Asian Journal of Philosophy

Simion and Kelp on Trustworthy AI

--Manuscript Draft--

Manuscript Number:	
Full Title:	Simion and Kelp on Trustworthy AI
Article Type:	Book Symposium
Funding Information:	Arts and Humanities Research Council (AH/W008424/1) — Dr J Adam Carter
Abstract:	Simion and Kelp offer a prima facie very promising account of trustworthy AI. One benefit of the account is that it can straightforwardly explain trustworthiness in the case of cancer diagnostic AIs. What I query in this brief note is just how far beyond this specific case type – a case type which involves the acquisition by the AI of a representational etiological function – their account can be extended without overpredicting untrustworthiness.
Corresponding Author:	J Adam Carter University of Glasgow Glasgow, Glasgow UNITED KINGDOM
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	University of Glasgow
Corresponding Author's Secondary Institution:	
First Author:	J Adam Carter
First Author Secondary Information:	
Order of Authors:	J Adam Carter
Order of Authors Secondary Information:	
Author Comments:	The submitted critical piece is for the article symposium on *Trustworthy AI* by Mona Simion and Chris Kelp". There was no option under "book symposium" to click: *Article Symposium: Trustworthy AI*.

Simion and Kelp on Trustworthy AI

J Adam Carter
Reader in Philosophy
University of Glasgow
67-69 Oakfield Avenue
Glasgow, Scotland
United Kingdom

adam.carter@glasgow.ac.uk

Orcid ID: 0000-0002-1222-8331

Abstract:

Simion and Kelp offer a prima facie very promising account of trustworthy AI. One benefit of the account is that it can straightforwardly explain trustworthiness in the case of cancer diagnostic AIs. What I query in this brief note is just how far beyond this specific case type – a case type which involves the acquisition by the AI of a representational etiological function – their account can be extended without overpredicting untrustworthiness.

Keywords:
Trustworthy AI; trustworthiness; etiological functionalism; artificial intelligence

Data availability N/A

Code availability: N/A

Conflicts of interest: no conflict of interest to report.

Funding: Thanks to the AHRC *Digital Knowledge* (AH/W008424/1) project for supporting this work.

Simion and Kelp on Trustworthy AI

Abstract

Simion and Kelp offer a prima facie very promising account of trustworthy AI. One benefit of the account is that it can straightforwardly explain trustworthiness in the case of cancer diagnostic AIs. What I query in this brief note is just how far beyond this specific case type – a case type which involves the acquisition by the AI of a representational etiological function – their account can be extended without overpredicting untrustworthiness.

1. Introduction

Increasingly, the question of whether – and if so under what conditions – artificial intelligence (AI) can be 'trustworthy' (as opposed to merely reliable or unreliable) is being debated by researchers across various disciplines with a stake in the matter, from computer science and medicine to psychology and politics.[1]

Given that the nature and norms of trustworthiness itself have been of longstanding interest in philosophy[2], philosophers of trust are well situated to help make progress on this question. In their paper "Trustworthy Artificial Intelligence" (2023), Simion and Kelp (hereafter, S&K) aim to do just this.

I think they largely succeed. That said, in this short note I am going to quibble with a few details. In short, I worry that their reliance on function-generated obligations in their account of trustworthy AI helps their proposal get exactly the right result in certain central AI cases, such as cancer diagnostic AIs, but at the potential cost of overpredicting untrustworthiness across a range of other AIs.

[1] For some recent reviews, see, e.g., Kaur et al. (2022); and Kaur, Uslu, and Durresi (2020).
[2] See, e.g., Hardin (1996); Jones (2012); Frost-Arnold (2014); O'Neill (2018); Simion and Kelp (2022); Carter (2022). See also Carter and Simion (2020) for a review.

Here's the plan for the paper. In §2 I'll provide a brief overview of S&K's account of trustworthy AI, emphasising the core desiderata they take themselves to have met. In §3, I'll then raise some potential worries, and discuss and critique some lines of reply.

