Trusting AI: Explainability vs. Trustworthiness

Draft of July 30 2023

Mona Simion & Christopher Willard-Kyle

Cogito Epistemology Research Centre, University of Glasgow

*§1 Introduction*

It’s one thing for an AI system to be reliable at what it does. It’s another thing for us to *trust* it to do what it’s supposed to. Finally, it is yet another thing to *rationally trust* it to do what it’s supposed to do. Nevertheless, the reliability and trustworthiness of an AI system can’t be fully separated in practice. AI can only succeed at its task if it is *used*, and it will only be used if its users trust it: ‘If the users do not trust a model,’ Ribiero *et al* (2016: 1135) warn, ‘…they will not use it.’ And it is crucial that some AI *is* used—as one example, AI systems have proven very effective at screening certain cancers (see, e.g., Harmon *et al* (2020)).

What stops people from trusting AIs? There’s no one answer to this question, but it’s normal to look to the *black box problem* as the among the primary culprits. The black box problem, in short, is that AI systems are often able to generate a *verdict* but unable to generate an *explanation* of how the system came to that verdict. The black box problem is nicely illustrated in the case of Lucie from Cappellen and Dever (2021): Lucie is confused after her request for a loan is denied due to the AI system (‘SmartCredit’) determining that she has a very low credit score. Lucie asks the bank to explain *why* she has been assigned a low credit score. They can’t give an answer: SmartCredit makes reliable predictions (we are to suppose) about who is likely to pay back their loans, but not even its creators can look under the hood and understand why Lucie’s score, in particular, is low. Lucie pushes back:

Perhaps she [Lucie] can at least be told what it was in her history that SmartCredit found objectionable. Was it her low annual income that was responsible? Was it those late credit card payments in her early twenties? Or was it the fact that she follows a number of fans of French film on Twitter…

Unfortunately, the bank doesn’t have much helpful to say about this, either. It’s easy enough to spot particular variables in the initial data set—the bank can show her where in the input her annual income is, and where her credit card payment history is, and where her Twitter follows are. But they don’t have much to say about how SmartCredit then assesses these different factors. All they can do is point again to the cascading sequence of calculations—there are the initial numbers, and then there are millions upon millions of mathematical operations on those initial numbers, eventually dropping out a final output number. The bank explains that that huge sequence of mathematical operations is just too long and complicated to be humanly understood—there’s just no point in trying to follow the details of what’s going on. No one could hold all of those numbers in their head, and even if they could, it’s not clear that doing so would lead to any real insight into what features of the case led to the final credit score. (Cappellen and Dever 2021: 9–10)

Lucie’s frustration is understandable. And—whether or not she *should* trust SmartCredit—it’s easy to imagine the inability her (or anyone else’s) inability to comprehend SmartCredit as partly responsible for her felt difficulty in *trusting* SmartCredit. As von Eschenbach (2021: 1620) puts things:

Trust requires that we have reason to believe both that AI is reliable and acting in our behalf, and so transparency into how AI operates and reaches its outcomes and predictions is needed in order for us to be able to judge AI to be trustworthy.

And indeed, a growing number of theorists have been calling for the development of *explainable* AI (XAI), explicitly as a means to building trust in AI. If we want people to trust (predictive) AI, the AI should be transparent: we should be able, at least in general, to understand why the AI makes the predictions it does.

Cases like Lucie’s motivate the following **uniformity assumption** for trusting AI:

**Uniformity Assumption:** Typically in AI communication, user rational trust (in AI) requires user understanding. And user understanding requires an explanation of why the AI has reached the conclusion it has.

In turn, the Uniformity Assumption motivates the following very popular view on what user rational trust in AI requires:

**The Explanation View of AI Trust**: User rational trust in AI requires an explanation of why the AI has reached the conclusion it has.

We think that Uniformity Assumptionis wrong, and, mutatis mutandis, that we should abandon the Explanation View. It’s not true of trust in general that rational trust (even typically) requires understanding why, and it’s not the case that AI communication generates any special normative requirement that there should be an explanation why that grounds rational trust.

