What is it like to trust a rock? A functionalist perspective on trust and trustworthiness in artificial intelligence

Peter R. Lewis *, Stephen Marsh
Ontario Tech University, Oshawa, ON, Canada

A R T I C L E I N F O
Keywords:
Trust
Trustworthiness
Artificial intelligence

A B S T R A C T
The trustworthiness (or otherwise) of AI has been much in discussion of late, not least because of the recent publication of the EU Guidelines for Trustworthy AI. Discussions range from how we might make people trust AI to AI being not possible to trust, with many points inbetween. In this article, we question whether or not these discussions somewhat miss the point, which is that people are going ahead and basically doing their own thing anyway, and that we should probably help them. Acknowledging that trust is a heuristic that is widely used by humans in a range of situations, we lean on the literature concerning how humans make trust decisions, to arrive at a general model of how people might consider trust in AI (and other artefacts) for specific purposes in a human world. We then use a series of thought experiments and observations of trust and trustworthiness, to illustrate the use of the model in taking a functionalist perspective on trust decisions, including with machines. Our hope is that this forms a useful basis upon which to develop intelligent systems in a way that considers how and when people may trust them, and in doing so empowers people to make better trust decisions about AI.

1. Introduction
There is no doubt that Artificial Intelligence (AI) has become, in the recent past, something more than the dream of well-meaning academics. If we follow Boden’s definition of AI as making ‘computers that do the sorts of things that minds can do’ (Boden, 2016, page 1), we must in fact conclude that AI has been around us for some considerable time, notwithstanding the fact that the explosion of data, processing capabilities, and business needs have led to a new generation of decision making and problem solving capabilities by machines. Examples of how artificial decision making affects humans abound, from autonomous vehicles in the road and in the air, to autonomous drones as killing machines, and from humble Chess and Go champions to algorithms to determine things like recidivism or the provision of subsidised housing to people. These tools exist, now, and we humans live with them. But what exactly does that mean? What is our relationship to, for instance, our car, when it decides to follow lanes or avoid pedestrians in the road?

These basic questions, and more complex ones, concern the fundamental relationship that humans have with technology in their midst. In that sense they are not remotely new: humans and machines have co-existed for centuries, or more, depending on how you might define technology. On a more fundamental level, humans have also co-existed with dogs, cats, horses, and any number of other domesticated animals for millenia. What, if anything, might we learn from this?

We think that we can learn that AI is nothing special in the grand scheme of human socio-technical societies. At least, as far as trust might be concerned. Drawing on Dennett’s intentional stance (Dennett, 1971), Sloman’s anti-reductive virtual machine functionalism (Sloman, 1996, 2013; Sloman & Chrisley, 2003), and Reeves and Nass’s Media Equation (Reeves & Nass, 1996), this paper examines, from basic principles, what we might mean by this, how humans might think about technology, and why some basic principles that already exist might still hold when the technology we are interacting with is more autonomous and potentially less understood and less under our control than extant technologies.

2. Trust, trustworthiness, and artificial intelligence

It behooves us in the beginning to think carefully about what we mean by the words we use. Definitional problems aside, the fact that trust is such an overloaded term in the first place (McKnight & Chervany, 1996) is only exacerbated by the way in which it is regularly abused in speech and print. To make matters worse, trust is
often confused or conflated with confidence, predictability, or cooperation (Mayer et al., 1995). In the realm of trust and artificial intelligence, a variety of perspectives can be discerned. Let us consider, for example:

- There is ongoing discussion in different fora around making people trust AI more.
- A question posed on Twitter by Accenture in 2019 was ‘Do you trust AI?’
- The EU has recently released ethics guidelines on how to build ‘Trustworthy AI’ (European Commission, 2019), and elsewhere ethics-based governance and standards initiatives have been argued as essential to ‘building trust’ in AI (Winfield & Jirokta, 2018).
- Some (e.g. Bryson (2018) and Ryan (2020)) have argued that no one should trust AI, and even that to do so is a category error. Others think it depends on the type of AI, for example Sutrop (2019) admits trust as a valid notion where the AI is ‘capable of deciding the best action to achieve a complex goal’.
- Many online services provide ratings or seals of approval on things as disparate as professors and chairs(’), but also including online services that include AI. The implication here is: ‘this one has a good rating, you can trust it/her/him/them.’

To ground the discussion, consider further: trust is the act of choosing to put oneself into a situation of risk, where the outcomes are dependent on the actions of another (cf. Gambetta (1990), Luhmann (1979, 1984), Marsh (1994) and McKnight and Chervany (1996)). It is an action taken by a person as the result of a decision making process that involved multiple pieces of evidence. For the sake of clarity and disambiguation in this paper, we call this the decision perspective on trust. Accordingly, while one may certainly be forced into the situation of depending on someone or something outside of our control, it is not possible to ‘make’ humans decide to trust something, except by coercion or deception, let alone ‘more’. Approaching this ethically, the best we can do is to give the human reasons to trust. And sometimes, to expose why they might not want to. On the other hand, unethical behaviour might involve deceiving or otherwise convincing people into deciding to depend on something, against their better judgement.

Furthermore, consider the question: what might I trust you for? Whilst it is definitely possible to have some foundational level of trust in specific others, trust is inherently contextual. I may trust my brother to drive me to the airport, but I wouldn’t trust him to fly the plane (he’s not a pilot). Asking if you trust AI is akin to asking if you trust the human race: perhaps a nice headline but in any event, meaningless. And, even if we answer ‘no’, what does that actually mean, given the situation we are already in?

It then starts to become apparent that what many are presenting becomes less of an ‘ask’ and more of a ‘tell’: much talk of ‘trusting AI’ is in fact trying to convince people to trust something (in some context). Dywer (2011) refers to this as Trust Empowerment. Trust Empowerment tries to make you trust based on the data it presents (and thus can incentivise the projection of a particular image), whereas Trust Empowerment provides the data you might need to make a trust decision. The difference is a powerful one given how personal and contextual trust is.

There is a great deal of discussion around making AI more trustworthy, perhaps through introducing ethical guidelines into how AI is developed and/or used (e.g. those from the EU European Commission, 2019). This work is particularly important, not necessarily because it will result in more trustworthy AI systems (it may) but because it raises the question of what we might mean by trustworthiness and what we might mean by things like responsibility and associated things like justice, when AI systems err (they have and they will again). AI can be made more trustworthy, this is not in doubt, but as we note above, this does not compel anyone to actually trust it in any way, for any thing. The question runs deeper, then: why would you want to trust it and what for?

One primary input into a trust decision is indeed trustworthiness (Mayer et al., 1995; McKnight & Chervany, 1996). Trustworthiness is an attribute, something that we (as humans or animals or AI systems) have, and it guides whether others choose to trust us, and what happens afterwards. It is based on any number of traits, beliefs, desires, intentions, competencies, and so on, and it is not something that we can easily prove to others. One option for others to obtain evidence of our trustworthiness can come through observing the way we behave in some context – we shall call this black-box evidence of trustworthiness – and in some cases, this information may be second-hand, for example through things like reputation systems. Another option is by providing insight into how we work, think, and value, perhaps through transparency or explanations; we shall call this white-box evidence of trustworthiness. Potential trustors use this evidence to form their own subjective beliefs about our trustworthiness as a potential trustee; these are typically called trusting beliefs (McKnight & Chervany, 2000).

The various thrusts to bring around ‘trustworthy’ AI also draw heavily on the idea of trustworthiness by proxy, where we might decide to trust an artefact because it is built by people whom we trust; trustworthy AI might be seen to arise when built by people who are similarly trustworthy (we accept that untrustworthy people may build trustworthy systems, and vice versa, but this is what codes of ethics are for).

However, regardless of an individual’s trustworthiness, another may choose to trust them in a specific context anyway. To put it another way: people don’t (always) trust something or someone because they believe them to be trustworthy, sometimes they do it because their disposition is such that they just want to, or because they consider the alternative even worse (and for many other reasons, besides, cf. Deutsch, 1962). The relationship between trust and trustworthiness is not necessarily straightforward.

Finally, there is the problem of responsibility. As we noted above, untrustworthy people may build trustworthy systems, and indeed there are combinations of trustworthy/untrustworthy here that we can play with. It is also worth noting that since the systems being built are often both modular and far too complex for any single human to be able to create and/or grasp the intricacies of each part themselves, complex software is built by teams, and often even multiple teams who do not work together in a traditional sense.

We thus arrive at an interesting situation: there is no doubt that when people talk of trusting AI or making AI more trustworthy they are sincere, but we are not convinced that they are asking the right questions, or reaching the right conclusions. Similarly, arguments that people ought not trust AI typically come from a deep concern for the potential for deception by machines (and the companies that build them), the information and power asymmetries that drive and may be fed by this, and the resulting human and social cost. But rather than driving us to avoid questions of trust and AI, we believe that such a concern only makes engagement with these fundamental questions more important, and the sort of analysis presented in this paper more valuable.

3. Perspectives on trust

A literature exploring whether people should or even can trust AI has sprung up. Bryson’s arguments, in particular, are some of the strongest in favour of a non-role for trust in the interactions between humans and so-called intelligent machines (Bryson, 2018). These arguments are powerful in the face of opposing efforts by some technology companies and labs to ‘get people to trust AI’. Ryan (2020) comes to similar conclusions, arguing that ‘even complex machines such as AI should not be viewed as trustworthy’ (emphasis ours). Essentially, these arguments rely on trust being a relationship between peers, where the depended-upon peer can bear responsibility. Therefore, if AI is considered (i) to be merely a designed and manufactured set of processes, models, and tools, and is therefore not a peer, and (ii) can bear no responsibility; then we are outside of the domain of trust altogether.
3.1. The accountability perspective

We believe Bryson is right to attempt to focus a debate on the trustworthiness of AI companies, regulators, and others who use AI on their behalf in society, rather than simply the machines themselves. Her conclusion, that we ought to engineer AI for accountability, is undeniable: as with any other manufactured product, when AI causes damage, ‘we need to know we can hold the human beings behind that system to account’ (Bryson, 2018). To do otherwise, would be to attempt to hold machines accountable (incorrectly, she argues) for their actions, instead of those humans who ought to be. Further, this has the potential to enable malicious actors to evade responsibility for the systems they build, on the (perhaps spurious) grounds of autonomy or even consciousness. For the avoidance of doubt: we fully agree that humans should be held responsible (and accountable) when a manufactured product such as today’s AI systems does something ‘wrong’. This is not in question.

However, extrapolating from the position that a) we ought not deflect accountability away from manufacturers into their products since those products themselves cannot be held accountable to b) trust is invalid when considering non-humans, is a leap we find problematic. Contrasting this with the decision perspective, we might term the presence of this additional requirement on trustees the accountability perspective on trust: it requires that a trustee be morally able to accept blame in order to consider trust a valid possibility. We note that this requires a different theoretical underpinning than that of the more widely accepted one of Gambetta (1990), Luhmann (1990), Marsh (1994), McKnight and Chervany (1996) and others.

While we are convinced by the sincerity and social importance of Bryson’s arguments, we do not believe that they preclude the discussions in this paper, for three reasons. First, even if we tell people that they ought not or cannot trust AI, many will decide to go ahead and do just that anyway, making the argument rather moot. Secondly, trust is about much more than just attribution of responsibility and blame: it is an essential mechanism for the coordination of groups. And third, will it always be the case that the people behind a particular AI system will be able to assume responsibility for its actions? The answer is only yes if we restrict the incorporation of yet-to-be-invented intelligent systems over a certain complexity level into society. It could also be argued that the threshold in this third point has already been surpassed.

Thus, the argument that we ought not construct a mask of anthropomorphism, and in doing so allow those in positions of responsibility to evade that responsibility, is a strong one with which we take no issue whatsoever. However this is not the point of this article. Instead, people will use trust to reason about machines as they choose and have long chosen to trust animals, children, machines real and imagined, perhaps in limited and specific ways given their nature, for millennia. We argue that we ought to understand this phenomenon, and acknowledge it in how machines are designed. At least, we would hope that it is uncontroversial to state that machines should not be designed to take advantage of the trust heuristic, perhaps deceiving the user in order to gain from them unethically. Ideally, however, we would also build machines that empower people through their own mental models and decision-making heuristics (such as trust), so that we help them reason about complex machines.

