when B is identified as an artificial agent, conceptual clarification or related conceptual manoeuvres (as suggested in Section 4) may have to be effected. Environmental factors (e.g. an inclusive climate of trust as opposed to an oppressive climate of distrust), the domain under consideration, and the consequences of trusting or distrusting are among the variables that determine the appropriate default stance to adopt within the spectrum that extends from full-fledged trust to distrust.

In the second instance, we must recognize the variety of approaches to AI research and the variety of systems to which these distinct approaches give rise, while engaging seriously with the possibility of a an AI-relevant theory or collection of theories that extends over its diverse manifestations. Whatever intelligence may ultimately be, AI is constituted by artificial entities capable of simulating intelligence or exhibiting certain intelligence-relevant mental traits, often by performing tasks normally thought to require intelligence. These tasks include but are not limited to the following: the successful navigation in a physical space, knowledge representation and reasoning, learning and generalization, image recognition, problem-solving, inference-making, and natural language processing and understanding. AI systems are sufficiently advanced technologies that have been designed for a range of tasks, the successful performance of which tends to conduce to certain ends or purposes that matter to us. Certain expressions in the English language suggest a natural connection between trust and care: we say both that the trustor normally entrusts the trustee with something she cares about and that the trustor entrusts that cared-for something to the care of the trustee.

AI systems have been designed and introduced because there are certain things such as the goal G that we care about (i.e. certain things that matter to us), specific tasks such as ϕ the successful performance of which will tend to promote these things that we care about, a recognition that there are limits to our agential powers,

[16] According to the motivation-attributing account of trust (of which the AI-relevant theory is a modification), trust is an attitude of optimism that the goodwill and competence of the trustee B will extend to cover the relevant domain in question, plus the normative expectation that B will be directly and favourably moved by the thought that the trustor A is counting on her (Jones, 1996).

and the hopeful trust that tapping on the resources and competences of sufficiently advanced technologies will promote the successful performance of these tasks and help us along our way to G. The point about the trustor A (presumably a human being) accepting limitations on her own agency is especially crucial: after all, if A could easily and directly bring about some desired end, then hoping for that end would be out of place, since A would simply act so as to achieve it (McGeer, 2008). A's hope that B will perform the desirable and G-relevant action ϕ and A's coming to terms with the limitations of her own agency go hand-in-hand here.

As AI systems progress through the design, testing, implementation, and use phases, various stakeholders (viz. big tech companies, AI researchers, industry partners, organizations, users, etc.) will become involved in differing capacities. This collective will of the relevant stakeholders could be tapped on in the ethics and epistemology of trust: we simply need to ensure that we have what generally passes muster as goodwill across these various stakeholders. Whereas the value-neutrality skeptic is only prepared to consider human users of AI systems as candidate trustees, the implication here is that a larger group of human stakeholders will have to be countenanced. Perhaps it may be more appropriate to characterize the trustee B in the quaternary trust relation R(A, B, ϕ, G) as an AI ecosystem rather than an AI system.[17] In addition, it has been hypothesized that a set of drives may be identified across a broad class of AI systems and will emerge as a result of convergent paths of AI development (Omohundro, 2008; Bostrom, 2014). These drives include the drive to self-improvement, the drive to greater rationality, the drive to self-preservation, and the drive to resource acquisition and the efficient use of these resources. These basic AI drives are analogues of the human will, render AI systems purposive and goal-driven, and enhance their agential aspects. We should remind ourselves of the collective will of the relevant stakeholders and the basic AI drives, whenever skeptical questions are raised about volition, autonomy, rationality, moral responsibility, intentionality, and consciousness in the context of AI systems.

6 Trust-Engineering

The *practical* problem concerns how we should design and engineer our AI systems to address any worries and concerns about their trustworthiness.[18] Although trust and

[17] Such a philosophical move will preserve the intimate connections between the derived intentionality of AI systems and the original intentionality of human stakeholders who design, implement, and use them for various purposes, especially where the AI-relevant theory of trust relies on the distinction between derived and original intentionality.