This doesn’t mean that we think there’s nothing to be gained by XAI—we prefer explainability all else being equal! But understanding how to increase trust (when appropriate) in AI requires the right diagnosis. In order to understand how to increase trust in AI, we think it’s better to focus not on AI *explainability* but instead on AI *trustworthiness*. That is, in this paper, we will defend what we call the Simple View:

**The Simple View of AI Trust**: User rational trust in AI requires AI trustworthiness.

In §2, we’ll explain why we think focusing too narrowly on AI *explainablity* as a means to increasing public trust in AI is problematic. In §3, we turn from *explainability* to *trustworthiness*, but we observe that most extant accounts of trustworthiness are not applicable to AI because of the *anthropocentric challenge*. In 4, we’ll turn to our positive proposal: we’ll develop an account of AI *trustworthiness*. We’ll also defend a view about those circumstances in which it is appropriate to *attribute* trustworthiness to AI. As a whole, the paper suggests that re-orienting discussions of AI trust around AI *trustworthiness* rather than *explainability* both better diagnoses failures of trust in AI and makes available a wider range of corrective interventions.

*§2 Trust without Explainability*

The Black Box Problem is intellectually mesmerizing—it’s strange, even eerie, that we should come possess new information without any clear idea whence the information comes. And, in fairness to the XAI movement, we should note that in typical conversations, we *can* ask for explanations from our interlocutors. Indeed, some philosophers have thought (though we are sceptical) that asserting *p* inherently involves committing to provide a justificatory defence of *p*—the sort of thing that offering an explanation why for paradigmatically is (cf. Brandom 1983: 641). One way of framing the black box problem is that we *can’t* engage in this type of exchange, even in principle, when an assertion comes from the relevant kind of AI.

For all that, we’re sceptical that there is any deep connection between trust and explainability. Consider an ordinary case of testimony:

A: Which way is the nearest subway stop?

B: Take a left at Great Western.

A: Ok, thanks!

At no point does B ask A for an explanation of either why or how they know that the nearest subway is to the left on Great Western. Nevertheless, it seems perfectly appropriate (in normal circumstances) for A to *trust* B that the nearest subway is where they say it is—to take them at their word. It’s a perfectly ordinary case of trust, and explanation just doesn’t come into it. Furthermore, all theories of testimonial justification give (or strive to give) this prediction.[[1]](#footnote-1)

Of course, this exchange is not very much like testimony from AI. A can easily imagine how B could in principle explain why [they believe that] the subway is to the left on Great Western (‘I saw it on a map,’ ‘I pass it every day on my way to work’). The modest point is that it’s not as though actual explanation needs to be a part of trust in general. In fact, we often *don’t* ask for explanation precisely *because* we trust someone enough to take them at their word. If I say that I trust my child to walk to school by themselves, but in fact I constantly track their phone every time they leave the front door, it’s not clear I *really* trust them to walk to school by themselves: a robust enough trust wouldn’t require independent verification.[[2]](#footnote-2) Similarly, if I say that I trust someone to tell me the truth, but in fact I always check what they say by seeking an explanation of why, it’s not clear that I *really* trust them to tell me the truth: a robust enough trust wouldn’t require independent verification.

One might object: it’s true that there’s no *actual* explanation why offered in the example of trust above, but B’s assertion is explainablein principle. What’s distinctive of AI testimony in cases like Lucie’s credit score is that the answer is not even explainable in principle.

Nevertheless, one can trust testimony even when explanation is not even available in principle to one due to the e.g. complexity of the issue. This often happens when we trust expert testimony. Maybe *I*’ll never be able to understand quantum mechanics. For me (given my training and capacity for understanding, etc.), the field of quantum mechanics is effectively a black box. My situation with respect to QM is roughly the same as Lucie’s with respect to SmartCredit. I can see that physicists put some values into some equations that generate certain outputs, but I can’t follow the calculations to their end, nor do I really understand why physicists treat the input values the way that they do. Even so, I still trust scientific experts when they tell me that quantum mechanics accurately describes reality, or make a prediction on the basis of QM.

Here’s another way to put the same point. When we trust others by accepting their say-so, we *can* seek understanding. But often, knowledge is enough. When we come to know what someone says by accepting their testimony, we succeed through trusting them, even if we do not at all understand *why* the content of their testimony is true.