3.2. The coordination perspective

Another argument made for a non-role for trust with intelligent machines is the proposal that trust (only) exists in a relationship between peers. Bryson’s definition of trust, for example, is “a relationship between peers in which the trusting party, while not knowing for certain what the trusted party will do, believes any promises being made” (Bryson, 2018). This is a carefully chosen and worded definition of trust that suggests that it only operates amongst peers, and only on promises. A form of trust based on promises is clearly a relationship between peers, peers of the sort that at least both have the cognitive and linguistic ability to participate in and understand the above expectation reconciliation process (Baier, 1986) (and possibly the ability to reason at the metalevel about the cognitive capabilities about a particular peer, much as we do in this article). Similarly, we can consider what Sako and Helper (1998) call ‘contractual trust’, trust based on the expectation that parties keep to an agreement negotiated between them, or perhaps a documented ethical or behavioural standard, and ‘transactional trust’ (Reina & Reina, 1999) where amongst other things, trust is given in order to build or maintain trustworthiness.

Thus, we can see a coordination perspective, where trust is seen as a coordination of expectations, requiring communication to establish these expectations, and hence the trustee is by definition aware they are being trusted, and is sufficiently peer-like to be able to do this.

According to the broader trust literature, and considering all the ways that humans decide to trust, we find this is overly narrow and restrictive as a general definition of trust. There is little doubt that humans are not the peers of their animal companions, and yet trust (in the sense described by Gambetta (1990), Luhmann (1979), and others) exists between human and, for instance, a service dog — even to the extent that the dog is trusted by the human not to lead them onto a road where there is an oncoming car. Moreover, the relationship is notably bidirectional. But note here that no promises were ever made by either party; and if there were, such promises would not be understood in any case, so the point of whether any were made at all is moot. The definition is not just problematic because it glosses over and cherry-picks the situations it wants. It is problematic because it implies human exceptionalism: it refuses to accept the possibility that non-human trusters and trustees could exist, and this is incompatible with a functionalist view on intelligence and agency.

Certainly, the sort of trust considered by the coordination perspective has the potential to lead to deeper trusting relationships. And while readers will have their own views about the above definitional discussion, there is no need to discount the heuristic notion of trust decisions, which brings with it a lower barrier to entry. Many people make trust decisions concerning various specific others, without those others knowing or perhaps fully understanding, daily.

3.3. The goodwill perspective

A further perspective identifiable in the literature (e.g. Baier (1986)) is what Sako and Helper (1998) call ‘goodwill trust’: that trust begins when the reliance depends on the goodwill of another. This view is carried through Lagerspetz (1998) into the recent work of Sutrop (2019) on trust an AI. One idea here is that we can usually tell if we were trusting someone or ‘merely depending on them’, based on our reaction when we are let down by them: respectively, do we feel betrayed, or are we just disappointed? According to this perspective, inanimate objects without the capacity to decide to act positively in the trustor’s interest (or not), also lack the capacity to betray. Thus, reliance based on competence alone can lead us only to a sense of disappointment. When instead a trustee reaches the level of accepting a dependency on the trustee’s willingness to do what was stated or further, what was agreed, we can see why there can be a feeling of betrayal when the trustee does otherwise — although, consider that even in this case, there are almost always new goals that can arrive, and in doing so, can trump what was agreed. It is even acceptable to act on these without notifying the trustee in some cases, although this may not stop the feeling of betrayal until and unless the trustee finds out what happened.

Sutrop (2019) argues that the feeling of disappointment rather than betrayal is common with machines, because regardless of the machine’s capabilities, we can always tell ourselves that it was ‘only a machine’ (cf. ‘only an animal’, ‘only a child’). But even for her this does not mean that there is no place for trust between humans and machines. By placing the act of communication between the trustee and trustor in establishing expectations at the core of the trust decision, Sutrop
concludes that this ‘goodwill’ trust can be appropriate in machines, given the machine has a certain level of cognitive capability. And in many cases, we may naturally feel as if our partner, horse, or autonomous car is capable of ‘trying to meet our needs’ through its actions, and approach it with the attitude of trust, as a result. As Dennett (1971) argues, this intentional stance can be a useful way to reason about complex objects, since it allows us to re-use exiting cognitive scaffolding (we suggest, such as trust). Notwithstanding the uneasiness of some to adopt this mindset, it is now commonplace in research, engineering, and use of artificial agents, and many people reason in terms of the intentionality of (intelligent) machines.

3.4. The decision perspective

Recall from Section 2 that the decision perspective considers trust a process of intention formation, based on evidence. As Luhmann (1979) notes, trust is a mental shortcut: a heuristic we use to reduce complexity in decision making, and one that is particularly useful in unanticipated circumstances (Sako & Helper, 1998). It does not require the trustee to be aware of the decision, or even that a decision is being made. It is also prevalent in human decision making, where there is a sense of agency in the other, that it will naturally find its way into our interactions with machines we perceive to be agent-like. It is, we argue in this paper, a natural extension of the intentional stance to use this decision heuristic with complex artefacts such as machines.

Indeed, we see little point in trying to support an argument in favour of excluding certain parts of this notion from ‘what trust really is’. What is more important is to understand how these things, all commonly referred to as trust, relate to each other, what they mean, and what that means for AI and people.

Generally then, we proceed with the most widely accepted and inclusive perspective found in the literature: to consider that trust can be seen simply as the decision to accept a dependency on some other, in the face of risk, and with expectation of positive outcomes. This expectation of positive outcomes may come from an belief of positive intention arrived at through communication, or through a belief in the presence of moral responsibility in the other, or through a belief in the goodwill of the trustee, or simply in a belief in its ability and likelihood to perform a certain action. But these are sub-cases, neither of which alone is sufficient to capture all the ways that humans do trust.

If we were to instead restrict the permitted discussion of trust to only those cases where there is deemed a sufficient “peer-like” level of cognitive or moral capacity in the trustee, we are faced with the uncomfortable question of: deemed by whom? And what if the person doing the trusting deems something else? With a focus on trust empowerment through education and transparency, but not judgement of others who choose to trust in their own personal circumstances, we are also unconvinced by the validity of arguments framed in terms of one person’s view on whether another ought or ought not trust. If we were to restrict the literature that follows in ways that only considered a subset of the issues explored in this section, we will have failed all those who simply go ahead and trust anyway.

It is also worth mentioning that sometimes mental models of a process or action are quite valuable, even if they abstract away from the actual workings of something. As an example, consider the car. One depresses the ‘gas’ pedal and it goes (or increases in speed). The model is quite simple and has the added advantage of removing the need to understand Venturi effects or fuel injection systems. Moreover, it abstracts even further when one sits in an electric car that uses no gas at all — press the pedal to go. The actual workings are sometimes complex and often of no interest to the driver. The same might be said of complex systems where trust is employed as a means to determine if the system is working properly or not. If there is consistent, ‘correct’ behaviour, trust is indeed merited (as part of an abstract model of something working). If the behaviour becomes erratic, it seems fair to trust it less. There is little to no anthropomorphism here, and we would venture to suggest no real harm in the model. Whilst it is possible to educate the human as to the reason why things happen, or how things are done, it’s quite reasonable to expect that many people simply won’t care as long as it works.

There are of course examples where this might go wrong and the person might feel betrayed in some way, but we would venture to suggest that this is neither unique to computational systems nor indeed to advanced technology in general.

4. A model: The human thinking about the AI

Having dispatched with some of the philosophical issues around the rationale for this approach, we now turn to the mechanics of the trust heuristic. In this article we are interested in how humans make trust decisions: how they decide whether to trust a particular X that they cannot fully understand, predict, or control. First then, we turn to literature concerning this.

4.1. How do humans decide whether to trust?

Recall that trust is, essentially, choosing to put yourself in a position where outcomes that matter to you are in the hands of another. It is a response to risk where we do not have control (Cofa, 2007; Deutsch, 1962). To put it another way, trust is a way of intentionally accepting that some things are out of our control. We are unable to control, but we can to a certain extent try to predict (or expect) certain hoped-for outcomes. Trust also allows us to accept and reason about the agency of others in our society (Luhmann, 1979). It allows us to capture salient features not only of the others with whom we are dealing, but of the world around us which, again, we can’t control (Luhmann, 1979). We do this by considering the trustworthiness of the other(s), the environmental situation, the risk associated with alternative outcomes (see for example Boon and Holmes (1991), Govier (1992) and Luhmann (1990)), and our attitude towards it. We accept here that these things are subjective, which is entirely in keeping with the way in which many researchers see trust decisions (Deutsch, 1973; Gambetta, 1990; Marsh, 1994). We also accept that risk estimation is not something at which humans are very good (Zeckhauser & Viscusi, 1990). Trust decisions are also personal: they are guided by situational and personal factors, including our own biases, attitudes, and prior beliefs, as well as the context of the trust decision.

Architecturally, there is strong agreement in the literature on a separation between trustworthiness judgements and trust decisions that can follow from them. This can be seen in the trust model proposed by Mayer, Davis, and Schoorman (Mayer et al., 1995), depicted in Fig. 1. In this model, three primary features are factored into a trustworthiness judgement: the potential trustee’s ability to do what you need them to, their benevolence in deciding whether to do it, and their integrity in terms of honouring any agreement made between you concerning whether they would do it.

A potential trustor makes a trust decision by drawing on this assessment of the trustworthiness of the potential trustee, as well as the risk, and specifically their own propensity to trust. This propensity captures the notion that some individuals are naturally more predisposed to trusting, essentially being more optimistic.

Should the trustor decide to trust the trustee, then they enter into a relationship that contains risk. Ultimately, either the trustee does what the trustor needed, or they fail to do so, leading to an outcome for the trustor. This outcome, whether positive or negative, provides new information about the trustworthiness of the trustee, which can be used for future judgements.

Based on an extensive review of the available literature at the time, McKnight and Chervany (1996) developed a similar model, with the aim of describing how various areas of trust research relate to each other as part of the overall process of making a trust decision. This is shown in Fig. 2. Theirs is compatible with the Mayer–Davis–Schoorman
model, yet captures some additional features, while abstracting away others (but still dealing with them in the text).

Specifically, while the process of generating a subjective belief about the trustworthiness of another is still present (the Belief Formation Processes leading to Trusting Beliefs), the Mayer–Davis–Schoorman model’s features of trustworthiness (ability, benevolence, integrity are omitted), in favour of a broader and perhaps more nuanced discussion in the text. Now here is the notion that a trusting disposition comes not only from the attitude (e.g. optimism) of the potential trustor, but also from the situation. They note that, in general, people tend to trust more in some situations than others, despite equivalent risk and trustworthiness of the other. Finally, we also see that trust decisions factor in System Trust. This refers to the broader (i.e. impersonal) social context in which trust decisions are made. For example, some societies tend to foster trust more than others, and the existence of institutions and social norms may lead to a greater or lesser propensity to trust (Luhmann, 1990). System Trust captures the extent to which one believes that proper impersonal structures are in place to enable one to anticipate a successful future endeavour. Given the nature of the McKnight–Chervany model as a map of relationships between concepts, rather than a schematic of a decision-making process, other features, such as the feedback loop and risk information, were omitted. It also provides a means for the members of the system to rely on it to enforce required actions on pain of, perhaps, punishment (see for instance Etalle et al. (2007)).

Much has been written about how humans make trust decisions in specific contexts. System or institutional trust is of great interest in sociology, economics, and business. To cite two very different examples, system trust plays a key role in business-to-consumer commerce, and can be fostered through specific action by vendors (Pennington et al., 2003); somewhat differently, trust (or lack thereof) of institutions affects the wellbeing of individuals associated with those institutions (Hudson, 2006), as well as having a causal effect on the levels of system trust in society at large (Sønderskov & Dinesen, 2016). It is beyond the scope of this article to survey all forms of trust captured in these models. Reviews for many of the specific features already exist, e.g. Goto (1996) and Helm (2004) provide insight into dispositional trust, while Jones (2012) explores trustworthiness more deeply.