[18] In what follows, you may get a lingering sense that while the theoretical framework for addressing trust-relevant issues in AI is sufficiently well-developed, the discussion of the *practical* problem of trust is less so. My principal aims in this paper with respect to the latter are modest: I wish to identify the *practical* problem of how we ought to go about designing and engineering our AI systems to address trust-relevant concerns, indicate connections between the *theoretical* problem of trust and the *practical* problem of trust in an AI-relevant context, and introduce the notion of trust-engineering. As a sidenote, I have recently been awarded a grant as a Co-I to further research on trustworthiness with respect to AI systems and have every intention of delivering a full articulation of the second candidate approach to trust-engineering. Such an articulation, given its scale, would however warrant a second paper.

trustworthiness are interlocking, they are also categorically distinct (Hardin, 2006; Nickel et al., 2010). Trust is a mental state that A holds toward B with respect to the performance of a G-relevant ϕ. Trustworthiness, on the other hand, is a quality in B that satisfies this mental state and helps to make it appropriate. Trustworthiness is an epistemic virtue that can be cultivated in human beings. Philosophers of trust typically defend trustworthiness as an epistemic virtue and characterize the bearer of this virtue as someone who can be counted on to avoid unduly violating the normative expectations that others rely upon her to meet (Frost-Arnold, 2014). Trustworthiness is also associated with other laudable traits such as loyalty, virtue, and prudence (Pettit, 1995). An individual who develops her dispositions to be loyal, virtuous, or prudent will cultivate her trustworthiness. It is not obvious how AI systems can be designed to develop their dispositions to be loyal, virtuous, and prudent, in a manner analogous to human beings. Given this important disanalogy between human beings and AI systems, we appear to have grounds to remain skeptical about whether the *practical* problem can be solved.[19]

We may arrest this skepticism by acknowledging a position that is intermediate between the modified pure rational-choice account (Section 2) and an account that gives rise to trustworthy AI. Let us call this middle-of-the-road position 'trust-engineering'. According to this account, while trustworthiness (viz. the possession of certain laudable traits) is a form of trust-reliability, it is not its only form. There is also trust-responsiveness, which is a disposition to prove reliable under the trust of others. Robust and reliable AI systems are apt to demonstrate competence in the domain wherein they are being counted on to perform particular tasks. Trust-engineers could therefore work at improving the robustness and reliability of these AI systems (e.g. with respect to out-of-distribution test examples, adversarial perturbations, absent supervisors) (Amodei et al., 2016). Beyond the mere fact of competence, it should be recognized that trust-responsiveness is a disposition to take the fact that another is counting on one to be a positive reason to act in a trustworthy manner, absent other independent reasons to be trustworthy.[20] The idea is that the presence of trust itself generates a reason to be trustworthy. Such responsiveness to trust requires a conscious awareness that one is being counted on. Trust-engineers should therefore also work at developing AI systems with the AI-encompassing account of intentionality in mind and steadfastly refrain from conflating trust-responsiveness with mere competence.[21]

One candidate trust-engineering approach may rely on certain physiognomic biases. To be clear, physiognomy (or the assessment of an individual's character or personality from her outer appearance and especially her facial features) is a junk science. However, there appears to be a number of physiognomic biases that relate to trustworthiness. Faces with low inner eyebrows, shallow cheekbones, and thin chins

[19] As skepticism about whether the *practical* problem grows, skepticism about the *theoretical* problem (as outlined in Section 2) may also be renewed.

[20] For more on trust-responsiveness and its ecological character, see McGeer and Pettit (2017).

[21] I am grateful to an anonymous reviewer for *Philosophy & Technology* for the guidance provided in this instance.

are perceived as untrustworthy, individuals who are attractive or appear happy (with smiling being used as a signal of the intention to cooperate) tend to be viewed as trustworthy, faces that resemble a baby are viewed as non-threatening, and we tend to trust people who appear similar to our tribe and distrust others who appear dissimilar (Scharlemann et al., 2001; Todorov et al., 2008; Sofer et al., 2017).

The ability that AI designers have to alter the outer appearance of AI systems in line with these physiognomic biases may be contrasted with the general inability that human beings have to alter their outer appearance (expect by plastic surgery or other extreme means). From an understanding of how these physiognomic biases work, a blueprint for designing the outer appearance of AI systems could be developed, such that the chances for these systems to be perceived as trustworthy are maximized.[22]