To sum up, ordinary cases of human trust—and particularly human trust in testimony—don’t seem to demand explainability as a condition on (rational) trust. So, it’s not as though there’s any *in principle* link between trust and explainability. We should be suspicious, therefore, that the mere fact that some AI testimonies are (to us) inexplainable by itself explains failures to trust.

*§3 Trustworthy AI: The Antrhopocentric Challenge*

Instead of focusing on whether AI is explainable, we should focus on whether AI is *trustworthy*.

Sometimes we trust the untrustworthy. When we do, that constitutes a mistake—even if the untrustworthy person happens to come through on the particular occasion. Trust is properly given only if the entrusted party is trustworthy (cf. Carter 2022). Rationally speaking, then, we endorse the following principle:

**Rational Trust**: S is rational to trust R to φ only if S is justified to believe R is trustworthy with respect to φ-ing.

One might think Rational Trust is too strong: can’t I rationally trust you to φ merely in virtue of being justified to believe that you will φ? The answer is ‘no’: justified belief that you will φ and justifiedly trusting you to φ come apart. I can justifiably believe that you will φ merely in virtue of knowing that I will force you to φ. That is hardly tantamount to trusting you to φ.

Alternatively, one might think Rational Trust is too weak. [[3]](#footnote-3) What one needs isn’t (merely) to justificably believe that R is trustworthy with respect to φ-ing but also hat R *will* φ. Two things about this: first, crucially, Rational Trust merely states a necessary condition on rational trust. Second, we think that requiring justified belief that R will φ would be too strong: Trust (often) makes us vulnerable. It can be rational to trust even in circumstances in which we do not have any further justification (other than the evidence of trustworthiness) to believe that our trust will be satisfied.

Here’s one way that we think focusing on trustworthiness helps in the case of AI. Being trustworthy is widely taken to involve more than mere reliability.[[4]](#footnote-4) Suppose that your colleague always brings a lunch to work. They are, therefore, maximally reliable at bringing their lunch to work. But this doesn’t mean they are *trustworthy* with respect to bringing their lunch. After all, it’s not their job to bring lunch. If tomorrow, they were to go out to a restaurant for lunch instead of bringing lunch from home, that’d be a surprise, but it wouldn’t be a betrayal or breach of trust. You can *count* on them to bring their lunch to work, but you can’t (rationally) *trust* them too.

We think this gap between reliability and trustworthiness is an important place to look when trying to understand failures to trust AI. Sophisticated AI can be very reliable—but that, by itself, isn’t enough to make AI trustworthy. If Lucie ought to trust SmartCredit (and as a reminder: we haven’t made a verdict one way or the other), it isn’t *merely* because SmartCredit is reliable. Or, put another way, even if Lucie *knows* SmartCredit is super reliable, that won’t settle the question of whether it’s rational for her to *trust* SmartCredit.

That isn’t to say reliability doesn’t matter for trust. It does. But trustworthiness requires both reliability and some condition X—whatever it is, e.g., that would transform my colleague’s failure to bring lunch from a (mere) *surprise* to a *betrayal* (of trust).

But introducing condition X creates an immediate problem for the very concept of trustworthy AI. Extant accounts of condition X are very often anthropocentric—they are categories that don’t seem to apply to (a wide range of) AI. Insofar as one of those theories is right, it just *can’t* be rational to trust AI. AI wouldn’t be the kind of thing that can be rationally entrusted. Call this the *anthropocentric challenge* for trusting AI.

We think the anthropocentric challenge can be overcome. The right account of condition X is *not* anthropocentric and can countenance trustworthy AI (and trustworthy artefacts in general). Indeed, we think this is one of the virtues of our preferred account. But to fully appreciate the anthropocentric challenge, we begin with some competitors to our account, on which it is difficult to see how AI could be trustworthy (and, therefore, difficult to see how AI could be rationally trusted).

First, consider Annette Baier’s (1986) influential account of trustworthiness. On her view, trustworthiness is reliability due to *good will*: the trustworthy person not only reliably fulfils their commitments but does so in virtue of their goodwill toward the trustor.

