Asian Journal of Philosophy (2023) 2:8
https://doi.org/10.1007/s44204-023-00063-5

ARTICLE SYMPOSIUM

Trustworthy artificial intelligence

Mona Simion[1] · **Christoph Kelp**[1]

Received: 24 October 2022 / Accepted: 26 February 2023

Abstract
This paper develops an account of trustworthy AI. Its central idea is that whether AIs are trustworthy is a matter of whether they live up to their function-based obligations. We argue that this account serves to advance the literature in a couple of important ways. First, it serves to provide a rationale for why a range of properties that are widely assumed in the scientific literature, as well as in policy, to be required of trustworthy AI, such as safety, justice, and explainability, are properties (often) instantiated by trustworthy AI. Second, we connect the discussion on trustworthy AI in policy, industry, and the sciences with the philosophical discussion of trustworthiness. We argue that extant accounts of trustworthiness in the philosophy literature cannot make proper sense of trustworthy AI and that our account compares favourably with its competitors on this front.

Keywords Artificial intelligence · Trustworthy · Trust · AI · Function

1 Introduction

What is trustworthy AI? Policy makers and AI developers around the world have invested millions to answer this question. The motivation for this interest lies with the thought that societies will only ever be able to achieve the full potential of AI if trust can be established in its development, deployment, and use (IHLEGAI 2019). If, for example, neither physicians nor patients trust an AI-based system's diagnoses or treatment recommendations, it is unlikely that either of them will follow the recommendations, even if the treatments may increase the patients' well-being (Thiebes et al., 2021). Similarly, if the general public doesn't trust autonomous cars, they will never replace common, manually steered cars (Condliffe, 2017). Rational trust,[1]

[1] See Carter and Simion (2020) for the nature and normativity of trust.

✉ Mona Simion
Mona.simion@glasgow.ac.uk

Christoph Kelp
Christoph.kelp@glasgow.ac.uk

[1] Cogito Epistemology Research Centre, University of Glasgow, Glasgow, UK

Published online: 13 March 2023

however, requires trustworthiness: Presumably, we should trust S to φ when they are *trustworthy* with respect to φ-ing. Indeed, paradigmatically good instances of trusting involve the trust of the truster *matching* the trustworthiness of the trustee (Carter Forthcoming, O'Neill, 2018). As such, if we are to expect users to trust a particular AI, we first need to understand what makes AIs trustworthy.

Several proposals in the form of 'lists' of features that make for trustworthy AIs can be found upon a simple Google search.[2] These proposals list features allegedly constituting AI trustworthiness without also aiming to offer an underlying, unificatory rationale. It is claimed that trustworthy AI is, for instance, safe, just, explainable, human-centred, beneficent, autonomous, robust, fair, transparent, non-discriminatory, promoting social and environmental wellbeing, non-malificent, etc.

As with all list-based theories, unsurprisingly, these trustworthy AI frameworks suffer from two main problems. The first problem has to do with explanatory adequacy: say that your preferred list of trustworthiness-making properties seems impeccably extensionally adequate — in that it seems to infallibly predict an AI is trustworthy when it is, and conversely, that it is not to be trusted when it is not. The question as to why your theory got it right remains unanswered: what is the trustworthy-making underlying property that delivers one particular list rather than another? Why should we think, for instance, that explainability belongs on the list, while transparency does not? Conversely, if we think that, on closer inspection, we should include transparency as well, why is that so? Short of having an answer to this question, we run the risk that our list merely covers paradigmatic cases of trustworthy AIs, rather than the nature thereof, In turn, if this is so, we run the risk of relying on untrustworthy non-paradigmatic AIs and, conversely, of not trusting trustworthy non-paradigmatic incarnations thereof.

The second problem has to do with the distinction, well-researched in philosophy but hardly ever mentioned in AI research and practice, between trustworthiness and mere reliability. Reliance is ubiquitous: You rely on the weather not to suddenly drop by 20 degrees, leaving you shivering; you rely on your colleague at work to help you with your jammed printer, because they're just better at this stuff; you rely on the shop at the corner to still be there tomorrow when you need to buy milk. Trust, the thought goes, is a more precious and less ubiquitous commodity. For most philosophers, trust involves reliance "plus some extra factor" (Hawley, 2014: 5.). The question as to what this extra factor might be has generated impressive amounts of literature in the ethics and epistemology of trust (see e.g. Carter and Simion for an overview). In contrast, this distinction has been ignored in AI research. If trust is