While predictability is not the same as trust or trustworthiness (Mayer et al., 1995), most models that describe how humans arrive at their beliefs about the trustworthiness of others do factor in predictability as a contributing feature (McKnight & Chervany, 1995). Thus, trustworthiness is often extended beyond the three features of the Mayer–Davis–Schoorman model, with a fourth, predictability.

4.2. A general model for trust decisions

As a synthesis of the above concepts and relationships, we now propose a general model of trust decisions. Our model describes a subjective, situational trustworthiness judgement based on four primary trustworthiness features, plus proxy trust. Subsequently, trust decisions are based on this trustworthiness judgement, plus risk and ones own attitude to it, and dispositional factors, which may be situational, personal, and systemic. The model is general because it applies not only to trust between peers, but the use of the trust decision heuristic in any situation. The model is depicted in Fig. 3.

To use the model, we position ourselves in the role of the potential trustor, and then attempt to ascertain features of trustworthiness and proxy trust (i.e. consider each of the boxes in the left-most column). To do this, we use available information about whatever we are assessing, in order to arrive at a trustworthiness judgement. It should be clear that our trustworthiness judgements are therefore highly dependent on the amount and sorts of information that are available, perhaps provided, visible, or able to be ascertained by examination or interrogation. We have seen how trustworthiness judgements are typically based on the trustor’s beliefs about four aspects of the trustee: competence, predictability, willingness, and integrity. These are what McKnight and Chervany call trusting beliefs (McKnight & Chervany, 2000). Let us explore each of these boxes, and the information they draw on, in turn.
4.2.1. Competence

Competence is concerned with whether the trustee can in fact do the task asked of them, in their situation. In assessing competence, the trustor looks for evidence of such an ability. Recalling our notions of black-box evidence of trustworthiness and white-box evidence of trustworthiness from Section 2, evidence of competence comes in both black-box and white-box forms. In the black-box sense, we can ask: has the trustee previously demonstrated the ability? Further, has anything changed since then (e.g. in the environment, or the degradation of a faculty, body part, or component)? This can be complemented with white-box evidence. For instance, when considering another human we might ask: do they appear physically and mentally capable of performing the task, given the trustee has some understanding of what the task involves? Do they have the required level of education, training, skill, or knowledge? These can be probed using questioning, or by deferring to certifications. For non-humans, are its machinery and mechanisms for carrying out the task sufficient and understood? To what extent have they been tested or verified, either in constituent parts, or as a whole?

4.2.2. Predictability

Predictability is concerned with statistical descriptions of the potential trustee’s behaviour, decisions, or outputs. As McKnight and Chervany put it, “predictability means one’s actions are consistent enough that another can forecast what one will do in a given situation”. Essentially, it is concerned with the question: do I have good statistical evidence of what the potential trustee does? This includes reliability of task performance, as well as uncertainty associated with its correctness, bias, and major determining features. We may also draw on social sources of predictability evidence. For example, reputation information provides an aggregation of others’ subjective experiences of the trustee’s behaviour. This aggregation represents a crowd-sourced statistical description, which, while susceptible to hacking through misleading information and other attacks, can be valuable.

4.2.3. Willingness & benevolence

Willingness expresses compliance with a need. As a trustworthiness feature, it is concerned with the alignment of the potential trustee’s motivations (e.g., goals, drives) with the trustor’s needs, or a decision by the potential trustee to create such alignment. Essentially, is the trustee going to take actions in order to attempt to (i.e. is it willing to, set up to, or actively pursuing a course of action to) do what I need it to do? Drawing on the intentional stance (Dennett, 1971), we may also interpret what is captured by human willingness for this purpose in non-humans, including goal-directed machines and other creatures where, although their information processing may be mind-like (or treated as such by the trustor) in some respects, it is less common to consider the creature ‘willing’ to do something. Functionally, in terms of trustworthiness judgements, willingness therefore captures the selection of actions directed for some purpose. In contributing to trustworthiness judgements, the trustor is assessing whether that purpose is aligned with her needs. Willingness in trust is often (e.g. Mayer et al. (1995)) also considered in terms of benevolence: that this is an act the trustee knows is in your interest rather than necessarily their own (maybe even against their own), and therefore their decision is driven by your needs rather than theirs. The trustee asks: is this the case? And do I know when this will change?

4.2.4. Honesty & integrity

Here we are concerned with whether the potential trustee will make agreements, and subsequently adhere to them. Firstly, when assessing this trustworthiness feature, the trustor needs to determine whether indeed the potential trustee has the cognitive capacity to make and honour agreements or promises. Secondly, the trustor asks whether such an agreement is in place for what it is considering relying on the trustee to do. If so, how strong is this agreement? Does the trustor understand when the trustee will break it, when new things come along? As an example, I may trust someone to go to the store to buy some drinks for a party, but I will normally be completely understanding if they fail to honour this agreement, because they needed to urgently attend to an injured spouse instead. Finally, will the trustee deceive the trustor, or withhold the truth in some circumstances? We must also acknowledge that in some cases (e.g. storytelling, for example by a grandparent or by an intelligent game character) this may be desirable (Castelfranchi, 2000).

4.2.5. Proxy trust

Finally, a further concept that emerges in the literature is that of proxy trust. People sometimes refer to levels of trust in artefacts that appear to not warrant it, yet without seeming nonsensical. When assessed in-and-of themself according to the above four features, there is typically little cause for high levels of trustworthiness. But through proxy trust, the object can also act as a facade for other trust relationships. For example, I can find a doctor trustworthy, even though I have never met them before or seen any evidence of their work on other patients, because I know that they were trained at a reputable medical school, and that medical education and practice are heavily regulated, and I have already chosen to trust those individuals responsible for medical education and regulation.

Similar examples in technology abound. For example, in cloud computing, Lansing and Sunyaev (2016) describe the “duality of trust” in both the technological artefact itself as well as the providing organization. They further highlight the importance of social context (specifically, in terms of inter-organizational trust and system trust) in determining the latter, as well as exploring the interactions between the two objects of trust on each other.

Sutrop (2019) argues (and we agree) that we ought at least be clear about the object of our trust: is it the machine itself, or the network of individuals behind it, or both? The notion of proxy trust, supported by this clarity, allows us to take an already made trust decision concerning one or more trustees, and to project it into the trustworthiness judgement of another.

4.2.6. Meta-features, entanglement, and extension

Each of the trustworthiness feature classes can also be considered at the meta-level: how they apply not to the trustee’s behaviour at the task, but to their trustworthiness behaviour itself, as captured in these features.

Consider, an individual may be honest, but may be highly unpredictable in its honesty; an individual may be willing to do the task, but not willing to be predictable in terms of how (perhaps this involves additional effort); an individual may not be competent, and also not honest about its lack of competence (potentially leading to mistrust); an individual may not be honest, but may not have the competence to be fully honest.

These trustworthiness factors are also inherently entangled. For example, a greater capacity for honesty is clearly a competence (more precisely, a meta-competence, in the model). A greater competence in providing evidence of its situational competence (cf. explainable AI)
also gives the trustor white-box information about that situational competence. Similarly, predictability also provides evidence about other features: if the trustor can build a statistical model of the trustee’s behaviour in terms of the other features, this gives her black-box information about those features.

Finally, we do not claim that our model is complete, rather that we hope it will be useful, and potentially extended and further concretised where necessary. While McKnight and Chervany’s surveys (McKnight & Chervany, 1995, 1996) found the four features we use here to be the most often used determinants of trustworthiness, they also found evidence that we draw on many others too. These included openness, openmindedness, the perception of the trustee being careful or safe, whether there was a sense of shared understanding, and even personal attraction.

Further, it is typically useful to concretize trustworthiness features in each specific domain. The literature is replete with domain-specific trust models in a variety of technological contexts, from geospatial data (Lush et al., 2018) to cloud computing (Lansing & Sunyaev, 2016). We highlight these two examples since they demonstrate particularly systematic, stakeholder-driven studies of trustworthiness features in their respective domains.

4.3. A trustworthiness feature space

Each of four feature classes in our model describes a type of information that we use to make assessments of trustworthiness. It is important to consider that not all four are required for an assessment of trustworthiness to lead to a good trust decision. Indeed, in some cases, the absence of one or more can be compensated by strength in others, i.e. their combination is not purely multiplicative. As an example, for a potential trustee we believe to have high integrity and competence, we may be prepared to discount predictability, if an agreement between the trustee and trustor is in place. On the other hand, for a potential trustee displaying high competence, predictability, and willingness, we may be prepared to dispose of any need for it to make an agreement with us to carry out the task. Consider that a potential trustee with low competence at the task, but high levels of honesty, could lead to a valuable trustworthiness judgement by the trustor, contributing to a good trust decision in the situation.

Thus, we must distinguish in each feature class between the amount of evidence we have for each feature (i.e. our confidence in assessing it) and the amount of that feature that we can subsequently attribute to the trustee in our trustworthiness. For example, having little information about the trustee’s competence or willingness is not the same as having information that tells us that the trustee has low levels of these. The model therefore captures how trustworthiness contributes to Marsh and Dibben’s (Marsh & Dibben, 2005) distinction between trust, distrust and mistrust, on the one hand, which are all active trust decisions (positively or negatively) based on sufficient trustworthiness evidence; and untrust on the other hand. Untrust is the situation that arises when the trustor either doesn’t have enough evidence to decide, needs further reassurance to help reduce risk (for instance the signing of a contract) or has evidence that the other agent is not competent in this situation.

To be fully complete, our feature space should also capture meta-predictability, meta-competence, meta-honesty, and meta-willingness as applied to the other features. We leave this analysis for later consideration and exploration; at this stage our aim is to illustrate how the model’s feature space already captures a broad range of trust decisions.

5. So, what is it like to trust a rock?

Just as humans and other animals are mostly made of carbon, hauling water around the Earth, computers are, for the most part, made from rocks.

For sure, these rocks have been reprocessed and arranged in a particular, complex, and purposeful fashion, and as those who carried out this arrangement, we humans are often predisposed to seeing them in terms of the set of components they are, and the processes used to assemble them. Yet as complexity increases, this perspective often shifts to be more whole-system oriented; indeed viewing computing machines as entities at varying levels of abstraction is a key part of modern computer science, software engineering, and human–computer interaction. One very useful example of this is the notion of computational agents, such as robots and software agents. This perspective is sometimes even necessarily and automatically taken by the majority of people who are not well-versed in the building or analysis of complex computing machinery. We have already discussed the prevalence of the intentional stance Dennett (1971), above.

Notwithstanding the efforts of psychology to understand the “software” of humans, neuroscience to understand our “hardware”, and biology to understand our “mechanics”, day-to-day as humans we reason about each other (and animals) as atomic entities: Jill, Dad, Dr Chauhan, Bill the pony, etc. When we evaluate the trustworthiness of one of these, we do not usually consider their constituent parts. Instead, we lean on our understanding of that individual as a whole: how we have seen them behave, what we believe them to be capable of, our understanding of how they think and make decisions, and (using empathy) what it might be like to be them.

On the last of these points, we cannot escape but consider what Dennett described as “the most widely cited and influential thought experiment about consciousness” (Dennett, 2011), in which Nagel (1974) teases us to consider: what is it like to be a bat? Nagel’s argument is that the subjective nature of consciousness precludes its explanation in reductionist terms, i.e. by resolving the phenomenon only to physical factors that do not take a subjective view. The use of a bat as the primary subject of the thought experiment is intended to simultaneously bring into focus the notions that a) as mammals, they would ordinarily be considered by us to have some form of consciousness, and b) their experience is unfathomable to us, as it is gained from very different perception capabilities (e.g. echolocation) and physical abilities (e.g. flying). Thus, while we might be able to imagine what it is like to be a bat, we can never truly assume the mindset of one.

Nagel’s arguments sparked decades of debate around consciousness, which we do not cover here, since consciousness is not our question. However, in one rather tongue-in-cheek response, Sloman (1996) proposes to replace what he (rather unapologetically) described as “deep-sounding unanswerable, time-wasting pseudo-questions” (i.e. but what is it really like to be a bat, from the inside?), instead with “deep, factual, discovery-driving questions about the nature and content of internal states of more or less intelligent entities of various kinds”. His approach, to explore the information processing mechanisms, cognitive states, and semantic spaces present in (very broadly speaking) agents of various sorts, is illustrated by a series of questions of the form “what is it like to be an X?”