This account makes it difficult to see how AI could be trustworthy. SmartCredit doesn’t have goodwill toward Lucie, for instance. SmartCredit is completely indifferent toward Lucie. Perhaps a *very* sophisticated AI could have goodwill toward Lucie—sci-fi AI like C-3PO or a Hal 9000 might develop the right sort of affective capacities to host a good or (in the latter case) bad will toward people. But most of our current AI is, we presume, not like that. So, on Baier’s account, most AIs are not candidates for trustworthiness or (via **rational trust**)proper objects of (rational) trust.[[5]](#footnote-5)

Crucially, we have argued in previous work (REFERENCE) there are problems for Baier’s account even when setting the anthropocentric challenge aside. Suppose that George has promised Ann that he will wash the dishes every weeknight, and he deeply resents Ann for this. He washes the dishes out of a sense of duty, be he hates doing so and comes to deeply resent Ann for (as he sees it) extracting this promise from him. George does not have goodwill toward Ann. His only feeling for Ann is deep resentment. Nevertheless, given that he always *does* wash the dishes just like he promised, and because he promised, it seems that he *is* trustworthy with respect to washing the dishes. Ann could justly criticize George for his unjust sense of resentment, but she can’t fault him for being untrustworthy with respect to washing the dishes. And, indeed, it’s rational for Ann to continue to trust George to wash the dishes, despite his antagonism.

Substituting a good will for condition X, as Baier does, makes rational trust in (most) AI impossible. But a good will is the wrong substitute anyway. It’s time to turn to other accounts.

According to Nancy Potter (2002: 25, emphases ours), being trustworthy amounts to being ‘one who can be counted on, *as a matter of the sort of person he or she is*, to take care of those things that others entrust to one’. Being the right sort of person involves having a good or virtuous character. A virtuous *character* is a robust thing. It’s not clear at all that (most contemporary) AI systems like SmartCredit have anything like virtuous characters, or any moral characters whatsoever. Accordingly, on Potter’s account, AI systems are not candidates for trustworthiness. Given **Rational Trust**, they aren’t candidates to be rational objects of trust either.[[6]](#footnote-6)

But once again, there are independent reasons to be sceptical of Potter’s precise proposal. Recall Ann and George. It’s not clear that George has a good character—he seems like a jerk. But in the original version, we could at least say this for him: he washed the dishes out of a sense of duty. And perhaps that’s enough to attribute to him a kind of (thin) good character, however unfair his attitude toward Ann. But we can tweak the example further to ensure he does not meet Potter’s condition. Imagine that George doesn’t do the dishes out of a sense of duty at all, but out of a sense of fear. He believes that God will punish him if he doesn’t fulfil his promises to Ann; otherwise, he wouldn’t wash the dishes. Even so, if George really does reliably and stably fulfil his promise to Ann to wash the dishes, it seems that he *is* trustworthy with respect to washing the dishes. Ann could justly criticize George for his bad character, but she can’t fault him for being untrustworthy with respect to washing the dishes. And, indeed, it’s rational for Ann to continue to trust George to wash the dishes, despite his flawed character.

Other accounts of trustworthiness face similar challenges. Paul Faulkner (2014: 1978) argues that affective trust requires thinking that the entrusted will be motivated to φ on the basis that the trustor is depending on the entrusted to φ. Again, it’s not clear that AI have the right sort of motivational profile to be the objects of affective trust on Faulkner’s account. But it’s also unclear why we shouldn’t say that Ann can trust George to wash the dishes (assuming he reliably fulfills his promise to do so) even if George is not in any way motivated by the thought that Ann is depending on him to do so.

Focusing on the inner will, character, or motivation of the entrusted has led us into problems. That suggests a more externalist approach. We think Hawley’s (2020) work is near the mark. According to Hawley, trustworthiness is simply reliability in avoiding unfulfilled commitments: this requires both caution in incurring new commitments and diligence in fulfilling existing commitments.

Can AI be trustworthy in Hawley’s sense? It seems so. After all, AI can avoid unfulfilled commitments. In fact, if anything, one might worry that Hawley’s view makes it rather *too* easy for AI to be trustworthy. That’s because Hawley’s account of trustworthiness is *negative*, and trust is achieved by being reliable at *not* violating commitments. It’s not clear that (most contemporary) AI are capable of making (or having) commitments at all. If so, then it’s almost trivial that AI are trustworthy—they are extremely reliable at avoiding unfulfilled commitments because they are extremely reliable at not making commitments in the first place!