[2] Several such list-based frameworks have been developed and published by researchers, industry, and policymakers in the recent past. For a comprehensive overview, see Hagendorff (2020) and for particular proposals see, e.g., https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.htmlhttps://www.ibm.com/watson/trustworthy-ai, https://www.ericsson.com/en/blog/2020/12/trustworthy-aihttps://www.microsoft.com/en-us/research/project/trustworthy-ai/ The Assessment List for Trustworthy Artificial Intelligence (ALTAI), for instance —written by the High-Level Expert Group on AI set up by the European Commission—consists of seven requirements: Human Agency and Oversight; Technical Robustness and Safety; Privacy and Data Governance; Transparency; Diversity, Non-discrimination and Fairness; Societal and Environmental Well-being; Accountability..

not mere reliance, though, neither is trustworthiness mere reliability: more needs to be the case. However, it's not clear that the features proposed in list-based frameworks of trustworthy AI will be able to account for trustworthy AI proper, rather than merely reliable AI.

This paper aims to innovate on both fronts: we start by looking into the literature on trust and reliance, and we argue that the conditions put forth to distinguish one from the other are too anthropocentric to do the job of accounting for trustworthiness in the case of AI (Sect. 2). Second, we propose an account of trustworthy AI that does the job (Sect. 3). Further, we argue that our account provides a unified, independent theoretical rationale for generating objective list frameworks for different AIs in different contexts (Sect. 4).

2 Trust and reliance

2.1 The psychological view

Classic accounts of what differentiates trustworthiness from mere reliability explain this distinction in psychological terms: trustworthiness, on these views, is reliability sourced in a particular good-making psychological trait. On Annette Baier's (e.g. 1986) *goodwill*-based account (see also Jones, 1996; Cogley, 2012), for instance, while the reliable person merely fulfils their commitments reliably, the trustworthy person fulfils their commitments reliably *in virtue of* their goodwill towards the trustor. This view, according to Baier, makes good sense of the intuition that trust differs from mere reliance in that trust, but not mere reliance, can be betrayed.

An alternative to the goodwill account that also attempts to explain the difference between trustworthiness and mere reliability in psychological terms is Nancy Potter's view (2002). According to Potter, trustworthiness is a virtue, i.e. a disposition to respond to trust in appropriate ways, given "who one is in relation" to and given other virtues that one possesses or ought to possess (e.g. justice, compassion) (2002: 25). A trustworthy person is "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." Potter's view purports to account for the intuition that mere reliability is not enough for trustworthiness by imposing a good character condition on trustworthiness.

The worry about these accounts is that they are not easily generalisable to artificial intelligence because the psychological assumptions underlying them are too anthropocentric: do AIs have something that is recognizable as goodwill? Can AIs host character virtues? Or, to put it more precisely, is it correct to think that AI capacity for trustworthiness co-varies with their capacity for hosting a will or character virtues? And more generally, should we, upon finding out that a particular artefact that we thought trustworthy is indeed incapable of hosting these psychological traits, revise our trust attitudes towards said artefact? Here is Simon Blackburn putting this point succinctly:

> We are often content to trust without knowing much about the psychology of the one-trusted, supposing merely that they have […] traits sufficient to get the job done" (Blackburn, 1998).

But, of course, there is excellent reason to think that what goes for trust holds, *mutatis mutandis*, for trustworthiness. Otherwise, in the kind of case Blackburn describes here, our trust would have to be misplaced in an important sense. And that doesn't appear to be the case. By the same token, there is reason to think that these accounts are too anthropocentric. In fact, the case of trustworthy AI serves to make this point particularly forcefully. Think of paradigm cases of AI such as manufacturing robots and self-driving cars. Questions about trustworthiness clearly have answers that they can go both ways here. Crucially, they do so even if it turns out that manufacturing robots or self-driving cars don't have goodwill or good character, simply because they do not have the right kind of psychology.

2.2 The Commitment view

Katherine Hawley's (2019) new account of trustworthiness departs abruptly from the tradition of taking trustworthiness to be psychologically demanding. According to Hawley, trustworthiness is simply a matter of avoiding unfulfilled commitments, which requires both caution in incurring new commitments and diligence in fulfilling existing commitments. Hawley's is a negative account of trustworthiness, which means that one can be trustworthy whilst avoiding commitments as far as possible. A trustworthy person, on Hawley's view, must not allow her commitments to outstrip her competence.