In doing so he extends the thought experiment to consider, from a functionalist perspective, what it is like to be an infant, someone with autism, a sunflower, a robot, a bat, and ultimately, a rock. This method is designed to help uncover, as he describes it, “some of the internal ontology, the virtual machine requirements, the abstract data- and operations on them, which we need to think about if we are to design human-like systems” (Sloman, 1996). When considering white-box trustworthiness at least, this is indeed what we are trying to do. Sloman’s approach is a powerful one. It admits no ‘special status’ for one type of intelligent agent over any other, but clearly acknowledges the variety that exists in forms of cognition and intelligent behaviour between humans, animals, and machines of different sorts. Further, since we can now consider that the semantic spaces between these forms will (in general) not overlap, the approach also provides us with a concrete way of considering the implications of these forms on the way others can perceive them. And this is our question.

As an example, let us take willingness (one of our trustworthiness features) into consideration: must a potential trustee really be
“willing”? And what does this mean? Or, by taking a functionalist perspective, can we admit that a range of different possible mechanisms that provide for willingness-like features, of different degrees, contribute to judgements of trustworthiness that people make?

We argue that they can, at least insofar as it is useful in reasoning about the way humans make trust decisions concerning a range of things that may be taken as expressing something functionally akin to willingness, in one way or another. If we accept this, and the generalisation of this example, then the remainder of our thought experiment becomes admissible — and, we hope, useful as an illustration to willingness, in one way or another. If we accept this, and the things that may be taken as expressing something functionally akin about the way humans make trust decisions concerning a range of contribute to judgements of trustworthiness that people make?

5.1. Straightforward rocks

So what is it like to trust a rock? For a start, the answer is perhaps “not very interesting or very risky”, but as with everything, it depends on who you are and what you are considering trusting it for. The behaviour and information processing capabilities of rocks, limited as they are, are well understood, and highly predictable. For example, a rock has the capability to retain its shape under stress, but under sufficient force it will crack or shatter. Depending on the rock, when water or wind consistently interacts with it, the rock may become deformed; in some sense, it therefore has the capability of preserving historical information about its environment. At high temperatures, a rock will melt. When a rock is thrown in the air, unless interrupted, it will fall to Earth.

A rock has nothing at all like what we would describe as agency or autonomy. And thus it has no capacity to be functionally willing or able to represent, communicate, or agree intentions, or to pursue them (i.e. honesty is also not in the semantic space of a rock). Most people understand this sufficiently well to find judgements of trustworthiness based on either of these trustworthiness features rather absurd.

A trustworthiness judgement concerning a rock depends on the trustor’s understanding and mental model of these things. So, for the majority of people, it is like considering rocks highly trustworthy in terms of ones belief in what the rock is competent to do, and the predictability of its behaviour. Here, we are so rarely surprised (black-box) and have enough understanding of how its internals operate (white-box), to be able to make good trustworthiness judgements.

Thus, to a large degree, people have sufficient understanding and control to simply depend on rocks based on their understanding of the rock’s (limited) competence and (high) predictability alone, and relax. After all, many people rely on the property of rocks to remain where they were placed (within some tolerance) under reasonable load, when building a house. Some rely on small crystalline rocks to keep them on time for social engagements.

Proxy trust also has a role here: in the case of the house, the homebuyer may not have sufficient understanding of the chemical make-up and density of house-bricks, but they do trust a surveyor who does. Similarly, most people do not understand the physical and information processing nature of quartz, yet trust watch manufacturers to have a sufficient understanding to create something that will meet their needs. In some ways, this is similar to the proxy trust present in software development, except that the complexity is substantially lower.

Such a high-quality mental model on the part of the trustor may not always be taken for granted, however. What it is like to trust a rock also depends on who the trustor is. Consider how trustworthiness judgements might be made in the absence of a lifetime of black-box examples or access to others through whom to proxy trust. Imagine, for example, how a prehistoric human might reason about being on the 14th floor of a brick-built building. Or consider the wonder of a child discovering the piezoelectric effect in a science class, without yet having sufficient understanding to be able to explain it; would they trust that it would always behave that way, in order to not be late for school? Perhaps the surprising behaviour they discovered might seem rather uncanny, perhaps the result of an unknown force, and the addition of this new uncertainty make them find the rock seem less trustworthy.

But most rocks are simple enough for most people to have functionally useful mental models of them. With good information concerning the first two features in our trustworthiness model (competence and predictability), and also high confidence in the lack of presence of the third or fourth dimensions (willingness or honesty), we are in the domain of what Sako and Helper (1998) call ‘competency trust’. Some (e.g. Avižienis et al., 2004) consider this be outside of the domain of trust altogether; requiring trust decisions to be made also on the basis of an expectation of ‘good will’ or at least (an interpretation of) willingness of the trustee (see Section 3.3 above). With a zero level for willingness, this competency trust is sometimes instead referred to as only an accepted dependency (Avižienis et al., 2004).

Our view is that this is largely a terminology issue: competency trust and accepted dependency, where a decision is made based only on evidence of competency, amount to the same thing. The ‘trustor’ (for want of a more general word) is still making a decision in the face of risk, based on an expectation of a positive outcome dependent on the other; only here the data they are able to draw on, and possibly the ability of other to choose to do otherwise, are very limited. In terms of our model, judgements of this type represent one edge of the trustworthiness feature space: perhaps it can be considered a zeroth level of trustworthiness.

5.2. More complex rocks

Now, suppose we pass a few of these rocks through an electric arc furnace, cast them most carefully into a particular form, to produce a silicon chip. Of course, what becomes interesting here (and perhaps always was) is not what the thing is made of, or where it came from, but what informational things flow through and are transformed by it, and what causal relationships this information has with the object and its world. In short, we are interested in its software.

Verification and validation of computer programs are common in software engineering. Extensive, often automated software testing allows us to have confidence that the software does what we want it to do, in a broad range of situations. Formal verification goes a step further and allows us to prove the correctness of a piece of computer software or hardware, guaranteeing that it will behave as expected (according to its specification) for all possible inputs. For example, a simple piece of software that takes a number as input and adds one to it, can be proved correct over the set of inputs that your chosen computer can represent as numbers (i.e. based on however many bits it uses). In cases of this sort, we often talk of there being no need to trust the machine to behave as expected, because its competence and predictability are fully accounted for (i.e. there is no risk, leading to the ‘Perceived Risk & Attitude Towards It’ box being zeroed out, essentially collapsing the model). Again, we find ourselves on the edge of the trustworthiness feature space established by our model.

Note however that this is still only a guarantee of expected behaviour to people who fully understand the behaviour specification. Any misalignment between a user’s own expectations and the expectations captured in the specification may still lead to user surprise. Thus, either sufficient expertise and time on the part of a user and/or proxy trust via others, come into play. The software and hardware industries make extensive use of proxy trust in cases such as this, and more complex generalisations. For example, IEEE Standard 1012 (IEEE 1012, 2016) is concerned with verification and validation of software and hardware for specific purposes. Badging a piece of software as certified (e.g. by the IEEE) may serve as a trust trigger for users, similarly to how badging is used in (for example) e-commerce (Lumsden, 2009) and assessing data provenance (Lush et al., 2018). These support users in deciding whether to depend on a piece of technology, when what is ‘behind the scenes’ is not accessible.
Formally proving the properties of most realistic software is prohibitively costly. For this reason, typically full formal verification is only even considered in safety-critical systems or those with high security requirements, where the risk profile warrants it. Aircraft control systems are an obvious example (Bochot et al., 2009). Other experimental examples include CompCert, a verified C compiler, and seL4, a verified microkernel. There is no doubt that the scope of software able to be verified has dramatically increased over the years, and continues to do so, through techniques such as abstraction (Clarke et al., 1994) and probabilistic model checking (Kwiatkowska et al., 2011). Indeed, modern verification can provide strong guarantees about systems even with distributed autonomy (Chen et al., 2013) and those that can adapt their own software architecture (Cámara et al., 2015).

But it is probably safe to say that most software that most people use for most purposes has not been formally verified, and further, that testing is at ‘best effort’. Further, at least for economic reasons, this is likely to remain the case for the foreseeable future. As and alluded to in Section 3 above, even if there were to be a collective global decision to require all complex technological systems to be verified before use, this would come at the cost of at least restricting the release of technology until both the technological ability and economic case for verifying each new intelligent system existed; the corollary of this is that some yet-to-be-conceived intelligent systems that are currently more familiar in science fiction than science fact, would never be permitted. Perhaps the existential threat posed by the use and incorporation of AI into society really is sufficient to warrant such a global treaty, but in the meantime, we note that the pervasive nature of computing-based technologies is only growing, and impacting the lives of ever more people, most of whom are non-experts.

There is no doubt that guarantees, properly understood, can empower people to depend on software with less need for trust. Indeed, if it is possible for a user to fully understand, control, or predict what piece of technology will do, for example by conducting a formal verification analysis or trusting someone else who has, then there is no need for direct trust decisions concerning the machine. This is somewhat akin to Barber’s considerations of fiduciary trust (Barber, 1985): one trusts that a professional will have one’s best interest at heart (here, the professional is the coder, or the verifier). But this might not be the case for any given machine-user pair: as Clarke’s third law notes, any sufficiently advanced technology is indistinguishable from magic, and a (perhaps prehistoric, uncontacted, or simply infant) human without such a mental model or anyone to proxy trust through, would not have the basis upon which to arrive at such a trust-free dependence. In the absence of understanding, the complexity-reducing (Luhmann, 1979) trust heuristic reigns.

We should also take a few words to mention blockchain, often referred to as a ‘trustless’ system. Blockchain is a now almost ubiquitous technology that it is argued potentially removes the need for trust, at least in third parties. For example, with a currency based on blockchain, there is no longer any need to trust a government, or a bank. But blockchain does not do away with the need for trust: if we decide to use it, we are still deciding to trust the developers who implemented the protocol, those who create the smart contracts that may run on it, and perhaps the organisation that holds our wallet (noting that at least one major Bitcoin wallet holder has collapsed in circumstances where the people behind it were found to be less than trustworthy).

5.3. What it is like

Our rock example allows us to explore what happens at and near the limits of the need for trust, and to illustrate how the generality and inclusiveness of the space defined by our model permits this analysis. At the limit, full knowledge and confidence in our judgement of the trustworthiness features as they apply to a rock would amount to there being no need for a trust decision. There is no need for most people to use the trust heuristic with the rock, but this is only since its workings are so simple as to be fully accountable for, and there is no risk. Indeed, trusting a rock is like feeling confident, it is like feeling in control, it is like feeling that you understand.

As the complexity of our ‘rocks’ increases, understanding becomes harder. As Avižienis et al. (2004) note, technical systems can, and usually do, fail. Moving away from trivial full-knowledge edge cases, trust is always relevant. Trusting a complex technical system is therefore like making extensive use of proxy trust, trusting others who have the expertise, tools, and time to develop an understanding beyond our own. It is like looking for evidence of verification and validation, to provide insight into the technology’s competence and predictability. It is usually like finding this incomplete, leaving us with uncertainty around these features. It is sometimes like having a sense that the technology is ‘trying’ to achieve something (i.e. has a goal), or may have an ability to make an agreement in some restricted formal sense, in the case of agent-like systems. It is like feeling that these things may help us to understand when the system will behave in one way rather than another. And as with all trust decisions, it is like weighing up all these factors in the context of what we are considering trusting the system for, alongside the severity of the risk if it were to not to do this.

Even in the presence of this uncertainty, risk, and partial understanding, we very often choose not to stay clear of these artefacts. We often accept the risk and engage anyway, armed with the best information we can get about it. These are personal trust decisions, and to provide people with evidence against these features above is trust empowerment. It would seem to us that there is a moral imperative on those who build and release such systems into society to provide such evidence.

6. Trust and animals

Now we are familiar with the model and its use, from the previous two sections respectively, we continue, first by exploring its applicability with animals, and then in the following section, with machines of various kinds.