Unsurprisingly, the negative character of Hawley’s proposal is its most controversial element. Suppose George meticulously avoids ever making any promises or commitments to Ann, even implicitly, even though Ann is his friend and in urgent need of help. Is this, by itself, enough to make George a trustworthy friend to Ann? Not obviously.

For this reason, we favour a more positive account of trustworthiness, that is, nonetheless, similar to Hawley’s in its externalist leanings: it is not requiring agents to have any particular character or motivational state in order to count as trustworthy. We now turn to our preferred approach.

*§4 Trustworthy AI: An Ought-Based Account*

Several of the definitions of trustworthiness in the last section made ‘trustworthy AI’ something of a category mistake. AIs *can’t* be trustworthy because they (typically) don’t have the right kind of inner lives or moral characters. Methodologically, we prefer to go the other direction. We think it’s intuitive that AI *can* be trustworthy. In this way, we agree with, e.g., Cappellen and Devers (2021) that accounts of trustworthiness ought to be de-anthropocentrized.

On our preferred account, then, here is what it is to be maximally trustworthy:

**Maximal Trustworthiness** One is maximally trustworthy with regard to φ-ing if and only if one has a maximally strong disposition to fulfil one’s obligations toφ. (Kelp & Simion 2023)

One important clarification: in some natural languages (English included, and in contrast with e.g. most Romance languages) ‘obligation’ is often used in a normatively ‘oomphy’ fashion. Here, we use obligation in a technical, non-normatively-thick sense, as per standard deontinc logic, to simply map on to *ought*:[[7]](#footnote-7) being under an obligation to φ, then, does not imply, in our sense, anything more normatively ‘oompmhy’ than simply being the subject of an ought to φ.

Consider, again, George and his level of trustworthiness with respect to washing the dishes. If George has a maximally strong disposition to fulfill his obligations to wash the dishes, then he is maximally trustworthy to do so. It doesn’t matter *why* he is disposed to fulfill those obligations. It *could* be feelings of goodwill (as for Baier), a good character (as for Potter), or being motivated by the fact that someone is depending on him to do so (as for Faulkner), but none of these *must* be the reason George is disposed to fulfill his obligations. Having the relevant disposition is enough, regardless of what inner reason (if any) motivates George to have that disposition.

Trustworthiness, on this account too, comes apart from reliability: Recall the case of my colleague who always brings a lunch from home. They are *reliable* with respect to bringing a lunch from but not *trustworthy* with respect to bringing a lunch from home because bringing a lunch from home does not fulfil any relevant obligation, since he is not under ay relevant obligation. On the other hand, if your colleague *promised* you that they would bring lunch—thereby incurring an obligation to do so—then their strong disposition to bring a lunch would make them trustworthy with respect to bringing a lunch.

Let’s clarify a few details before pressing on. Dispositions have both trigger conditions and manifestation conditions. The ‘trigger’ condition for trustworthiness is having an obligation to φ: obligations occasion the opportunity for trustworthiness. The ‘manifest’ condition for trustworthiness to φ is simply φ-ing, or fulfilling one’s obligation to φ.

Dispositions should be evaluated relative to suitable conditions. One can be maximally trustworthy with respect to φ-ing even if one fails to fulfill one’s obligations to φ in unsuitable situations. For example, it wouldn’t count against George’s trustworthiness to wash the dishes if he failed to manifest his disposition to do so when the water stopped working.

So far, we’ve given an account of maximal trustworthiness. But trustworthiness is a matter of degree. How trustworthy one is, we think, is just a function of how close someone is to being maximally trustworthy. Suppose George reliably does all his homework (when obligated to do so) unless his favourite football team have a match whereas Ann reliably does all her homework (when obligated to do so) even if her favourite football have a match. But neither finishes all their homework when the Eurovision finals are on. Neither is maximally trustworthy with respect to doing their homework. Nevertheless, Ann is more trustworthy than George at doing all her homework because Ann is *closer* to being maximally trustworthy.