2. S&K's line of argument

A natural strategy for giving an account of trustworthy AI will be a kind of 'application' strategy: (i) give a compelling account of trustworthiness simpliciter and then (ii) apply it to AI, and make explicit what follows, illuminating trustworthy AI in the process.

But, as S&K note, there is a problem that faces many extant accounts of trustworthiness that might try to opt for that strategy. The problem is this: many accounts of trustworthiness are such that the psychological assumptions underlying them (e.g., that being trustworthy involves something like a good will or virtue) are simply too anthropocentric.

As S&K ask:

> 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? (p. 4).

The situation seems to be this: an account of trustworthiness with strongly anthropocentric psychological features 'baked in' will either *not* be generalisable to AI (if AI lacks good will, virtue, etc.), or it will be generalisable only by those willing to embrace further strong positions about AI.

Ceteris paribus, a more 'generalisable' account of trustworthiness, when it comes to an application to AI specifically, will be a less anthropocentric one that could sidestep the above problem. One candidate such account they identify is Hawley's (2019) negative account of trustworthiness, on which trustworthiness is a matter of avoiding unfulfilled commitments.[3]

[3] Another recent account of trustworthiness that is arguably not too anthropocentric that it would be difficult to generalise over to AI is the performance-theoretic account of trustworthiness. See here, Carter (2022).

S&K have argued elsewhere[4] at length for a different -- and similarly non overtly anthropocentric -- account of trustworthiness, which they take to have advantages (I won't summarise these here) over Hawley's: on S&K's preferred account, trustworthiness is understood as a disposition to fulfil one's *obligations*.

What is prima facie attractive about an obligation-centric account of trustworthiness, for the purpose of generalising that account to trustworthy AI, is that (i) artifacts can have functions; and (ii) functions can generate obligations.

Let's look at the first point first. S&K distinguish between *design functions* (d-functions), sourced in the designer's intentions, and *etiological functions* (e-functions), sourced in a history of success, noting that artefacts can acquire both kinds of functions. S&K use the example of a knife to capture this point:

> My knife, for instance, has the design function to cut because that was, plausibly, the intention of its designer. At the same time, my 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 (p. 9).

While d-functions and e-functions (i.e., proper functioning) will often line up, these functions can come apart (e.g., when artifacts are designed to work in non-e-function-filling ways). When they don't line up, S&K maintain that e-functions generally override. As they put it:

> what we usually see in cases of divergence is that norms governing proper-functioning tend to be incorporated in

[4] See Kelp and Simion (2022).

> 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 (Ibid., p. 9).

So we have in view now S&K's thinking behind the idea that artifacts (of which AI is an instance) can acquire functions. What about the next component of the view: that functions can generate obligations?

The crux of the idea is that a species of obligation, function-generated obligation, is implicated by facts about what it is for something to fulfil its e-function. The heart has a purely e-function generated obligation to pump blood in normal conditions (the conditions under which pumping blood contributed to explanation of its continued existence). In maintaining this, on S&K's line, we aren't doing anything objectionably anthropocentric, any more than when we say a heart *should* (qua heart) pump blood. We can easily extend this kind of obligation talk over to artifacts, then: just as a heart is malfunctioning (and so not meeting its e-functionally sourced obligations) if it stops pumping blood, a diagnostic AI is malfunctioning (and not meeting its e-functionally sourced obligations) if it stops recognising simple tumours by their appearance, and miscategorises them.

Against this background, then, S&K define an AI's being *maximally* trustworthy at phi-ing as being a matter of having a "maximally strong disposition to meet its functional norms-sourced obligations to phi." The conditions for *outright* AI trustworthiness attributions can then be characterised in terms of maximal AI trustworthiness in the following way: an outright attribution of trustworthiness to an AI is true in a context c iff that AI approximates "maximal trustworthiness to phi"[5] closely enough to surpass a threshold on degrees of trustworthiness determined by c, where the closer x approximates maximal trustworthiness to phi, the higher x's degree of trustworthiness to phi.