This candidate approach to trust-engineering poses serious problems. When the outer appearance of AI systems is altered in certain ways, trust may be evoked in the trustor, albeit through a mechanism that, due to its biased and irrational nature, does not in fact contribute to making the trustee more trustworthy. Where approaches to trust-engineering manipulate weaknesses in the human condition to engender trust, it makes perfect sense to ask: should such systems be developed in the first place? A far better mechanism on which to rely than a set of physiognomic biases is a psychological mechanism that has been identified in Pettit (1995).[23] Trustors rely on this mechanism of an esteem-seeking desire for the good opinion of others to get trustees to act in a trust-responsive fashion. This mechanism works because it is a fact of human psychology that people seek esteem and generally wish to be well thought of by others (Pettit, 1995). Broadly, desiring the esteem of the trustor A promotes trust, while at the same time promoting trustworthiness insofar as this esteem can be reliably maintained if the trustee B is also motivated to be trustworthy.[24]

Another candidate approach to trust-engineering that is more in line with Pettit's theory could involve working the esteem of others into the reward function of our AI systems. More specifically, if the esteem of others (both trustors who are counting on an AI system and reliable and interested third parties) could be captured and represented in an appropriate manner in the reward function, reinforcement learning-based agents might infer an optimal policy that selects actions in a trust-responsive

[22] Trust-engineers planning to exploit these physiognomic biases should however be aware of a much-observed phenomenon in robotics: the uncanny valley effect (Mori et al., 2012). According to this phenomenon, people's responses to humanlike robots will abruptly shift from empathy to revulsion as these robots approach but fail to attain a lifelike appearance. The uncanny valley is this descent into a sense of coldness and eeriness that characterize the emotional responses to humanlike robots.

[23] I should note in passing that there is much to be gained from a more sustained engagement by the broader public of philosophy of technology and the AI research community with Pettit's theory and his way of thinking about trust.

[24] While physiognomic trust cues are a type of cunning, they are distinct from the type of cunning with which Pettit's theory is concerned. Pettit's cunning of trust is a Hegelian cunning that transforms an a-moral instinct, desire, or disposition into a motivation that approaches the moral. I am grateful to an anonymous reviewer of a previous version of this manuscript for pointing this out to me.

fashion.[25] This could have important implications for the trust-responsiveness of our AI systems.

The reinforcement learning approach can be represented in terms of a Markov decision process. Each Markov decision process or MDP is a tuple of the form (S, Φ, P, γ, r), where S is the set of states (including the initial state s_0), Φ is the set of possible courses of action available to an agent (e.g. ϕ_1, ϕ_2, …, ϕ_n), P is the transition probability matrix (viz. P(s_1 | s_0, ϕ_i)), γ is the discount factor, and r is the reward function. My Pettit-inspired trust-engineering approach involves working esteem-based considerations into r. Alternatively, we could get our AI system to infer the reward function r by observing the behaviour of trustworthy exemplars. This approach is known as inverse reinforcement learning and is yet another tool that could be used in trust-engineering (Abbeel & Ng, 2004; Vasquez et al., 2014). More specifically, we could rely on inverse reinforcement learning to derive an esteem-sensitive reward function r from the policy π or behaviour of trustworthy exemplars, before working this esteem-sensitive r explicitly into our reinforcement learning-based approach.

Consider an AI-based credit scoring system. Its trustors $\mathrm{A}_1, \cdots, \mathrm{A}_\mathrm{n}$ include lenders and various financial institutions. Reliable and interested third parties may include economists and government agencies. When AI-assisted decisions are made by lenders to extend or deny credit, the esteem of $\mathrm{A}_1, \cdots, \mathrm{A}_\mathrm{n}$ and these reliable and interested third parties could be captured and represented after the manner of restaurant and hotel reviews. This esteem, aggregated across parties (whose scores may be weighted according to the degree of esteem in which they are themselves held) and rendered computationally feasible, gives rise to a particular reward or feedback signal that will indicate how well the AI system (the trustee B) is doing trust-wise at a particular step. Through its interactions with the environment, we have good reason to expect a reinforcement learning-based AI system to end up selecting actions in a trust-responsive fashion. Where inverse reinforcement learning is feasible, we could get our artificial agent to observe the policy π or behaviour of a trustworthy non-AI-based credit scoring exemplar (B*) and infer an esteem-sensitive reward function r.