Someone can be trustworthy without being maximally trustworthy. That’s because ‘trustworthy,’ like ‘tall,’ is a context-sensitive term. Whether one is outright trustworthy to φ (as assessed in a given context) depends on whether one is trustworthy *enough*:

**Attributions of Outright Trustworthiness to φ**

‘*S* is trustworthy to φ’ is true in context *c* if and only if *S* approximates maximal trustworthiness to φ closely enough to surpass a threshold on degrees of trustworthiness determined by *c*. (Kelp & Simion 2023)

It shouldn’t be surprising, on reflection, that trustworthiness works this way. Someone might be trustworthy to watch your kids in one context (when you’re in the room next door), but not in another (when you’re out of the country for a week).

This account of trustworthiness is defended in more detail elsewhere.[[8]](#footnote-8) But here are a few quick reasons we like it. First, as already noted, it nicely explains why and how trust and reliability come apart. We can only be trusted to do things that fulfill obligations. That’s why I don’t *trust* my colleague to bring a packed lunch even though they are perfectly reliable at doing so: I merely *expect* them to.

Second, our account is like Hawley’s in that it doesn’t refer to anything about the inner life, motivations, or character of the relevant agent. This de-anthropocentrized approach rightly leaves open the possibility of AI trustworthiness, and it also avoids the troublesome cases of trustworthy but motivationally atypical agents of the sort discussed in §3.

But third, unlike Hawley’s view, our account is not negative. People (or AIs) can be untrustworthy with respect to φ-ing even if they fail to actively make a commitment to φ. That’s because we are subject to some obligations (e.g., in the case of humans, moral norms) whether we commit to them or not. Again, this seems right. I don’t get to be ‘off the hook’ for being untrustworthy with respect to not pushing people down the stairs just because I fail to *commit* to not pushing people down the stairs.

But wait—does our account really predict that AI can be (non-trivially) trustworthy? After all, our view says that AI can only manifest trustworthiness to φ when they have an *obligation* to φ. But are AI programs the sorts of things that can be subject to obligations?

Two things about this: first, recall that we use obligation in a technical, non-normatively-thick sense, as per standard deontinc logic, to simply map on to *ought*:[[9]](#footnote-9) being under an obligation to φ, then, does not imply, in our sense, anything more normatively ‘oompmhy’ than simply being the subject of an ought to φ.

Second: Are AIs subject to oughts? Yes they are. Artefacts, just like traits and activities, can be governed my norms sourced in their *functions*. There are, for instance, properly and improperly functioning hearts, democracies, and washing machines. Just so, there are properly and improperly functioning AI systems.

There are two kinds of functions that are relevant for us here— the *design* function (the function intended by the designer) and the *etiological* function (the function that explains the thing’s continued existence, acquired through a ‘feedback loop’ of continued success). AI systems can have both kinds of functions. It’s perhaps more obvious that AI systems can have *design* functions. AI systems tend to be recent creations. And unlike, e.g., the inventor of the wheel, we can often identify the person or group who first designed them. But it’s possible for artefacts to develop etiological functions that differ from their design functions: Suppose a programmer created an AI system to do X, but it turns out to do Y very well, and that’s why people keep using the relevant AI. When this happens, the etiological function (typically) overrides the normative significance of the design function.

So AI systems *are* subject to oughts A diagnostic cancer AI *ought* to identify cancer—that’s what its function is! And so, our attempt to de-anthropocize an account of trust that is applicable to AI has been successful. But we can clarify things by tweaking the definitions to make it explicit that the relevant kind of obligations are function-generated:[[10]](#footnote-10)

**Maximal Trustworthiness for AI** For all x where x is an AI, x is maximally trustworthy with regard to φ-ing if and only if x has a maximally strong disposition to meet its functional norms-sourced obligations to φ.

**Attributions of Outright Trustworthiness to φ for AI** For all x where x is an AI, ‘x is trustworthy to φ’ is true in context *c* if and only if x approximates maximal trustworthiness to φ closely enough to surpass a threshold on degrees of trustworthiness determined by *c*.

**Rational Trust** **in AI** For all x where x is an AI, S is rational to trust x to φ only if S is justified to believe that x is (outright) trustworthy with respect to φ-ing.