Crucially, on this view, one can be trustworthy regardless of one's motives for fulfilling one's commitments, and regardless of whether one is displaying virtues or not in the process: no particular psychological basis for reliability is necessary. At the same time, on this view, trustworthiness differs from mere reliability. This is because Hawley accounts for trustworthiness in terms of commitments and people can be reliable *phi*-ers even though the question of committing to *phi*-ing never arises: I can reliably buy my coffee at the same shop every morning, without ever having committed to do so. If so, on Hawley's view, I'm a reliable buyer, but I'm not the proper target of trustworthiness attributions.

Hawley's account is less psychologically demanding than its predecessors, and, in this, it might be thought to generalise more easily to Artificial Intelligence. Of course, AIs don't strictly speaking commit to doing things. That said, one might think that they incur commitments nevertheless in virtue of their design: a cancer diagnostic AI, for instance is "committed" to working in particular ways to the aim of identifying tumours because that's how it's supposed to work, by design. If so, designer's intentions will spell out the set of commitments a particular AI undertakes, and which, in turn, they need to fulfil in order to be trustworthy.

One problem for an account of trustworthy AI along these lines is the problem of bad design. This problem pertains to commitments, as it were, that the AI in question should have taken on but didn't — i.e. things that are supposed to be part of

its design plan, but are not. If my cancer diagnostic AI has a design flaw that renders it incapable to recognise the simplest, garden-variety tumours that does not seem to absolve it from its being supposed to be able to recognise garden-variety tumours. However, on the version of the commitment view that we are now considering, according to which AI commitments are grounded in design plans, my cancer diagnostic AI will come out just as trustworthy as any other diagnostic AI that does have the capacity to recognise garden-variety tumours: after all, both AIs work as their design says they are supposed to work, and thereby fulfil their design-sourced commitments.

In a similar vein, consider also bad commitments: say that not only does my diagnostic AI lack a design plan that enables it to recognise simple tumours, but its design explicitly features a line that will cause it to crash whenever presented with a garden-variety tumour. Again, intuitively, my diagnostic AI is not equally trustworthy to your diagnostic AI that lacks this annoying feature.

In response to the problem of commitments one should have taken on, Hawley appeals to commitments we take on indirectly, by entering into particular roles and relationships. More specifically, Hawley argues that we may and often do take on meta-commitments — commitments to incur future commitments, by entering into particular relationships and inhabiting certain social roles. One can maybe try to generalise this reply to the case of Artificial Intelligence as follows: my diagnostic AI undertakes certain meta-commitments in virtue of being a member of its kind. In virtue of these meta-commitments, the design plan of my diagnostic AI should feature particular stipulations — including that it should recognise simple tumours.

The problem now, however, is how to trace the source of these meta-commitments. Since they seem to be triggered by AI being a member of the kind 'cancer diagnostic AI', and since we started off by taking AI commitments to be sourced in design features, it should presumably be the case that, in turn, these meta-commitments will be sourced in the design plan of the kind that my AI pertains to: the kind 'cancer diagnostic AI'. The problem now, however, is that there is no such thing as a recognisable design plan pertaining to the kind: particular artefacts and particular types of artefacts come with design plans, but general kinds thereof do not. If so, the commitment account leaves commitments that should have been in place in virtue of belonging to a certain kind unexplained after all.

Finally, and relatedly, trustworthy AI serves to put some pressure on Hawley's distinctively negative account of trustworthiness. Recall that according to Hawley, one way to be trustworthy is to avoid taking on commitments altogether. Crucially, once we start thinking about cases of trustworthy AI, it is just not clear that this is a plausible view to have. Consider a self-driving car that simply will not take on the commitment to take you to City Hall once you have told it to, perhaps by design. In fact, it will not take on any commitments at all. Any such item would not be an instance of trustworthy AI. Or, at the very least, it would be less trustworthy than a self-driving car that takes on the relevant commitments but doesn't always live up to them. However, that's not something that a negative account of trustworthiness can make sense of. In this way, there

is further reason to think that Hawley's account will struggle to generalise in a satisfactory manner to trustworthy AI.[3]

3 Trustworthiness as disposition to fulfil one's obligations

In previous work (Kelp and Simion Forthcoming), we have developed an account of trustworthiness as a disposition to fulfil one's obligations. More precisely, on our account:

Outright trustworthiness attribution "*S* is trustworthy" is true in context *c* if and only if *S* approximates maximal trustworthiness to *phi* for all *phi* closely enough to surpass a threshold on degrees of trustworthiness determined by *c*.