6.1. What it is like to trust a dog?

Both the authors of this paper, as is the case with many humans both now and throughout history, live with animals. As he was writing his paper, one of Steve’s cats disturbed him, sitting at the back door and meowing. On getting up and opening the door, Steve was led through the limits of the need for trust, and to illustrate how the generality and inclusiveness of the space defined by our model permits this analysis. At the limit, full knowledge and confidence in our judgement of the trustworthiness features as they apply to a rock would amount to there being no need for a trust decision. There is no need for most people to use the trust heuristic with the rock, but this is only since its workings are so simple as to be fully accountable for, and there is no risk. Indeed, trusting a rock is like feeling confident, it is like feeling in control, it is like feeling that you understand.

As the complexity of our ‘rocks’ increases, understanding becomes harder. As Avižienis et al. (2004) note, technical systems can, and usually do, fail. Moving away from trivial full-knowledge edge cases, trust is always relevant. Trusting a complex technical system is therefore like making extensive use of proxy trust, trusting others who have the expertise, tools, and time to develop an understanding beyond our own. It is like looking for evidence of verification and validation, to provide insight into the technology’s competence and predictability. It is usually like finding this incomplete, leaving us with uncertainty around these features. It is sometimes like having a sense that the technology is ‘trying’ to achieve something (i.e. has a goal), or may have an ability to make an agreement in some restricted formal sense, in the case of agent-like systems. It is like feeling that these things may help us to understand when the system will behave in one way rather than another. And as with all trust decisions, it is like weighing up all these factors in the context of what we are considering trusting the system for, alongside the severity of the risk if it were to not to do this.

Even in the presence of this uncertainty, risk, and partial understanding, we very often choose not to stay clear of these artefacts. We often accept the risk and engage anyway, armed with the best information we can get about it. These are personal trust decisions, and to provide people with evidence against these features above is trust empowerment. It would seem to us that there is a moral imperative on those who build and release such systems into society to provide such evidence.

6. Trust and animals

Now we are familiar with the model and its use, from the previous two sections respectively, we continue, first by exploring its applicability with animals, and then in the following section, with machines of various kinds.

6.1. What it is like to trust a dog?

Both the authors of this paper, as is the case with many humans both now and throughout history, live with animals. As he was writing his paper, one of Steve’s cats disturbed him, sitting at the back door and meowing. On getting up and opening the door, Steve was led through the back porch to the garage door and had to open that too. It is winter and some of the cats like to go in and out through the garage door as opposed to the (closer) door to the outside from the porch. Marmalade made it perfectly clear that the outside door was undesirable at this time, and happily went out the other way. What does this story tell us? First, that cats communicate needs or wants, second, that they have preferences, and third, that they can express those preferences quite

6. And we note that such an existential threat is far more likely to come from the malevolent or careless use of AI systems by humans, than by any doomsday scenario foretold by sensationalist science fiction. Although even this does not negate the threat of AI: compare the trope ‘guns don’t kill people, people kill people’, and the claim that guns (and AI) are just tools, like spoons. As Stephen King so lucidly pointed out, ‘let’s see you try to kill twenty schoolkids with a fucking spoon’ (King, 2013).

5. Of course, Gehm’s corollary, that “any technology distinguishable from magic is insufficiently advanced” might also tell us a thing or two about the current state of artificial intelligence.

6. https://www.bloomberg.com/news/articles/2019-03-15/former-bitcoin-baron-mark-karpeles-gets-suspended-jail-term.
well. All of this goes through Steve's interpretation of the observed phenomena.

The point of this brief sojourn into Steve's domestic life is this: functionally, it doesn't matter if the observed phenomena are based on a well-mapped-out mechanistic response, or something less well understood, potentially more cognitive, or even some kind of utterly alien preference and decision making mechanism beyond our current knowledge and understanding. If the observed phenomena are predictable, stable, and interpretable sufficiently to be made sense of, then they can be used to aid decision making. When it has snowed, Steve's cats use a specific door. When it has not, they use another entirely. Steve, as someone familiar with these cats, can predict this behaviour or preference, depending on the cat.

Steve's Service Dog, whose name is Jessie, trusts Steve. She goes with him to many places, some of which are deeply uncomfortable for her (or any dog), but she goes because he is here and she knows he will keep her safe. For people who are privileged enough to have a Service Animal, the bond is strong and based on trust: firstly that the animal will do what is expected in even the most difficult circumstances (especially then, actually) — stopping someone who is visibly impaired from walking into traffic, or even from walking a regular route when there are insurmountable obstacles — spotting the onset of an anxiety attack in the midst of an entirely new situation, and so on. This list is endless, but the animal is trusted because they are trained and predictable and because in turn, they trust their partner to listen. Both partners in this partnership need training. It is unfortunate as well as telling that in the current pandemic many service dogs are finding their lives difficult as they aren't going out as much, and certainly not to places where their abilities are challenged. Why is this? Are they bored? Are they forgetting what the huge trust bond is like? Are they missing being trusted? What is it like to be them in that situation?

Jessie is also a therapy dog with the Saint John Ambulance. She visits people and sits with them to help them through tough times, or to raise their spirits, and so on. Her behaviour in these circumstances is both predictable and wonderful — she will find the people who need her and make sure she is there for them, even if she doesn't know them. She can be trusted to do what is right for people in different circumstances because she has demonstrated time and again that she does. However, she does this willingly because Steve asks her to, and she is quite capable of letting him know when she is tired or just doesn't want to go. Of course, it is quite possible to argue that Jessie does stuff because she lives in fear of Steve's temper, or because she thinks she may get a treat, or because it's nice to be cuddled (actually, Jessie doesn't like to be cuddled). The thing is, from a functional point of view, she does the things she does. She also appears to enjoy them, being quite willing, often excited (if that's what a dog's tail tells us) to carry on doing them.

Ultimately, the point is this: Steve's animals trust him, just as animals have trusted humans around the world, for generations. At least, according to the definitions discussed above (see Sections 2 and 3.4), they appear to, and behave accordingly. Undoubtedly, also, some empathise, but in the way of sensing emotion and acting accordingly and predictably, hence the therapy animal. The point though is that they trust because they judge by their own 'needs' and standards. These are not 'our' standards, they belong to the animals concerned, but they are no less important for that.

A further point, and the reason for using the device of an overly personal account of this sort of behaviour, is to highlight just how familiar these sorts of interactions between dogs, cats, and humans are to us. As humans, we often feel like we have some sort of shared understanding, some sort of shared semantic space, with animals that we live with, those with which we are familiar. We suspect the above passages would not seem alien to readers with a similar cultural background to Steve.

Of course, other people distrust dogs, either specific ones (based on evidence of what they have seen that dog, or that breed of dog, do), or in general. Young children, for example, often find dogs unpredictable, and coupled with their capability to cause serious injury or even death, this leads to a decision by many that (perhaps some) dogs are not to be trusted. Many people would think it an entirely sensible decision to never leave a dog alone with a baby, for example.

So, humans trust (and distrust) dogs, and dogs trust (and distrust) humans. Let us use our model to ask: what is this like? Well, in many situations, it is like believing that the dog can do what you need it to, since most of us have pretty good understanding of the sorts of things dogs are capable of. It is like believing there to be some predictability in its behaviour, and being able to increase this through training or familiarity. But here, confidence can vary based on the trustor and the dog. It is like having no expectation of the dog having any ability to make or honour formal or otherwise communicated agreements, though as discussed above, there may be tacit normalisations of mutual behaviour, as expectations are built up over time. It is often like believing that the dog is trying to do what you want, in a limited sense that is certainly of a lesser degree than could be expected from a human, but greater than would be expected of something inert like a rock. It is like feeling concerned about the consequences of bad outcomes (e.g. a dog biting a child), but with varying beliefs about the chances of this happening, depending on the situation, the trustor's disposition and the dog in question. Proxy trust also comes into play, since it is also often like believing the dog to be more trustworthy if we trust the people who trained the dog.

6.2. What's it like to trust an octopus?

Octopuses are also intelligent, roughly speaking perhaps, on a par with dogs. They “acquire information, calculate what they can do with the world around them and use flexible problem-solving strategies” (Mather, 2019). They appear to intentionally hide, wiggling partially concealed arms in order to lure unsuspecting prey. They have sufficient fine motor control and creative problem solving to open jars, untie knots, and solve mazes, trying a variety of different strategies until they succeed. They anticipate, plan, and find and use objects as tools, for example coconut shells (Finn et al., 2009), to help them achieve their goals. They play (Mather & Anderson, 1999), and have been observed to have different personalities from each other. For example, some individuals seem more more playful than others (Mather & Anderson, 1993). Octopuses are mostly solitary, and have limited social interaction. Cannibalism is prevalent in octopus populations. When octopuses do interact, they tend to do it by extending arms towards each other, rather than touching (Scheel et al., 2016). Octopuses are motivated by exploration, and by fear.

While octopuses do have brains, their neurons are not organised in the way we are accustomed to from studying mammals: more than half of their neurons are distributed about their body, primarily throughout their legs (Kier & Smith, 1985). Nevertheless, based on the notion that we have “enough information to evaluate how octopuses might have a unity behind the diversity when organizing their world”, Mather (2019) argues that octopuses have minds. Octopuses have not been observed (Mather et al., 2018) to pass the mirror self-recognition test (Gallup, 1970), though creatures like magpies have passed it (Prior et al., 2008). Yet octopuses can distinguish between humans, based on facial recognition. Perhaps, in summary, octopuses are both intelligent when viewed from a human perspective, and yet also have ‘a different way of being’ in the world (Mather, 2019).

Why is this relevant? For a start most of these interesting observations about octopuses would probably count as new or surprising information to most people. Whereas humans have lived alongside dogs for millennia, and developed a broad and deep familiarity, the same
is not true of our relationship with octopuses. Yet octopuses possess high levels of intelligence, comparable with dogs. Even if we were to spend more time around octopuses, relationship-forming might still prove difficult, as it is probably the case that their semantic space has less overlap with ours. As with Nagel’s bat, its perceptual and motor machinery, as well as the world it inhabits, are radically different from our own. It is difficult for most humans to understand how octopuses are intelligent, because their intelligence is so alien. It is not obvious how one might construct an intelligence test for octopuses, for example; the Turing Test clearly won’t do.

The notion of overlapping semantic spaces gives us a somewhat more formal way to approach the notion of peer-ness, how functionally “like a peer” some other might be. As we have seen, this is important in trustworthiness judgements.

Let us consider how a human might approach a trustworthiness judgement concerning an octopus. Most people would not understand its behaviour well enough to predict it well, nor would they have a useful model of an octopus’s motivations. As such, despite being similarly ‘intelligent’ to a dog, most people would find octopuses harder to trust, due to their lack of familiarity and the disjunction between our worlds and semantic spaces. Of course, an expert in octopus behaviour, or someone who has simply spent large amounts of time around them, may approach reasoning about their trustworthiness differently, and come to different conclusions. In summary, the reason for our brief digression into the world of octopuses is this: trustworthiness and trust decisions are not only tied to perceived intelligence in the potential trustee; familiarity plays just as important a role in accruing evidence of trustworthiness and perceived risk.

7. Trust and machines

Let us move on from rocks and animals, to consider the use of our trust model with more complex human-made artefacts.

7.1. What is it like to trust a car?

Suppose we combine our melted and filtered rocks with the remains of some billion-year-old bacteria, and reassemble them quite carefully. One possible outcome of that is the modern motor car.

Cars, of the non-autonomous kind, are also an instructive example. Suppose you are driving your car and are faced with deep water on the road. Neither you nor the car have been in this situation before, so what do you do? Do you avoid it, concerned for the risk of losing or damaging the car, or perhaps getting stuck? Or do you trust the car to get you through, aware of and despite this risk? Let us first explore our primary trustworthiness features. There is an absence of behavioural data of that might be used to assess predictability, since this is a new situation. Though we might have a hunch, we have little insight into the car’s capability to do such things, and our uncertainty about the depth of the water, or if there is anything hidden in it, adds additional doubt. We would not normally expect a car to be able to assess the situation and be honest with us about whether it intends to get through the water (like with the rock, this idea sounds silly). We also wouldn’t expect it to ‘change its mind’ half-way across (this is also outside our semantic space), though it may have some automatic control features that kick in and change its behaviour without me knowing about them. And I have confidence that any algorithms used to mediate my direct manipulation of the controls have the goal of safety rather than disobedience or recklessness.