This AI-based credit scoring system could take the form of a computation-cum-transducer system. Its transducers or 'sense organs' allow it to interact with borrowers, lenders, and interested third parties. The states of this system could be directed at real-time credit information on individuals and businesses, transactions and bank statements, reports from lenders and other sources, the esteem reports of reliable and interested third parties, and other states of affairs in the world. These states will have conditions of satisfaction: when the system requests a bank statement and biometric information from the borrower, this request utterance will not be satisfied until the borrower obliges accordingly. These states have directions of fit:

[25] I am aware of the problem of reward gaming or reward hacking, wherein an artificial agent may exploit certain loopholes or weaknesses in how the reward function has been (mis)specified to help itself to more than its fair share of the reward. The problem of reward gaming or reward hacking is however a general and technical aspect of the value alignment problem that confronts *all* AI researchers. It does not follow from a proposal's not being entirely foolproof that it ought not receive due and serious consideration.

world-to-world (in the case of declarative utterances about the borrower's creditworthiness) or world-to-word (in the case of imperative utterances, when commands or instructions are issued in the credit-scoring process). Given the right input from the world and the right history and context, this AI-based credit scoring system could have the sort of intentional states that we find described in Bynum (1985). Contra the no-intentionality skeptic, it will be far from the case that AI systems lack the relevant intentionality to qualify as appropriate trustees. Contra the reliance-without-trust skeptic, we may have good reason to trust these systems in the way that we might other human beings, given that they have intentional states and are apt to respond in a trust-responsive fashion.

7 Conclusion

In this paper, I have identified two problems of trust: a *theoretical* problem and a *practical* one. In Section 2, I identified trust as both a quaternary relation R(A, B, ϕ, G) and a mental state that the trustor A holds toward a trustee B with respect to the performance of a G-relevant ϕ. I also argued that a theory of trust typically identifies an accompanying set of conditions that must be satisfied before we have sufficient grounds for believing that a relation of trust exists between A and B. In Section 3, I entertained skeptical notes vis-à-vis AI-relevant theories of trust from four camps: the reliance-without-trust skeptic, the no-intentionality skeptic, the value-neutrality skeptic, and the distrust-as-default-position skeptic. In Section 4, I addressed these skeptical challenges to the *theoretical* problem and proposed an account of intentionality and related conceptual manoeuvres (e.g. the overruling of the fact-value distinction) that could be made against these skeptical overtures. In Section 5, I identified the scope challenge confronting any AI-relevant theory of trust or collection of theories that purports to be representationally adequate to the multifarious forms of trust and AI. In Section 6, I identified two candidate approaches to trust-engineering in response to the *practical* problem, one involving a certain reliance on a set of physiognomic biases and another relying on a set of esteem-seeking psychological mechanisms. Using Pettit's theory about the cunning of trust, I weighed in favour of the second candidate approach rather than the first. I then proposed the development of an esteem-sensitive reward function that may allow AI systems to respond in a trust-responsive fashion, aided by reinforcement learning and/or inverse reinforcement learning.

I have not excluded a more conservative reading of my position that would merely permit our speaking in terms of trustworthy AI ecosystems instead of trustworthy AI systems (Section 5). Given the disanalogy between human beings and AI systems, we could redirect our demands for trustworthiness, not to the AI systems themselves but rather to the AI ecosystems at large. We could encourage big tech companies, AI researchers, industry partners, organizations, and users to cultivate their dispositions to be loyal, virtuous, or prudent, extend our hopeful trust to them and galvanize members of the AI ecosystems to live up to our hopeful vision of what AI systems can do and be, and appeal to their goodwill. Nonetheless, the trust-engineering account may help us to secure a position that is intermediate between trustworthy

AI and reliance-without-trust. Botsman (2017) has argued that we are at the start of the third trust revolution in the history of human civilization. Distributed trust flows laterally between individuals, enabled by networks, platforms, and systems.[26] Distributed trust is the name of this third trust revolution: we trust other people through technology (including AI as its most advanced form), rate everything from chatbots to Uber drivers, and have come to rely on well-trained bots to give us advice, resolve our problems, and carry out our food orders (Botsman, 2017, p. 8). My philosophical attempt to address the *theoretical* and *practical* problems in an AI-relevant context may be situated within the vanguard of this third trust revolution.

Funding This research has been generously funded by the NTU Intramural Accelerating Creativity & Excellence (ACE) grant (c. SGD 200,000) and is part of a two-year interdisciplinary project on medical artificial intelligence.

References

Abbeel, P., & Ng, A. Y. (2004). Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning* (p. 1).

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. arXiv:1606.06565.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias. Propublica, May 23.

Baier, A. (1986). Trust and antitrust. *Ethics*, *96*(2), 231–260.

Birolini, A. (2013). *Reliability Engineering: Theory and Practice*. Springer Science & Business Media.

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

Botsman, R. (2017). Who Can You Trust? Penguin UK.