In turn, we define maximal trustworthiness to phi in dispositional terms:

Maximal trustworthiness to *Phi* One is maximally trustworthy with regard to *phi*-ing if and only if one has a maximally strong disposition to fulfil one's obligations to *phi*.

Correspondingly, degrees of trustworthiness are defined against the distance from maximal trustworthiness to phi:

Degrees of trustworthiness to *Phi* The degree of trustworthiness to *phi* of S is a function of the distance from maximal trustworthiness to *phi*: the closer one approximates maximal trustworthiness to *phi*, the higher one's degree of trustworthiness to *phi*.

According to this view, then, when Ann says "George is trustworthy," what she is saying is, very roughly, that he is trustworthy enough by the standards operative in the conversational context, which, in turn, means that he has a strong enough disposition to fulfil his contextually relevant obligations.

How does the contextual threshold get set? On our view, degrees of trustworthiness *simpliciter* can be measured along at least two dimensions, i.e. breadth and depth: we can measure on how many *phi*-s one is trustworthy on, and how well one approximates maximal trustworthiness to *phi* for each *phi* in question. In turn, both of these dimensions will influence how the threshold is set at a given context.

The depth dimension of the contextual threshold for trustworthiness simpliciter, which is given by the contextually appropriate degree of trustworthiness to *phi* for a particular *phi,* concerns the contextually appropriate strength of one's disposition to meet one's obligations to *phi*.

[3] Ryan (2020) makes a case against the application of In contrast to the view defended here, however, Ryan thinks that there is no account of trust (and trustworthiness), properly so-called, that we can apply to AI, and that we should speak merely in terms of reliance in relation to AI.

In turn, the threshold for breadth is to be understood in terms of a contextually determined set (or sets) of *phi*; that is, the set (or sets) of actions that are salient at the conversational context where the attribution is made. We want to allow that one can be trustworthy in different ways, i.e. by approximating maximal trustworthiness via different routes, as it were. To do achieve this, we may countenance sets of sets of phi-ings that are made salient such that one is close enough to maximal trustworthiness *simpliciter* just in case one is sufficiently trustworthy to *phi* for all *phi* in some such set of sets. We distinguish between two varieties of ascriptions of trustworthiness *simpliciter*, viz. predicative and attributive:

> Predicative ascriptions: George is trustworthy. Ann is trustworthy.
> Attributive ascriptions: George is a trustworthy babysitter. Ann in a trustworthy physician.[4]

How is the threshold for breadth determined in cases of attributive ascriptions of trustworthiness? As a first step, the relevant *phi*-s that are picked up by the conversational context are the *phi*-s pertaining to the domain of attribution, i.e. babysitting in the case of George (watching the kids, feeding them, etc.), and being a physician in the case of Ann (diagnosing health conditions, prescribing medication, etc.). That said, practical interests of the attributors will also make a difference: different phi-s will be picked out at the context depending on, e.g. whether the attributors are Ann's patients, or the stake holders of the hospital.

The way the threshold for attributive ascriptions of trustworthiness simpliciter is set at a context, then, is as follows: first, context delivers the *phi*-s that are relevant for the breadth dimension of the threshold against which the trustworthiness ascription is to be evaluated as true or false in accordance with the phi-s pertaining to the domain of attribution and practical interests. Second, after the set (or sets) of *phi*-s is established, the threshold for depth across the *phi*-s in question gets set: that is, in this second step, context determines how strong the disposition to fulfil one's obligations to *phi* for the relevant *phi*-s needs to be.

What about predicative ascriptions of trustworthiness simpliciter? In these cases, on our account, the threshold for breadth is set at a context in the following way: first, context delivers the *phi*-s that are relevant for the breadth dimension of the threshold against which the trustworthiness ascription is to be evaluated as true or false in accordance with the practical interests of the attributors. Second, after the set (or sets) of *phi*-s is established, the threshold for depth across the *phi*-s in question gets set: that is, in this second step, context determines how strong the disposition to fulfil one's obligations to *phi* for the relevant *phi*-s needs to be.

[4] For more on predicative vs. attributive ascriptions see (Geach 1956).