3. Critical Discussion

[5] A fuller exposition of the details of these ideas are spelled out in S&K work on trustworthiness more generally, i.e., in Kelp and Simion (2022).

I suspect that a typical place one might begin to poke to look for a hole in the above account would be the very idea that a machine could have an obligation in the first place. Imagine this line or reply: "But S&K have complained that extant accounts of trustworthiness that rely on 'virtue' and 'good will' as psychologically demanding prerequisites for being trustworthy are too anthropocentric to be generalisable to AI. But isn't being a candidate for an 'obligation' equally psychologically demanding and thereby anthropocentric? If so, haven't they failed their own generalisability desiderata by their own lights?"

The above might look superficially like the right way to press S&K, but I think such a line would be uncharitable, so much so that it's not worth pursuing. First, we humans often have our own obligations to others *sourced* in facts about ourselves (substantive moral agreements we make, etc.) that are themselves predicated on our having a kind of psychology that we're not yet ready to attribute to even our most impressive AI.

But S&K's argument is compatible with all of this – viz., with granting that obligations often times for creatures like us arise out of features AI lack. What matters for their argument is just that AI are candidates for e-function generated obligations, and it looks like this is something we can deny only on pain of denying either that AI can have e-functions, or that e-functions can generate norms.[6] I think we should simply grant both of these – rather than incur what looks like an explanatory burden to deny either.

The right place to press them, I think, is on the *scope* of the generalisability of their account. Here it will be helpful to consider again the case of a cancer-diagnostic AI which they use for illustrative purposes. The etiological function that such cancer diagnostic AIs acquire (which aligns with their d-function) is going to be a purely *representational* function. Cancer diagnostic algorithms are updated during the AI's supervised learning process (i.e., as is standard in deep learning) against the metric of representational accuracy; the aim here is reliably *accurately* identifying (and not misidentifying) e.g., tumours from images, and thus to maximise representational accuracy via sensitivity and specificity in its classifications. The AI becomes valuable to the

[6] For some representative defences of this idea, see, e.g., Graham (2014), Simion (2019), Kelp (2018), Kelp and Simion (2021).

designer when and only when, and to the extent that, this is achieved.

To use but one example, take the case of bladder cancer diagnosis. It is difficult using standard human tools to reliably predict the metastatic potential of disease from the appearance of tumors. Digital pathology via deep learning AI is now more reliable than humans at this task, and so can predict disease with greater accuracy than through use of human tools alone (see Harmon et al. 2020). This predictive accuracy explains the continued use (and further accuracy-aimed calibration by the designers) of such diagnostic AIs.

There are other non-diagnostic AIs with representational functions as their e-functions. An example is FaceNet, which is optimised for accuracy in identifying faces from images (Schroff, Kalenichenko, and Philbin 2015; William et al. 2019).

AIs with purely representational e-functions, however, are – perhaps not surprisingly – an outlier in AI more broadly. Let's begin here by considering just a few examples of the latest deep learning AI from Google's DeepMind. AlphaCode, for instance, is optimised not for representational accuracy but for practically useful coding. Supervised training, in this case, is not done against a representational (mind-to-world) metric, but against a kind of usefulness (world-to-mind) metric. In competitive coding competitions, for instance, AlphaCode's success (and what explains its continued existence) is developing coding solutions to practical coding problems and puzzles.