Bringsjord, S., & Govindarajulu, N. S. (2020). In E. N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University, summer 2020 edition*.

Buechner, J., & Tavani, H. T. (2011). Trust and multi-agent systems: Applying the diffuse, default model of trust to experiments involving artificial agents. *Ethics and Information Technology*, *13*(1), 39–51.

Bynum, T. W. (1985). Artificial intelligence, biology, and intentional states. *Metaphilosophy*, *16*(4), 355–377.

Carter, J. A., & Simion, M. (2020). The ethics and epistemology of trust. *Internet Encyclopedia of Philosophy*.

Castelfranchi, C., & Falcone, R. (1998). Principles of trust for MAS: Cognitive anatomy, social importance, and quantification. In *Proceedings International Conference on Multi Agent Systems (Cat. No. 98EX160)* (pp. 72–79). IEEE.

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

[26] Local trust exists between members of small communities and rests in specific individuals. Institutional trust flows upward to leaders, experts, and brands, and runs through institutions and intermediaries (e.g. courts, regulatory bodies, corporations). Both local trust and institutional trust are earlier chapters in the history of human civilization (Botsman, 2017, p. 257).

Daukas, N. (2011). Altogether now: A virtue-theoretic approach to pluralism in feminist epistemology. In *Feminist Epistemology & Philosophy of Science* (pp. 45–67). Berlin: Springer.

D'Cruz, J. (2019). Humble trust. *Philosophical Studies*, *176*(4), 933–953.

Dennett, D. C. (1987). *The Intentional Stance*. Cambridge: MIT Press.

Dennett, D. C. (1996). *Kinds of Minds*. Basic Books.

Ferrario, A., Loi, M., & Viganò, E. (2019). In AI we trust incrementally: A multi-layer model of trust to analyze human-artificial intelligence interactions. Philosophy & Technology, pp. 1–17.

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

Fricker, M. (2007). *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford: Oxford University Press.

Frost-Arnold, K. (2014). Imposters, tricksters, and trustworthiness as an epistemic virtue. *Hypatia*, *29*(4), 790–807.

Gambetta, D. (1998). Can we trust trust? In *Trust: Making & Breaking Cooperative Relations* (pp. 213–238). Blackwell.

Grodzinsky, F. S., Miller, K. W., & Wolf, M.J. (2011). Developing artificial agents worthy of trust: "would you buy a used car from this artificial agent?". *Ethics and Information Technology*, *13*(1), 17–27.

Hardin, R. (1992). The street-level epistemology of trust. *Analyse & Kritik*, *14*(2), 152–176.

Hardin, R. (2006). *Trust*. Polity.

Hieronymi, P. (2008). The reasons of trust. *Australasian Journal of Philosophy*, *86*(2), 213–236.

Himma, K. E. (2009). Artificial agency, consciousness, and the criteria for moral agency: What properties must an artificial agent have to be a moral agent? *Ethics & Information Technology*, *11*(1), 19–29.

Holton, R. (1994). Deciding to trust, coming to believe. *Australasian Journal of Philosophy*, *72*(1), 63–76.

Horsburgh, H. J. N. (1960). The ethics of trust. *The Philosophical Quarterly (1950-)*, *10*(41), 343–354.

Horsburgh, H. J. N. (1961). Trust and social objectives. *Ethics*, *72*(1), 28–40.

Ihde, D. (1990). *Technology and the Lifeworld: From Garden to Earth*. Indiana: Indiana University Press.

Johnson, D. G. (2006). Computer systems: Moral entities but not moral agents. *Ethics & Information Technology*, *8*(4), 195–204.

Jones, K. (1996). Trust as an affective attitude. *Ethics*, *107*(1), 4–25.

Jones, K. (2004). Trust and terror. In P. DesAutels, & M. U. Walker (Eds.) *Moral Psychology: Feminist Ethics & Social Theory* (pp. 4–25). Rowman & Littlefield.

Krishnamurthy, M. (2015). (White) tyranny and the democratic value of distrust. *The Monist*, *98*(4), 391–406.

Latour, B. (1992). Where are the missing masses? The sociology of a few mundane artifacts. In W. E. Bijker, & J. Law (Eds.) *Shaping Technology/Building Society* (pp. 225–258). MIT Press.

Mao, J., Gan, C., Kohli, P., Tenenbaum, J. B., & Wu, J. (2019). The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. arXiv:1904.12584.

Marcus, G. (2020). The next decade in AI: Four steps towards robust artificial intelligence. arXiv:2002.06177.