*§5 How to Trust AI*

Much of our work here has been theoretical. We’ve given an account of what trustworthiness for AI is and when it is rational to trust an AI. But part of what motivated this discussion was a practical problem. People often *don’t* trust AI, even when it’s very reliable. Does our theoretical work yield any practical hints for how we might properly increase trust in AI?

Before answering, we want to emphasize that it’s *not* the conclusion of our paper that people should trust AI more readily than they currently do. People should trust a particular AI *when they are justified to believe that it is trustworthy*. And while we’re optimistic about possibility of trustworthy AI, we also aren’t out to blame the layperson (like, perhaps, Lucie) who doesn’t immediately trust a particular AI system. Whether Lucie should trust SmartCredit depends on whether Smartcredit is justifiably believed by Lucie to be trustworthy.

Nevertheless, we don’t want people to *needlessly* fail to trust AI. When AI is trustworthy, does our account say anything about how to increase such rational trust?

Our first result is a negative one: we’re skeptical that explainability is the main culprit. As we showed in §2, there are plenty of ordinary cases of rational trust where explanation and even explainability are lacking from the perspective of the trustor. This doesn’t mean we think that explainability will never make it easier for someone to trust AI. We think the movement to make AI more understandable is a good thing: after all, understanding why an AI has delivered a particular result may be evidence for its trustworthiness, and thus feed into one’s justification to trust it. Furthermore, in some cases, offering an explanation may just be a context-relevant ought governing the AI at stake: plausibly, indeed, that is exactly the case of SmartCredit: a tool as SmartCredit shoulders an obligation to its users to explain why the application was unsuccessful, such that people live Lucie can come to know what they need to do to improve their credit score. What the depth of this explanation required is going ot be, of course, is going to vary with context, since the contextual need for understanding will vary, as will the stakes for the user. But we think explainability has been given an outsized role to play in discussions of trust and AI: often, such explanations will neither be sufficient nor necessary to render one justified to trust an AI.

According to our view, the (trustworthiness-generating) norms that AI are subject to are *functional* norms, generated either by the design plan or by the AI system’s etiological function. This is different from the norms that typically govern our interactions with fellow people (e.g., they are different from *moral* norms or *social* norms).[[11]](#footnote-11) Whereas we all know (to some degree, in a broad range of cases) the norms that govern e.g. communicative interactions among humans (Don’t lie! Assert only if you know!), we don’t always know the function of an artefact just by looking at it. One upshot of our picture is that, in order to increase rational trust in AI, the functions of artefacts should be made suitably obvious.

That raises the question: how do people normally acquire rational trust in artefacts? We suspect the main two avenues are mapping on to the ways in which people acquire justified beliefs in trustworthiness in general: via induction and testimony. The plausible equivalents for artefacts are *exposure* and *instruction*: For the latter, someone can rationally trust AI on the basis that they are *told* the relevant AI is trustworthy. This can be an important tool in *spreading* or *transmitting* trust in AI, but in some ways, it only pushes the problem back a step. How do we get testifiers to trust AI in the first place?

Exposure to AI means giving people the opportunity to witness (or, better, interact with) AI in environments where (trustworthy) AI can consistently manifest their obligation-meeting dispositions, and thus acquire justification via induction: People often learn about the disposition of an artefact to fulfil its function by seeing it at work (or, if possible, by trying it out for themselves). How do you learn that you can trust a kitchen knife to cut through food? By seeing it done, or using the knife oneself, again and again. Our suggestion is that this applies to AI artefacts as much as it does to kitchen artefacts. Give AI opportunities to display their trustworthiness in a wide-range of low-stake situations to allow observers (or users) to form a stable inductive basis.

Explanations can sometimes help. But we should expect that people will rationally trust AI, like most artefacts, when they understand what its function is, and when they see it manifest its disposition to fulfil its function: when they have seen it work consistently before. Understanding why the AI does what it does or says what it says needn’t come into it. That may seem intellectually lightweight, but given how difficult AI processes are to understand, it is also good news.

**References**

Baier, Annette (1986). ‘Trust and Antitrust,’ *Ethics* 96(2): 231–60.

Potter, Nancy N. (2002). *How Can I be Trusted?: A Virtue Theory of Trustworthiness*, Rowman & Littlefield.