Perhaps even more ambitiously, the research team at DeepMind is developing an AI optimised to 'interact' in human-like ways three dimensional space in a simulated 3-D world (Abramson et al. 2022). This AI is optimised in such a way that it will (given this aim) acquire an e-function that is at most only partly representational (e.g., reliably identifying certain kinds of behaviour cues), while also partly practical (moving objects in the 3-D world).[7]

Next, and perhaps most notably, consider – in this case due to the OpenAI research team – ChatGPT, a chatbot built on

[7] See, e.g., https://www.youtube.com/playlist?list=PLJ1sthn_UneUQ2avq5yCVszcbmcmbege6

OpenAI's GPT-3 language models, and which provides 'human-like' responses to a wide range of queries. Although ChatGPT is often used for purposes of 'fact finding' (e.g., you can ask ChatGPT to explain complex phenomena to you), it is not right to say that this AI has a representational e-function. On the contrary, ChatGPT is optimised for *conversational fluency*; to the extent that accuracy misaligns with conversational fluency, ChatGPT is optimised to favour the fluency metric.

Finally, consider a familiar AI – YouTube's recommender system – which is optimised against the metric of (in short) 'keeping people watching', and thus, generating advertising revenue (Alfano et al. 2020). When the accuracy of a recommendation choice (with respect to clustering towards videos of a similar content-type which the user has watched) misaligns with a choice more likely to keep the user watching more content, the algorithm is optimised to recommend the latter. This feature of YouTube's recommender system has been identified as playing a role in the disproportional recommendation of conspiratorial content on YouTube relative to viewers ex ante search queries.[8]

With the above short survey in mind, let's now return to the matter of the *scope* of the generalisability of their S&K's account of trustworthy AI. As I see it, at least, S&K's account can explain trustworthy AI in cases where AI acquires representational e-functions, such as the diagnostic AI example, and other AIs with representational functions, like FaceNet. But – and here is where I am less confident about their account – we've just seen that many of the most touted and promising recent AIs either lack a representational e-functions altogether (e.g., AlphaCode, ChatGPT, etc.) *or* have such a function but only alongside other practical e-functions (e.g., DeepMind's virtual world AI).

S&K seem to face a dilemma here. On the one hand, if e-function generated obligations of the sort that a disposition to fulfil them matters for AI trustworthiness are *not* limited to those obligations generated by *representational* e-functions (but also include obligations generated by *non-representational* e-functions), then it looks like the view – problematically – predicts that YouTube's recommender system, a known source of conspiratorial content, is maximally trustworthy *so long as* it is maximally fulfilling all the obligations generated by the e-function it has to 'keep viewers watching' (in turn, maximising ad revenue profits). I take it that

[8] See Alfano et al. (2020).

this result is a non-starter. Which brings us to the more plausible option and restrictive option: which is for a proponent of S&K's view of trustworthy AI to hold that e-function generated obligations of the sort that a disposition to fulfil them matters for AI trustworthiness are limited to those obligations generated by *representational* e-functions – such as, e.g., cancer diagnostic AIs, FaceNet, etc.

Let's assume this later more restrictive route is taken. On this assumption, we seem to get the result that, on S&K's view, all but the minority of AIs being developed (those like cancer diagnostic AIs, FaceNet, etc.) *fail* to meet the conditions for trustworthy AI. So does this result *overpredict* untrustworthiness in AI? Here is one reason for thinking that perhaps it does. Even if we grant that, e.g., YouTube's recommender system (in virtue of its documented propensity to recommend conspiratorial content, a propensity that aligns with its fulfilling its practical e-function) is an example of an 'untrustworthy AI' (and agree that S&K's view predicts untrustworthiness correctly here), it's less clear that, e.g., AlphaCode should get classed together with YouTube's recommender system. At least, it's not clear to me what resources S&K's proposal have for distinguishing them given that neither has been optimised to acquire a representational e-function. Without some additional story here, then, the concern is that S&K might overpredict untrustworthy AI even granting that the view diagnoses some cases of untrustworthy AI (e.g., YouTube's recommender system) as it should.