4 Trustworthy AI

In what follows, we will show how our account generalises to trustworthiness in Artificial Intelligence. To do this, we would like to begin via talking about how AIs acquire obligations through function acquisition.

4.1 Artefacts' functions and norms

Traits, activities, and artefacts alike are governed by norms sourced in their functions: there are, for instance, malfunctioning and properly functioning hearts, malfunctioning and properly functioning social practices, and malfunctioning and properly functioning washing machines.

Some functions and their corresponding norms are acquired etiologically (henceforth etiological functions, or e-functions[5]) via a history of success and positive feedback. My heart, for instance, has acquired the function of pumping blood in my circulatory system via successfully doing so in my ancestors, which benefitted them, and which, in turn, contributes to the explanation of the continuous existence of hearts. My heart is properly said to be malfunctioning — i.e. functioning badly, in breach of the norms governing hearts — when it fails to work in a way which, in normal conditions, leads to its reliably enough fulfilling its function of pumping blood in my circulatory system. My heart is malfunctioning, for instance, when it fails to beat at a normal rate. The content of 'proper functioning' is dictated by the way of working that reliably enough delivers function fulfilment in normal conditions.

Similarly, e.g. the social practice of telling things to each other is malfunctioning when done non-knowledgeably: plausibly, this practice has generated knowledge in hearers in the past, which benefitted them, and which explains the continuous existence of this practice. Further on, in normal conditions, assertions need to be knowledgeable in order to fulfil their function of reliably generating knowledge in hearers. Again, the content of 'proper functioning' is dictated by the way of working that reliably enough delivers function fulfilment in normal conditions.

Artefacts are, first and foremost, bearers of design functions. Design functions differ from etiological functions in that they are sourced in the intentions of the designers rather than in a history of success. Design functions, note, need not imply success: the Museum of Failed Inventions is filled with unsuccessful bearers of design functions. Correspondingly, what it is for artefacts to be properly functioning — i.e. functioning by the norm, or in the way in which they are supposed to — will be determined by their design as well, rather than sourced in reliable success.

That being said, often, artefacts also acquire etiological functions on top of their design functions. My knife, for instance, has the design function to cut because that was, plausibly, the intention of its designer. At the same time, my

[5] For more on etiological accounts of functions see e.g. (Millikan 1984, Neander 1991a, 1991b, Griffith 1993, Peter Godfrey-Smith 1993, 1994, and David Buller 1998. For alternatives, see e.g. (Bigelow and Pargetter 1987, Kitcher 1994 Denis Walsh and Ariew 1996).

knife also has an etiological function to cut: that is because tokens of its type have cut in the past, which was beneficial to my ancestors, and which contributes to the explanation of the continuous existence of knives.

When artefacts acquire etiological functions on top of their design functions, they thereby acquire a new set of norms governing their functioning, sourced in their etiological functions. Design-wise, my knife is properly functioning (henceforth properly d-functioning) insofar as it's working in the way in which its designer intended it to work. Etiologically, my knife is properly functioning (henceforth properly e-functioning) insofar as it works in a way that reliably leads to cutting in normal conditions.

In the happy cases, the two ways to function properly often go hand in hand: presumably, whoever designed my knife intended it to cut precisely in the way in which it reliably does in normal conditions. Proper d-functioning and proper e-functioning can come apart, however. This will happen, most often, in cases in which artefacts are designed to work in non-reliably function fulfilling ways.

In cases in which d-norms and e-norms of proper functioning come apart, the latter override the former at the context, and tend to be appropriated into the design of future generations. That is because reliable function fulfilment comes first in functional items, and proper e-functioning, but not proper d-functioning, delivers it. That's why what we usually see in cases of divergence is that norms governing proper e-functioning tend to be incorporated in design plans of future generations of tokens of the type: if we discover that there are more reliable ways for the artefact in question to fulfil its function, design will follow suit.

4.2 Trustworthy AI and proper function

On our account, trustworthy AI is AI that meets the norms associated with its proper functioning. In turn, the latter can be soured in its design functions, its etiological functions, or both.

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

In turn, we define maximal trustworthiness to phi for AI in dispositional terms:

AI maximal trustworthiness to *Phi* For all x where x is an AI, x is maximally trustworthy with regard to *phi*-ing if and only if x has a maximally strong disposition to meet its functional norms-sourced obligations to *phi*.

Once more, the functional norms at stake can be d-functional norms or e-functional norms or both, where the latter override the former in cases of conflict.