Marcus, G., & Davis, E. (2019). Rebooting AI: Building artificial intelligence we can trust. Vintage.

McGeer, V. (2008). Trust, hope and empowerment. *Australasian Journal of Philosophy*, *86*(2), 237–254.

McGeer, V., & Pettit, P. (2017). The empowering theory of trust. In P. Faulkner, & T. W. Simpson (Eds.) *The Philosophy of Trust* (pp. 14–34). Oxford: Oxford University Press.

Mori, M., MacDorman, K. F., & Kageki, N. (2012). The uncanny valley [from the field]. *IEEE Robotics & Automation Magazine*, *19*(2), 98–100.

Morrow, D. R. (2014). When technologies makes good people do bad things: Another argument against the value-neutrality of technologies. *Science and Engineering Ethics*, *20*(2), 329–343.

Nickel, P. J., Franssen, M., & Kroes, P. (2010). Can we make sense of the notion of trustworthy technology? Knowledge. *Technology & Policy*, *23*(3-4), 429–444.

Omohundro, S. M. (2008). The basic AI drives. In *AGI*, (Vol. 171 pp. 483-492).

Pettit, P. (1995). The cunning of trust. *Philosophy & Public Affairs*, *24*(3), 202–225.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the* 22nd *ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 1135–1144).

Russell, S. J., & Norvig, P. (2010). *Artificial Intelligence: A Modern Approach*. Pearson Education London.

Sayre, K. M. (1986). Intentionality and information processing: An alternative model for cognitive science. *Behavioral & Brain Sciences*, *9*(1), 121–38.

Scharlemann, J. P., Eckel, C. C., Kacelnik, A., & Wilson, R.K. (2001). The value of a smile: Game theory with a human face. *Journal of Economic Psychology*, *22*(5), 617–640.

Schwab, K. (2017). *The Fourth Industrial Revolution*. Currency.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral & Brain Sciences*, *3*(3), 417–57.

Searle, J. R. (1983). *Intentionality: An Essay in the Philosophy of Mind*. Cambridge: Cambridge University Press.

Searle, J. R. (1984). *Minds, Brains & Science*. Harvard: Harvard University Press.

Sofer, C., Dotsch, R., Oikawa, M., Oikawa, H., Wigboldus, D. H., & Todorov, A. (2017). For your local eyes only: Culture-specific face typicality influences perceptions of trustworthiness. *Perception*, *46*(8), 914–928.

Strawson, P. (1962). Freedom and resentment. In *Proceedings of the British Academy*, (Vol. 48 pp. 1–25).

Sung, J.-Y., Guo, L., Grinter, R. E., & Christensen, H.I. (2007). My Roomba is Rambo: Intimate home appliances. In *International Conference on Ubiquitous Computing* (pp. 145–162). Berlin: Springer.

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

Taddeo, M., & Floridi, L. (2011). The case for e-trust. *Ethics & Information Technology*, *13*(1), 1–3.

Todorov, A., Baron, S. G., & Oosterhof, N.N. (2008). Evaluating face trustworthiness: A model based approach. *Social Cognitive & Affective Neuroscience*, *3*(2), 119–127.

Turkle, S. (2005). *The Second Self: Computers & the Human Spirit*. Cambridge: MIT Press.

Ullmann-Margalit, E. (2004). Trust, distrust, and in between. In R. Hardin (Ed.) *Distrust* (pp. 60–82). Russell Sage Foundation.

Vasquez, D., Okal, B., & Arras, K.O. (2014). Inverse reinforcement learning algorithms and features for robot navigation in crowds: An experimental comparison. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems* (pp. 1341–1346). IEEE.

Verbeek, P.-P. (2008). Morality in design: Design ethics and the morality of technological artifacts. In P. E. Vermaas, P. Kroes, A. Light, & S. A. Moore (Eds.) *Philosophy & Design* (pp. 91–103). Berlin: Springer.

Walker, M. U. (2006). *Moral Repair: Reconstructing Moral Relations after Wrongdoing*. Cambridge: Cambridge University Press.

Wanderer, J., & Townsend, L. (2013). Is it rational to trust? *Philosophy Compass*, *8*(1), 1–14.

Zhu, J. (2009). *Intentional Systems & the Artificial Intelligence (AI) Hermeneutic Network: Agency & Intentionality in Expressive Computational Systems*. PhD thesis, Georgia Institute of Technology.