In this case, proxy trust often plays an important role: I may find the artefact more trustworthy because I trust others who were involved in its journey to this current situation. To continue the car example, most countries have a legal road-worthiness test, which the car will have recently passed. That test will have been carried out by qualified personnel, as will have been its most recent service, and those qualifications in turn provide additional trustworthiness because I decided to trust them as facades of a broader network of assessors and curriculum designers. Similarly, I may have chosen a particular make of car, because I trust the brand or manufacturing company, which again is a facade for a large network of qualified individuals, machinery, and certified processes, covered by relevant regulatory frameworks, operating as facades for further experts I am choosing to trust. Thus, while I might choose to trust the car to make it through the water today, the evidence for its trustworthiness draws on a large web of proxy trust relationships and their own trustworthiness features.

Note that this is subtly different from the concept of transitivity in trust relationships (Falcone & Castelfranchi, 2012; Josang et al., 2012). In trust networks, transitivity refers to the idea that if X trusts Y, and Y trusts Z, what trust relationship exists between X and Z? The answer draws on social information, to provide X with evidence (available information, in our model) about Z. In proxy trust, however, if X (say, a driver) trusts Y (say, a manufacturer of Z), we do not require Y to trust Z. Rather, Y’s trustworthiness to X is proxied or projected onto Z (the car manufactured by Z is found more trustworthy to X).

Now let us consider the car that drives itself. This is often used as an example of existent AI, and so it bears discussion. There are several models of such a vehicle, from different manufacturers and using different (or the same) technologies in different ways, but the basic idea is that, when asked (ok, told) they will hold lane position, distance from other vehicles, and steer. Some go further and can navigate, and so on (done properly, the latter, notably GM’s Superruise, relies on commendably detailed mapping of what exists in the world, and thus doesn’t work in areas that are not so mapped). We said ‘ask’ earlier purposely. In most, if not all, cases, the systems will not engage if there is a reason not to: speed, visual impairment in cameras, lack of relevant data, and so on. We must also assume that the system will disengage when it cannot perform properly, but therein lies a problem. If the system disengages when the driver is distracted, safety is not just compromised, it is likely destroyed.

There are many instances of trust to pick out of even this brief description. Before we do so, consider this: the trust in fact goes both ways already. The driver trusts the car to hold its lane, and the car trusts the driver to be paying attention just in case things go awry with the lane-following trick. Let us put this another way: driving is hard. If it was easy, we wouldn’t have tests to see if we could do it.

7.2. What’s it like to trust narrow AI?

Essentially, all the examples of AI that we see in practical use today can be described as ‘narrow’ AI. That is to say, they are conceived as, designed for, and evaluated on their ability to perform specific tasks in a limited and well-scoped domain. A characteristic of narrow AI is that it typically cannot generalise its ‘intelligence’ to other domains. For example, a virtual voice assistant may be reliable at interpreting spoken dates and transferring these into online calendar entries, perhaps responding to ask how to resolve clashes, or even suggesting venues or people to invite. But such a system would likely fail miserably at interpreting an X-ray or driving a car. Indeed, they would perhaps, worse-still, be ‘not even wrong’, since their semantic spaces would likely be utterly distinct.8 Similarly, the same is true in reverse: a doctor could not reasonably expect to bring the software behind her autonomous car into the workplace in order to help analyse X-ray images. Many people, and especially governments and businesses, are referring to narrow AI when they talk today of AI at all.

It perhaps now widely known that many (but clearly not all) of these contemporary intelligent systems have the abilities they do because they have been trained to do them, using machine learning, based on

---

8 As a fun experiment, try asking Siri or Alexa or any of the many chatbots available online to teach you to drive a car, including perhaps for feedback on where you went wrong, and see where you can take the conversation.
data. Typically, the data contains positive and negative examples of desired behaviour, and the machine discovers through trial and error the programming or statistical model that is required to solve the task sufficiently reliably. How is this different from the examples of complex software systems, discussed above? In many ways, decision support systems and software agents are not different from any other piece of complex software, the like of which we discussed in Section 5.2. However as we have seen, we are interested not only (or perhaps not even primarily) in how the system is composed, but how it functions, and how this is subsequently perceived and reasoned about by others. Recalling the idea of taking the intentional stance about discrete and abstract software entities, we consider them as having some form of agency. This is the case at least insofar as that we routinely reason about them having goals and other forms of motivation (see e.g. Hawes, 2011), the ability to reflect on themselves as discrete entities in the world (Lewis et al., 2011), and to make agreements (Smith, 1980), and many other things besides. We do this, not out of any strong philosophical belief that there is undeniable truth of their agency (e.g. some philosophers have a strong conviction that a machine cannot possess agency or intelligence, only simulate it, for example, though this opens up further questions about the definition of 'machine'). We assume the intentional stance as a mindset because it is often useful, and sometimes even natural (Reeves & Nass, 1996) to think about them in this way.

In these cases of more agent-like, adaptive, or interactive narrow AI-based systems, the experience can therefore move into the domain of considering willingness, as goals and intentions captured explicitly in the system come into play. Where an AI-based system is observed to be attempting to solve a problem or achieve a goal, it can be like feeling reassured that the system’s goal is aligned with your own. It can also be like feeling uncertain about the limit of the circumstanc es in which the competence of the system will enable it to achieve that goal, or if and when the goal may change, for example in the presence of new information. And for some AI-based systems of this type, it can be like making simple, formal agreements. Issuing commands or requests can lead to the feeling that the system has agreed to do as you ask. It is usually like expecting that these will be honoured, within the competence and predictability of the machine.

Not all narrow AI appears to us like an agent. Many AI-based tools today are embedded within broader software systems, where the decision or recommendation is only part of the tool’s use. What is it like to trust these AI-based tools? To reiterate the point, it depends on who the trustor is, and what they are considering trusting it for. Sometimes, it will be like having a well-justified belief that the AI technology will be competent far beyond what might be expected from a human (including oneself), but recognising that this is true only in a very specific and narrow domain. For example, I might believe that an AI-based chess opponent is a more competent chess player than myself, or indeed than any human being alive today. Similarly, a doctor might believe that the competence of a system to provide an accurate MRI analysis exceeds any human being alive today. Similarly, a doctor might believe that the competence of a system to provide an accurate MRI analysis exceeds any human being alive today. Similarly, a doctor might believe that the competence of a system to provide an accurate MRI analysis exceeds any human being alive today.

As an example, the EU’s ‘Ethics Guidelines for Trustworthy AI’ propose that trustworthy AI will be: (1) lawful – respecting all applicable laws and regulations; (2) ethical – respecting ethical principles and values; and (3) robust — both from a technical perspective and while taking into account its social environment. The guidelines propose high-level ‘requirements’ for trustworthy AI, namely: human agency and oversight, technical robustness and safety, privacy and data governance, transparency, diversity, non-discrimination and fairness, environmental and societal well-being, and accountability. A follow-up self-assessment checklist, published in 2020, contains important and seeking questions, of the type that we believe ought to be asked of anyone releasing an AI-based system into society. Examples range from “Did you inform end-users of the duration of security coverage and updates?” to “Could a low level of accuracy of the AI system result in critical, adversarial or damaging consequences?” In the case of an AI system using online or continual learning, the checklist includes asking whether designers have considered the potential negative consequences of the system achieving its goal through unexpected or unusual methods. Perhaps

---

9 See e.g. Andras et al. (2018) for a discussion on the relationship between explainable AI and trustworthiness judgements.

10 We use this term advisedly, since it is in general insufficient to think of those humans affected by, and perhaps relying on, any of the pervasive examples of AI in operation in society, through the lens of traditional ‘computer users’. We might as well just say ‘people’ or ‘citizens’.
trusting that system will achieve a goal will sometimes come without the accompanying reassurance that it will do it in a reasonable way. Questions around the trustworthiness of an AI-based system to acceptably balance the ends vs the means become relevant.

Guidelines such as the EU’s are important, not least because they unpack likely questions concerning trustworthiness for complex machines. Further, doing this in an expert-led and consensual manner does, we believe, help to build the basis for future proxy trust decisions. Being able to say, for example, that an independent audit of an AI system against the EU’s checklist led to a particular result, would likely help us to arrive at a more informed trustworthiness judgement concerning the system in question. Drawing on the notion of system trust, capturing processes such as this within a regulatory or other governance framework can help to build this trust further (Winfield & Jirota, 2018). So perhaps trusting narrow AI is also like looking to guidelines and experts to help you make your decision.

In the currently dominant domain of data-driven machine-learning-based AI, issues of fairness, bias, vulnerability to attack, and understandability by humans are often listed as paramount to the trustworthiness of the system. On the resistance of machine learning systems to adversarial attacks, such attacks come in many forms: they can be about exposing confidential, perhaps sensitive training data (Papernot et al., 2017), or manipulating the machine learnt model such that it makes decisions that the attacker desires (Jia & Gong, 2020). And since many machine learning systems are essentially black boxes, without an explicit attempt to provide an explanation or interpretation of the results (e.g. Bibeiro et al., 2016 is an obvious example), an attack can be hard to identify. On the topics of fairness and bias, what might once have been largely theoretical discussions around sampling have instead led to intense discussions about prejudice and fairness. Examples abound. To name two that have come to light in recent years, Amazon’s automated recruiting tool, trained on data from previous hiring decisions, discriminated based on gender for technical jobs (Reuter, 2018); while commonly used AI-based face detection technologies have been found to routinely discriminate based on race and gender (Buolamwini & Gebru, 2018).

Perhaps the general point here is that, with current machine learning techniques, trusting a machine-learning-based AI system is like trusting the machine to behave in a way that automates and propagates past decisions into the future. It is perhaps not like trusting a machine to have an understanding of what a decision ought to be, or a good understanding of its consequences. Indeed, Gil et al. (2019) argue that these issues of reliability, fairness, explainability, and safety will be “of paramount importance” as the competencies and capabilities of AI advance. This is not simply a technical question: these technologies raise issues of sufficient import that political pressure groups such as the Algorithmic Justice League11 and Algorithm Watch12 are raising them in the legal and political spheres.

It would be easy to interpret much of the above as foreseeing nothing but doom and gloom for the role of AI in society. Thankfully, the seriousness with which this issue is now being treated is, we believe, cause for hope. The areas of bias, attack mitigation, and human interpretability in machine learning are now thriving and important research topics. It is not the purpose of this article to survey the very wide array of work being done to enhance these features in machine-learning-based AI. Suffice to say that this is a challenge that is motivating intergovernmental agencies to act (European Commission, 2019), standards to be established,13 multi-national companies to orient their technological efforts (Deloitte, 2019), and notwithstanding the long history of trust research in technology (e.g. Lansing & Sunyaev, 2016; Marsh, 1994; Marsh & Dibben, 2005; Muir, 1994), new academic communities and labs to spring up.

Equally important, these issues are making their way into more mainstream media; one recent example is the 2020 movie Coded Bias.14 Given how trustworthiness judgements and trust decisions are, ultimately, made by people based on their own perception and understanding (an important point noted in the business world for example by Deloitte in their 2019 report (Deloitte, 2019) on building trust in AI), bringing this discussion into the public sphere is an essential part of trust empowerment.

When considering complex autonomous machines that generate and select behaviours to achieve goals in particular contexts, we can ask of a machine: can it do what I need it to, capturing competence and predictability; and will it (decide to) do what I need it to (and when), capturing willingness and honesty. We note that the majority of recent work on trustworthy AI considers the first question; surprisingly little research has been conducted on the second. Taking a functionalist perspective, questions of willingness and honesty in narrow AI systems relate to important factors concerning their goals and capacity to deceive users. Examples include: i) what runtime goals are present and how do they relate to mine, ii) are goals transparent or hidden, does the machine deceive, iii) when will goals change, and (iv) what is the space of possible goals for the machine? As described in our model in Section 4, these are features that people will look for in assessing trustworthiness.