Correspondingly, degrees of trustworthiness will be defined against the distance from maximal trustworthiness to phi:

AI degrees of trustworthiness to *Phi* For all x where x is an AI, the degree of trustworthiness to *phi* of x is a function of the distance from maximal trustworthiness to *phi*: the closer x approximates maximal trustworthiness to *phi*, the higher x's degree of trustworthiness to *phi*.

It is easy to see that this account delivers the non-anthropocentric view of trustworthiness we were looking for in an account of trustworthy AI: first and foremost, it does not require any highbrow stipulations concerning AI psychology: should we discover that it is implausible to think that AIs have anything resembling a will, or character traits, the possibility of AI trustworthiness will survive this discovery.

Furthermore, the account also explains why sometimes AIs fail to be trustworthy even though they meet their design impeccably: recall the case of my diagnostic AI that could not recognise simple tumours as a matter of design. On this account, although meeting all of its d-functionally sourced obligations, my diagnostic AI fails to meet its e-functionally sourced obligations: it is malfunctioning etiologically, in that the recognising of tumours by the type of artifact it belongs to contributes to the explanation of the continuous existence of cancer diagnostic AIs. In this case, in which d-norms and e-norms of proper functioning come apart, the latter override the former at the context, because reliable function fulfilment comes first in functional items, and proper e-functioning, but not proper d-functioning, delivers it.

Note, also, that the account also deals well with trustworthiness comparisons: recall that we said that even though your perfectly normal diagnostic AI and mine worked impeccably by their respective design plans, intuitively, yours was more trustworthy qua diagnostic AI than mine, for being able to recognise garden-variety tumours. The view delivers the result straightforwardly, in that your AI approximates maximal trustworthiness more closely than mine, via meeting more of its obligations to phi than mine — to wit, via meeting etiologically sourced obligations that mine fails to meet.

Last but not least: recall that we started this paper by noting that policy makers and AI developers around the world have put forth several 'list-based' proposals for what constitutes AI trustworthiness. It is claimed that trustworthy AI is, for instance, safe, just, explainable, human-centred, beneficent, autonomous, robust, fair, etc. We have said that, as with all list-based theories, unsurprisingly, these trustworthy AI frameworks suffer from lack of explanatory adequacy i.e. from lack of an underlying rationale for including a particular property on these lists. What is the trustworthy-making underlying property that delivers one particular objective list rather than another?

Our view delivers an independently motivated answer to this question: trustworthy-making properties for AIs are properties that map on to their having a disposition to fulfil their functionally sourced obligations. AIs should be safe, just, human-centred, and beneficent, for instance, insofar as this amounts to their being etiologically properly functioning i.e. functioning in a way that contributes to the explanation of their continuous existence. At the same time, it will be a matter of context and of the type of AI at

stake whether particular properties are trustworthy-makers: does the type of AI in question carry a functionally-sourced obligation to be e.g. explainable? Plausibly, for some AIs — such as e.g. creditworthiness scoring AIs — this obligation will be present and salient: people need to know why their mortgage was rejected, in order to figure out how to improve their credit score in the future. In contrast, many of my diagnostic AIs workings need not be particularly explainable to patients for it to qualify as trustworthy: after all, explanation has the function of generating understanding. In the case of complicated medical diagnostics, however, little to no understanding is available to laymen, no matter how much explanation is on offer.

Once more, what explains this variation is the particular way of functioning which triggered the continuous existence of the relevant AI.

5 Conclusion

We have argued that trustworthy artificial intelligence is artificial intelligence that fulfils its functionally sourced obligations, where the latter can be either design-sourced, or etiologically determined via the benefit that the relevant AI brings, and which contributes to the explanation of its continuous existence. We have also offered a contextualist semantics for AI trustworthiness attributions, together with a threshold-setting recipe for both attributive and predicative ascriptions.

References

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

Beaver, D. and Geurts, B. 2014. Presupposition. In Zalta, E. ed. *The Stanford Encyclopedia of Philosophy.* URL = <https://plato.stanford.edu/archives/win2014/entries/presupposition/>.

Bigelow, J., & Pargetter, R. (1987). Functions. *Journal of Philosophy, 86*, 181–196.

Bird, A. (1998). Dispositions and antidotes. *The Philosophical Quarterly, 48*, 227–234.

Blackburn, S. (1998). *Ruling passion: A Theory of practical reasoning*. Clarendon Press.