Brandom, Robert (1983). ‘Asserting,’ *Noûs* 17 (4): 637–50.

Cappellen, Herman & Josh Dever (2021). *Making AI Intelligible: Philosophical Foundations*, Oxford University Press.

Carter, Adam J. (forthcoming). ‘Trust and Trustworthiness,’ *Philosophy and Phenomenological Research*.

Carter, Adam J. (2023). ‘Simion and Kelp on Trustworthy AI,’ *Asian Journal of Philosophy* 2:18.

Faulkner, Paul (2014). ‘The Practical Rationality of Trust,’ *Sythnese* 191: 1975–1989.

Harmon, S. A., Sanford, T. H., Brown, G. T., Yang, C., Mehralivand, S., Jacob, J. M., Valera, V. A., et al. (2020). Multiresolution application of artificial intelligence in digital pathology for prediction of positive lymph nodes from primary tumors in bladder cancer. JCO Clinical Cancer Informatics, 4, 367–382. https:// doi. org/ 10. 1200/ CCI. 19. 00155

Hawley, Katherine (2020). *How to be Trustworthy*, Oxford University Press.

Kelp, Christoph & Mona Simion (forthcoming). ‘What is Trustworthiness?’ *Noûs*.

Ribiero, Marco Tulio, Sameer Singh & Carlos Guestrin (2016). ‘”Why Should I Turst You?” Explaining the Predictions of Any Classifier,’ *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*: 1135–44.

Simion, Mona & Christoph Kelp (2023). ‘Trustworthy artificial intelligence,’ *Asian Journal of Philosophy* 2:8.

Simion, Mona & Christopher Willard-Kyle (forthcoming). ‘Trust, Trustworthiness, and Obligation,’ *Philosophical Psychology*.

Von Exchenbach, Warren J. (2021). ‘Transparence and the Black Box Problem: Why We Do Not Trust AI,’ *Philosohy and Technology* 34: 1607–1622.

1. REFERENCES on testimony. [↑](#footnote-ref-1)
2. REFERENCES on trust and monitoring. [↑](#footnote-ref-2)
3. Rational Trust merely states a necessary condition. That is because one may be worried about cases where S is justified in believing both (1) that R is trustworthy w.r.t. phi-ing, and (2) that, for whatever reason, R will not successfully phi in this particular case. e.g., maybe I know (it’s likely that) an unusual, external factor will prevent R from doing what they are disposed to do. `For a biconditional to hold, a prima facie clause will be needed. For this paper, though, the necessity direction is all we need. [↑](#footnote-ref-3)
4. REFERENCES trust vs reliability. [↑](#footnote-ref-4)
5. This is perhaps a bit too quick. Lucie could rationally trust SmartCredit if Lucie justifiably believed (perhaps because of misleading expert testimony) that SmartCredit *does* have goodwill toward her and so (given that SmartCredit is also reliable) *is* trustworthy. But we’ll ignore this scenario: we think it should be obvious enough to Lucie that SmartCredit doesn’t have anything like goodwill toward her. Or at least: Lucie isn’t rationally justified to believe that SmartCredit *does* have goodwill toward her. [↑](#footnote-ref-5)
6. But see footnote 1. [check number before submission… fn beginning ‘perhaps this is a bit too quick’] [↑](#footnote-ref-6)
7. The operator OB is taken as primitive in standard deontic logic, and standardly taken to map on to necessity. See e.g. REFERENCE. [↑](#footnote-ref-7)
8. So far, we’ve just been talking about attributions of being trustworthy to φ. But sometimes, we attribute trustworthiness *simpliciter*. We think trustworthiness to φ is the fundamental notion, which is why we’ve focused on it here. But see Kelp & Simion (2023) for our preferred treatment of trustworthiness *simpliciter*. [↑](#footnote-ref-8)
9. The operator OB is taken as primitive in standard deontic logic, and standardly taken to map on to necessity. See e.g. REFERENCE. [↑](#footnote-ref-9)
10. See Simion and Kelp 2023. [↑](#footnote-ref-10)
11. This isn’t to deny that it’s *possible* for very sophisticated AI to be subject to moral and social norms as well—recall C-3PO and Hal 9000. [↑](#footnote-ref-11)