So what is it like to trust a modern, commercial AI technology? Well perhaps it is usually like believing that the system can do what I need it to, as long as I have a fair idea of what the limit of its operations are. It’s like finding it mostly predictable, but sometimes finding its behaviour a bit unusual and wondering why it did what it did. It’s sometimes like coming up against the limit of its operations when you’re trying to achieve something, and finding this frustrating or surprisingly funny. As a result, it is often like not quite finding it as trustworthy as you might have hoped to. Sometimes, it is like finding its decisions useful and trustworthy for quite some time, before later being disappointed (and yes, perhaps feeling a little betrayed) at realising that the decisions you were relying on were racially biased,15 or drawing heavily on factors you had hoped to exclude. When this happens, it can feel like the risks of trusting future AI-based technologies might be greater than would be immediately obvious. In issuing commands (or requests) to a contemporary AI technology, such as an autonomous driving module or a virtual assistant, it can be like believing that there

---

11 https://www.ajl.org/.
12 https://algorithmwatch.org.
13 https://www.utoronto.ca/news/u-t-s-schwartz-reisman-institute-and-ai-global-develop-global-certification-mark-trustworthy-ai.
14 https://codedbias.com/.
15 We are grateful to one of the reviewers of our paper for pointing out that we may appear to be dismissing the problem of bias and all of its ills in a rather flippant way. Such is not our intention. In all honesty, in the perspective-taking that we are attempting in this paper, we respectfully acknowledge that there is a limit to the ability of us (as white males) to properly empathize with the feelings of those on the receiving end of systems like this — we simply have not felt the weight of this problem in the same way others have. This does not mean that there is a limit to our ability to work to change things. As Ibram X. Kendi notes (Kendi, 2019), there is a problem here because of the ‘power and policies’ inherent in the systems developed today. Being who we are, we can all strive to make changes to these, whilst at the same time knowing that we will never be able to feel how it feels. Rather than attempt to expand this line of enquiry here, we merely highlight the existence and importance of the problem. We further suggest that the field would be very well served by amplifying the voices of those very people who are affected, and not, as in recent events, by attempting to silence them. We as AI researchers and practitioners have been given an immense opportunity and responsibility to actively make a difference and to not simply acknowledge that a problem exists (which is to be non-racist) but to make systems that actively change things for equality (which is anti-racist). The issue of bias in machine learning models is perhaps one of the most socially and technologically important challenges in today’s AI systems. It is well past time for all of us as practitioners and researchers to acknowledge this and work to change it.
is an agreement between you, when the system ‘accepts’ the command. It is like believing that the machine will honour this, but also understanding that there might be either things beyond its control to stop this (e.g. a loss in Internet connectivity or unforeseen environmental conditions). It might be like wondering if there is another goal within the machine that might override it, for example, a safety mode engages, a mobile robot’s charging cycle begins, or a goal placed there by the manufacturer kicks in to try to get you to change your behaviour. It might be like wondering if perhaps this is what the device was in fact intended to do.

7.3. What is it like to trust a human- or animal-like robot?

Suppose we put some of the most complex adaptive software people have yet been able to build into a machine with a broad scope of physical and interactive capabilities, for example to move about, to manipulate arbitrary objects, to perceive its environment, to vocalise utterances, and so forth. Suppose we then put this contraption in a previously unseen environment, with humans that behave as humans do. Making this machine behave either usefully or similarly to the humans is a goal of at least some in the field of artificial intelligence. What might it be like to trust such an entity? As we have seen, this will depend on the robot (most notably its software), the situation the trustor is in, and the perspective of the trustor.

Many would immediately view this as deception. Indeed, in perhaps one of the most well-known cases of humanoid robotics, the robot Sophia has drawn both widespread interest and criticism. For example, on the one hand, the Saudi government used the robot in a publicity stunt, when they made it a citizen (although it is not example, on the other hand, the Saudi government used the robot in a previously unseen environment, with humans that behave as humans do. Making this machine behave either usefully or similarly to the humans is a goal of at least some in the field of artificial intelligence. What might it be like to trust such an entity? As we have seen, this will depend on the robot (most notably its software), the situation the trustor is in, and the perspective of the trustor.

Many would immediately view this as deception. Indeed, in perhaps one of the most well-known cases of humanoid robotics, the robot Sophia has drawn both widespread interest and criticism. For example, on the one hand, the Saudi government used the robot in a publicity stunt, when they made it a citizen (although it is not quite clear what this means in this context). Hart (2018) argues that this serves to degrade the notion of citizenship and, in following, of the human rights that citizenship is intended to bestow. Meanwhile, Facebook’s Head of AI and Turing Award winner Yann LeCun rather angrily described Sophia as a puppet designed to deceive people.

Not all humanoid robots are designed to be quite as human-like. Perhaps one of the most well-known cases of humanoid robotics, the robot Sophia has drawn both widespread interest and criticism. For example, on the one hand, the Saudi government used the robot in a publicity stunt, when they made it a citizen (although it is not quite clear what this means in this context). Hart (2018) argues that this serves to degrade the notion of citizenship and, in following, of the human rights that citizenship is intended to bestow. Meanwhile, Facebook’s Head of AI and Turing Award winner Yann LeCun rather angrily described Sophia as a puppet designed to deceive people.

Not all humanoid robots are designed to be quite as human-like. Perhaps one of the most well-known cases of humanoid robotics, the robot Sophia has drawn both widespread interest and criticism. For example, on the one hand, the Saudi government used the robot in a publicity stunt, when they made it a citizen (although it is not quite clear what this means in this context). Hart (2018) argues that this serves to degrade the notion of citizenship and, in following, of the human rights that citizenship is intended to bestow. Meanwhile, Facebook’s Head of AI and Turing Award winner Yann LeCun rather angrily described Sophia as a puppet designed to deceive people.16 Not all humanoid robots are designed to be quite as human-like. Pepper (Pandey & Gelin, 2018), for example, is clearly designed to mimic human features, and to draw on these to guide interaction, while also intended to be sufficiently distinct in its appearance and behaviour to avoid direct likeness. Nevertheless, there is ample evidence of children, for example, being uncertain as to the limits of the competency of robots such as these.

Here we clearly open a window on the ethical issues, associated with drawing on people’s preconceptions, that come into play when intentionally making an entity with one set of competencies mimic another with a very different set of competencies. In these cases we see a set of narrow AI technologies appearing like a human. Questions of responsibility, especially considering the information and power asymmetries that exist in the world of commercial technology, cannot be ignored. Yet here still, the importance of understanding how humans make trust decisions regarding arbitrary others, including these machines, should be emphasised. This understanding will be an essential part of appreciating how such deception can both benefit and/or harm people.

Not quite human-like, but relevant nonetheless is the growth in so-called robotic animals, such as robotic police dogs. Clearly, the use of the term dog to describe a machine is intended to evoke a particular form of understanding and consideration by humans, drawing on our prior beliefs about dogs and their role in the police. The recent growth in the use of ‘robotic dogs’ by police, especially in the US, is interesting not because they are being trusted to carry out police work in any meaningful sense (they are not, yet), but because their presence is becoming normalised (Benson, 2021). Our familiarity with them will grow, and perhaps this is an intentional and necessary part of the trust building that would be needed for any broader rollout. A more cynical view might see this as conditioning people to accept technology within policing that they might otherwise not be willing for their government to consent to. We recall the discussion above concerning the distinction between trust enforcement and trust empowerment.

Assuming such an effort continues, what might it be like to trust a robotic police dog? Perhaps to many it would seem a little like trusting a real dog, or at least subjectively believing that it would be. And perhaps this would turn out to be a mistaken belief more often than we would like. Perhaps it would later be like feeling deceived, as we realise that real dogs don’t compare our faces against police databases, nor carry weapons, nor (usually) be willing to subject themselves to direct and serious harm solely for their handler’s benefit. Perhaps, on the other hand, it would make us feel more comfortable about the increasing prevalence of hi-tech policing; helping us to understand (and, yes, trust) the ways in which an unfamiliar technology was being used. It is interesting to speculate about how people might react differently to a hypothetical similarly programmed robotic octopus-like contraption, accompanying police officers.

What about human-like or animal-like robots beyond our current technology, but which may exist in the near or far future? Is it conceivable that a dog-like robot could be built that behaves as a real dog? We do not know how to do such a thing at the moment, but we may one day. How would we react to such a thing, how might we treat it, and what might the framing of our relationship with such a dog say about us? We do in fact have some clues, for example when we look at robotic pets in Japan.

Science fiction allows us to explore topics such as this. Star Trek, for example, encouraged us to consider how might we decide to trust both unfamiliar aliens as well as highly functional human-like machines. What is it like to trust Commander Data, for example? It would be different from how we trust a human, but it would almost certainly not be the case that there would be no trust. Indeed, Introna (2010) notes that Data has become a ‘trusted friend’ of the other characters, and in the story this trust clearly predates the issue of his legal status as a person being settled. An analysis of Data’s features against the trustworthiness story this trust clearly predates the issue of his legal status as a person being settled. An analysis of Data’s features against the trustworthiness features in the model, would likely yield a rich understanding of the likely outcomes of trusting ‘him’.

Perhaps more importantly to our discussion, Introna (2010) highlights that while an anthropocentric view of others encourages us to consider the extent to which another entity (animal or machine) is the same as a human, more generally we must consider all the range of ways in which entities of different types might be different from us and each other. Here, we see Sloman’s point about varieties of intelligence made again, as we discussed in Section 5. We would expect the trust heuristic, as something that evolved in primarily human–human social relationships, to indeed be biased towards anthropocentric considerations. But this does not mean that it excludes, or ought to exclude, all else. Indeed, as our understanding of other forms of beings and intelligences grows beyond Descartes’s human exceptionalism (Beckoff, 2011), perhaps trust as a decision-making heuristic has the potential to become more useful, not less.

One may decide to conclude that the gap between today’s realistic human-like robots and a conception such as a fictional android is just too great to make this question particularly important. But in that case, how would we know when we have created something ‘worthy’ of being considered for a trust decision in a similar way to how we consider other humans? And where would animals fit? If we discover or accept that there are animals capable of making simple agreements, for example, or adopting the goals of another, then is it not entirely reasonable to make trust-based decisions that take into account these features of them? As evolution makes clear, variations in species are by degree, not by kind (Beckoff, 2011). There are machines that can already do these things, or at least something functionally equivalent to them. And as we have seen, trustworthiness judgements are made

---

16 https://www.hansonrobotics.com/sophia/.
17 https://www.facebook.com/yann.lecun/posts/10155025943382143.
by a subject based on their understanding of an object’s behaviour and internal processes; they are not solely (or perhaps even primarily) determined by what that object is made from.

So, what is it like to trust a human-like robot? Sometimes, it is like believing that the robot can do what I need it to. And sometimes like being right about that, but often not feeling very sure about its limits. It may be like wondering if the robot’s real competencies have been masked by its image. Sometimes, it might be like believing that the robot will expose its capabilities to me, including the limits of them, and that this will give me some useful trustworthiness information. It is probably like believing its behaviour to be somewhat predictable, but being unsure how it will behave in new situations that is wasn’t programmed for or trained on. It is sometimes like believing that the robot exists to try to achieve goals given by you or another human, in a limited sense, but again being unsure of these limits. It is possibly sometimes like expecting the robot to adopt new goals relevant to you, your needs, or your situation. It might be like believing the robot will do this proactively.

For someone with training in modern AI, it might be like trying to reason about the robot’s goal-alignment: considering how it represents utility/reward/objective/cost/ error functions or goal states, and reasoning about these as if they represent what the robot is ‘attempting’ to do. It may be like believing these goals to be aligned with mine, but simultaneously believing that the robot would have no idea if it were to violate a norm in trying to achieve them. We imagine that it might be like feeling concerned about the consequences of this, depending on the robot and the situation, so it might be like not trusting it to operate completely autonomously around people. It might be like considering the robot to be ‘honest’ in some way, since it has the capability to expose its goals to me, and I believe it is doing so without a conflicting or overriding goal to deceive me about those capabilities. It might even be like believing that the robot is helping me to understand them. On the other hand, it might be like never being totally sure if it is being fully honest, and in this respect it might be a little like trusting a human. Perhaps the robot has an overriding goal I am not aware of and will deceive me through decision or design, i.e., it might be like wondering if it will it tell me that its goal is X, and then pursue goal Y instead.

Proxy trust is also often a crucial part of the picture here: making trustworthiness judgements of robots is often like drawing heavily on factors such as who made the robot, and what laws, regulations, and standards are in place for robots in the situation.