Buller, D. (1998). Etiological theories of function: A Geographical Survey. *Biology and Philosophy, 13*, 505–527.

Carter, J. A. (forthcoming). 'Trust and trustworthiness', *Philosophy and Phenomenological Research*.

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

Cogley, Z. (2012). Trust and the trickster problem. *Analytic Philosophy, 53*, 30–47.

Condliffe, J. (2017). A single autonomous Car has a huge impact on alleviating traffic. MIT technology review. Retrieved from https://www.technologyreview.com/s/607841/a-single-autonomous-car-has-a-huge-impact-on-alleviating-traffic/
Ericsson framework for trustworthy AI https://www.ericsson.com/en/blog/2020/12/trustworthy-ai
European Commission, Ethics Guidelines for Trustworthy AI https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html
Faulkner, P. (2007). A genealogy of trust. *Episteme, 4*(3), 305–321.
Geach, P. (1956). Good and Evil. *Analysis, 17*, 33–42.
Godfrey-Smith, P. (1994). A modern history theory of functions. *Noûs, 28*, 344–362.
Graham, P. (2012). Epistemic entitlement. *Noûs, 46*, 449–482.
Griffith, P. (1993). Functional analysis and proper functions. *British Journal for the Philosophy of Science, 44*, 409–422.
Hawley, K. (2014). Trust, distrust and commitment. *Noûs, 48*(1), 1– 20.
Hawley, K. (2019). *How to be trustworthy*. Oxford University Press.
Healey, R. (1991). *The philosophy of quantum mechanics: An interactive interpretation*. Cambridge University Press.
IBM Trustworthy AI Pramework (2021). https://www.ibm.com/watson/trustworthy-ai
IHEGAI (Independent High-Level Expert Group on Artificial Intelligence). (2019). *Ethics guidelines for trustworthy AI*. Brussels: European Commission Retrieved from https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419
Johnston, M. (1992). How to speak of the colors. *Philosophical Studies, 68*, 221–263.
Jones, K. (1996). Trust as an affective attitude. *Ethics, 107*, 4–25.
Kelp, C. and Simion, M. (Forthcoming). What Is Trustworthiness? *Nous*.
Kelp, C., & Simion, M. (2021). *Sharing knowledge: A functionalist account of assertion*. Cambridge University Press.
Kitcher, P. (1994). Function and design. *Midwest Studies in Philosophy, 18*, 379–397.
Lewis, D. (1997). Finkish dispositions. *The Philosophical Quarterly, 47*, 143–158.
McLeod, C. 2015. Trust. *The Stanford Encyclopedia of Philosophy.* In Zalta, E. ed., URL = <https://plato.stanford.edu/archives/fall2015/entries/trust/>.
McKitrick, J. (2003). A case for extrinsic dispositions. *Australasian Journal of Philosophy, 81*, 155–174.
Microsofst Trustworthy AI Project https://www.microsoft.com/en-us/research/project/trustworthy-ai/
Millikan, R. (1984). *Language, Thought, and other biological categories*. MIT Press.
Mumford, S. (1998). *Dispositions*. Oxford University Press.
Neander, K. (1991a). Functions as selected effects: The conceptual analyst's defence. *Philosophy of Science, 58*, 168–184.
Neander, K. (1991b). The teleological notion of 'function.' *Australasian Journal of Philosophy, 69*, 454–468.
O'Neill, O. (2018). Linking Trust to Trustworthiness. *International Journal of Philosophical Studies, 26*(2), 293–300.
Potter, N. N. (2002). *How can I be trusted?: A virtue theory of trustworthiness*. Rowman & Littlefield Publishers.
Ryan, M. (2020). In AI we trust: Ethics, artificial intelligence, and reliability. *Science and Engineering Ethics, 26*, 2749–2767.
Simion, M. (2021). *Shifty speech and independent thought: Epistemic Normativity in context*. Oxford University Press.
Sosa, E. (2015). *Judgment and agency*. Oxford University Press.
Suarez, M. (2007). Quantum propensities. *Studies in History and Philosophy of Modern Physics, 38*, 418–438.
Thiebes, S., Lins, S., & Sunyaev, A. (2021). Trustworthy artificial intelligence. *Electronic Markets, 31*, 447–464.
Walsh, D., & Ariew, A. (1996). A taxonomy of functions. *Canadian Journal of Philosophy, 26*, 493–514.