Finally, it will be like weighing up all these factors against the risk, the severity of the outcomes, one’s attitude to risk, and what one’s alternatives are in the situation.

It follows that, we cannot simply tell people that they ought to trust a particular robot, or that they ought not trust it. Nor should we be pursuing an agenda of trying to convince people to trust robots or any other kind of agent, or indeed rocks (i.e., trust enforcement), as this motivates deception. Conversely, neither can we escape the fact that whether we think that trust ought to be used or not in the domain of complex intelligent machines, many will go ahead and use trust-based reasoning to aid their decisions regardless. We cannot put it better than the Johns Hopkins University Institute for Assured Autonomy, when they argue\(^{18}\) that people “need to be able to trust the autonomous systems in their lives,” and further, “if people cannot depend on or trust AI, then they will not choose to have it in their lives. People choosing to have intelligent autonomous systems in their lives will therefore be evidence of the trust placed in them”. As the creators of today’s and tomorrow’s intelligent systems, it is our duty to build technology that empowers them to make this choice.

---

\(^{18}\) https://iaa.jhu.edu/.

8. A pragmatic functionalist perspective

At this point, the reader may be thinking: “Okay, yes the model can be applied to all sorts of artefacts, but isn’t the more important question whether it makes sense to think in terms of an artefact’s capacity for honesty, integrity, willingness, and benevolence?” Isn’t this the crux of the matter?

It is tempting to think that one ought apply trust-based reasoning, including judgements of trustworthiness, only to those with agency (the capacity to make decisions in a situation), or perhaps even moral agency (the capacity to be held responsible for ones decisions and actions). Indeed, as discussed above in Section 3.2, this is the starting point for much of the literature in social trust (e.g. Sztompka, 2000), where trust is taken strictly to be a relationship between peers. But then, what form of agency would be required?

If we take a hard view on a requirement for moral agency and free will in considering trustworthiness, then considering trusting (according to our model) any individual without free will leads to the only non-zero trustworthiness features being competence and predictability; both of these are readily amenable to analysis in machines, and we can arrive at a perfectly justifiable ‘trust’ decision based on that information alone. Many engineers do. And we can even decide that we prefer the term ‘accepted dependency’ to ‘competency trust’ to describe this. If, on the other hand, we also find it useful to consider functionally equivalent (perhaps proto) forms of honesty, integrity, willingness, and benevolence in machines and animals, then this gives us the opportunity to make more informed trustworthiness judgements. The model also allows us to capture how people may already be doing this in their personal interactions with animals and machines (whether we think they ought to or not). A functionalist perspective on these mental features, as illustrated in our examples in Sections Section 5, 6, and 7, admits this analysis. And after all, understanding how people are making trust decisions, and empowering them to make better trust decisions, is what this is all about.

In doing this, some may consider us to be choosing to treat an artefact as if it were an agent, simply because this helps us reason about it better, given our finite mental abilities. We reiterate that this is not a new idea. And to a certain extent we are indeed open to the accusation of anthropomorphising artefacts, applying a heuristic usually used with other humans in the non-human domain; though (following Dennett, 1971) this is neither necessarily a bad thing, nor is it something we will ever stop people from doing anyway (Reeves & Nass, 1996), and it is also certainly not new either (Marsh et al., 2011; Muir, 1994). Attributing agency to certain artefacts can indeed be useful, in a functional, evolutionary sense, even where this does not represent a complete or even accurate mental model of the object (Hutson, 2012). After all, the mind has evolved to enable humans to make good decisions in bounded situations; this is the reason why heuristics exist and dominate decision making. Further, we find it an unwarranted case of human exceptionalism to claim that trustworthiness can, in principle, only be applied to other human minds. Any such specially reserved role for trust is indeed insufficient: features used in assessing trustworthiness develop over time in children (Harbaugh et al., 2003); and trust is also found in the relationships of chimpanzees (Engelmann & Herrmann, 2016), demonstrating that it is not unique to humans. Trustworthiness is also a concept used effectively to reason about the behaviour of working dogs (Marker et al., 2005); earlier in this article we also discussed common personal experiences of trust with dogs. It is conceivable that one day we will be confronted with the challenge of deciding whether to trust unfamiliar aliens or machines not built by ourselves. Here, we are reminded of the ‘Other Minds reply’ to Searle’s Chinese Room, and the accusations of chauvinism that have been levelled against the human-centric nature of the Turing Test (Saygin et al., 2000).

We hope that we have demonstrated in this article that deciding whether we trust entities we encounter in the same way that we trust other humans is interesting but not the most important question. A better question is: when and how will we decide whether to trust intelligent
entities of various kinds, entities that we find ourselves in a position to interact with, both now and in the future?

It does behoove us to think carefully about the words we use: do we mean to convince people to use (our) AI because we believe in it? Or do we mean to create AI that many people truly find trustworthy? Or do we want to make AI that people can depend on, without the need for what some might consider misplaced) trust? Or perhaps we simply want to understand when and how humans do use the trust heuristic with AI, to help us design better socio-technical systems. At the start of this paper we brought up Boden’s definition of AI being about machines that do the sorts of things that minds can do (Boden, 2016). Consider also our lack of understanding about precisely what sort of mind-like entities we might encounter, and the added uncertainty from the fact that trustworthiness judgements are themselves subjective heuristic assessments of another’s ‘mind’. We are thus faced with a simple choice: do we choose to disengage from reasoning using trust and trustworthiness for artefacts that appear mind-like, when we cannot be fully sure they are capable of being trustworthy? Or indeed, and perhaps with a small sense of irony, do we choose to trust that this way of reasoning has the potential to be useful, in the context of systems that we cannot fully understand, predict, or control?

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Peter R. Lewis is supported by an NSERC Canada Research Chair grant on ‘Trustworthy Artificial Intelligence’ (number 950-233208). Stephen Marsh is supported by an NSERC Discovery grant ‘Information Transparency Through Trust’, number 005897-2018. The thoughts in this paper were born during many productive discussions at the Trusting Intelligent Machines (TIM) workshop at Schloss Raischoltzhausen in Germany. Our thanks to the fantastic workshop attendees and in particular to Michael Gucket, the staff at the Schloss and Universität Gießen for furnishing such a great place to think.

References

Andras, P., Esterle, L., Guckert, M., Han, T. A., Lewis, P. R., Milanovic, K., Terry Payne, C. P., Pitt, J., Powers, S. T., Urquhart, N., & Wells, S. (2018). Trusting intelligent machines: Deepening trust within socio-technical systems. IEEE Technology and Society Magazine, 37.

Avilés, A., Laprie, J.-C., Randell, B., & Landwehr, C. (2004). Basic concepts and taxonomy of dependable and secure computing. IEEE Transactions on Dependable and Secure Computing, 1(1).

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

Barber, B. (1983). Logic and limits of trust. New Jersey: Rutgers University Press.

Beckhoff, M. (2011). Animal minds and the feasible human exceptionalism. Psychology Today, URL: https://www.psychologytoday.com/us/blog/animal-emotions.

Benson, T. (2021). Robotic police officers are slowly being normalized, whether we like it or not. URL: https://www.digitaltrends.com/features/robot-law-enforcement-

Bochot, T., Vreteluizer, P., Waevelynck, H., & Wels, V. (2009). Model checking flight control systems: the Airbus experience. In Proceedings - International conference on software engineering.

Boden, M. A. (2016). AI: Its nature and future. Oxford University Press.

Boon, S. D., & Holmes, J. G. (1991). The dynamics of interpersonal trust: resolving uncertainty in the face of risk. In R. A. Hinde, & J. Groebel (Eds.), Cooperation and prosocial behaviour (pp. 190–211). Cambridge University Press.

Bryson, J. (2018). AI & global governance: No one should trust AI. United Nations University Centre for Policy Research Article. URL: https://cpr.unu.edu/publications/articles/ai-global-governance-no-one-should-trust-ai.html.

Bucalamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on fairness, accountability and transparency. Proceedings of Machine Learning Research.
Lansing, J., & Sunyaev, A. (2016). Trust in cloud computing: Conceptual typology and trust-building antecedents. ACM SIGMIS Database: The DATABASE for Advances in Information Systems, 47, 58-96.

Lewis, P. R., Chandra, A., Parsons, S., Robinson, E., Glette, K., Bahnoon, R., Torresen, J., & Yao, X. (2011). A survey of self-awareness and its application in computing systems. In Proceedings of the international conference on self-adaptive and self-organizing systems workshops (pp. 102–107).

Loggins, K. (2020). Here's what happens when an algorithm determines your work schedule. Vice.

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

Luhmann, N. (1990). Familiarity, confidence, trust: Problems and alternatives. In D. Gambetta (Ed.), Trust (pp. 94–107). Blackwell.

Lumsden, J. (2009). Triggering trust: To what extent does the question influence the answer when evaluating the perceived importance of trust triggers? In People and computers XXII: celebrating people and technology. RCS.

Lush, V., Lumsden, J., & Bastin, L. (2018). Visualisation of trust and quality information for geographical dataset selection and use: Drawing trust presentation comparisons with B2c e-commerce. In Proceedings of the 12th IFIP international conference on trust management (pp. 75–90). Springer.

Marker, L. L., Dickman, A. J., & Macdonald, D. W. (2005). Perceived effectiveness of livestock-guarding dogs placed on Namibian farms. Rangeland Ecology & Management, 58(4), 329–336.

Marsh, S. (1994). Formalising trust as a computational concept. (Ph.D. thesis), University of Stirling.

Marsh, S., Briggs, P., El-Khatib, K., Esfandiari, B., & Stewart, J. A. (2011). Defining trust as a computational concept. In Proceedings of the 5th international conference on learning representations. URL: https://arxiv.org/abs/1610.05755.

Pennington, R., Wilcox, H. D., & Grover, V. (2003). The role of system trust in business-to-consumer transactions. Journal of Management Information Systems, 20(3), 197–226.

Prior, H., Schwarz, A., & Güntürkün, O. (2008). Mirror-induced behavior in the Magpie (Pica pica): Evidence of self-recognition. PLoS Biology, 6(8), e202.

Reeves, B., & Nass, C. (1996). The media equation: How people treat computers, television, and new media like real people and places. Cambridge, U.K.: Cambridge University Press.

Reina, D. S., & Reina, M. L. (1999). Trust and Neterayal in the workplace: Building effective relationships in your organization. San Francisco: Berrett-Koehler Publishers.

Reuters (2018). Amazon ditched AI recruiting tool that favored men for technical jobs. The Guardian.

Ribero, M. T., Singh, S., & Gueschin, C. (2016). Why should I trust you?: Explaining the predictions of any classifier. In Knowledge discovery and data mining (pp. 1135–1144).

Ryan, M. (2020). In AI we trust: Ethics, artificial intelligence, and reliability. Science and Engineering Ethics, 26, 2749–2767.

Sako, M., & Helper, S. (1998). Determinants of trust in supplier relations: Evidence from the automotive industry in Japan and the United States. Journal of Economic Behaviour and Organization, 34, 387–417.

Saygin, A. P., Cicelli, L., & Akman, V. (2000). Turing test: 50 years later. Minds and Machines, 10, 463–518.

Schell, D., Godfrey-Smith, P., & Lawrence, M. (2016). Octopus use signals during frequent conspecific interactions. Current Biology, 26, 377–382.

Sloman, A. (1996). What is it like to be a rock? URL: https://www.cs.bham.ac.uk/research/projects/cogaff/misc/rock/.

Sloman, A. (2013). Virtual machine functionalism: The only form of functionalism worth taking seriously in Philosophy of mind. URL: https://www.cs.bham.ac.uk/research/projects/cogaff/misc/vm-functionalism.html.

Sloman, A., & Chislet, R. (2003). Virtual machines and consciousness. Journal of Consciousness Studies, 10, 133–172.

Smith, R. (1980). The contract net protocol: High-level communication and control in a distributed problem solver. IEEE Transactions on Computers, C-29(12), 1104–1113.

Sønderskov, K. M., & Dinesen, P. T. (2016). Trusting the state, trusting each other? In Proceedings of the 5th international conference on learning representations. URL: https://www.cs.ox.ac.uk/research/projects/cogaff/misc/...