4. Concluding remarks

Giving a plausible account of trustworthy AI is no easy task; it is no surprise that, at least in 2023, the themes of trustworthy and responsible AI are among the most widely funded [9] S&K's account offers a welcome intervention in this debate because it clarifies the kind of anthropocentric barrier to getting a plausible account up and running from the very beginning, and it offers an example of how such an account that avoids this problem might go. My quibbles with the scope of the account in §3 remain, but they should be understood as just that: quibbles that invite further

[9] Along with being a regular topic funded by academic funding agencies, the importance of the question is also reflected in government-level funding. See, e.g., this initiative by the UK government https://apply-for-innovation-funding.service.gov.uk/competition/1408/overview/e8b03fe9-ca18-415a-852d-343f4231c442

development of an account that is, on the whole, a promising one.[10]

[10] Acknowledgements [xxx].

References

Abramson, Josh, Arun Ahuja, Federico Carnevale, Petko Georgiev, Alex Goldin, Alden Hung, Jessica Landon, et al. 2022. 'Improving Multimodal Interactive Agents with Reinforcement Learning from Human Feedback'. arXiv. https://doi.org/10.48550/arXiv.2211.11602.

Alfano, Mark, Amir Ebrahimi Fard, J. Adam Carter, Peter Clutton, and Colin Klein. 2020. 'Technologically Scaffolded Atypical Cognition: The Case of YouTube's Recommender System'. *Synthese*, 1–24.

Carter, J. Adam. 2022. 'Trust and Trustworthiness'. *Philosophy and Phenomenological Research*.

Carter, J. Adam, and Mona Simion. 2020. 'The Ethics and Epistemology of Trust'. *Internet Encyclopedia of Philosophy*.

Frost-Arnold, Karen. 2014. 'Trustworthiness and Truth: The Epistemic Pitfalls of Internet Accountability'. *Episteme* 11 (1): 63–81. https://doi.org/10.1017/epi.2013.43.

Graham, Peter J. 2014. 'Functions, Warrant, History'. In *Naturalizing Epistemic Virtue*, edited by Abrol Fairweather and Owen Flanagan, 15–35. Cambridge University Press.

Hardin, Russell. 1996. 'Trustworthiness'. *Ethics* 107 (1): 26–42.

Harmon, Stephanie A., Thomas H. Sanford, G. Thomas Brown, Chris Yang, Sherif Mehralivand, Joseph M. Jacob, Vladimir A. Valera, 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 (April): CCI.19.00155. https://doi.org/10.1200/CCI.19.00155.

Hawley, Katherine. 2019. *How to Be Trustworthy*. Oxford University Press, USA.

Jones, Karen. 2012. 'Trustworthiness'. *Ethics* 123 (1): 61–85. https://doi.org/10.1086/667838.

Kaur, Davinder, Suleyman Uslu, and Arjan Durresi. 2020. 'Requirements for Trustworthy Artificial Intelligence–a Review'. In *International Conference on Network-Based Information Systems*, 105–15. Springer.

Kaur, Davinder, Suleyman Uslu, Kaley J. Rittichier, and Arjan Durresi. 2022. 'Trustworthy Artificial Intelligence: A Review'. *ACM Computing Surveys* 55 (2): 39:1-39:38. https://doi.org/10.1145/3491209.

Kelp, Christoph. 2018. 'Assertion: A Function First Account'. *Noûs* 52 (2): 411–42.

Kelp, Christoph, and Mona Simion. 2021. *Sharing Knowledge: A Functionalist Account of Assertion*. Cambridge University Press.

———. 2022. 'What Is Trustworthiness?' *Nous*.

O'Neill, Onora. 2018. 'Linking Trust to Trustworthiness'. *International Journal of Philosophical Studies* 26 (2): 293–300. https://doi.org/10.1080/09672559.2018.1454637.

Schroff, Florian, Dmitry Kalenichenko, and James Philbin. 2015. 'Facenet: A Unified Embedding for Face Recognition and Clustering'. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 815–23.

Simion, Mona. 2019. 'Knowledge-First Functionalism'. *Philosophical Issues* 29 (1): 254–67.

Simion, Mona, and Christoph Kelp. 2023. 'Trustworthy Artificial Intelligence'. *Asian Journal of Philosophy*, no. Special Inaugural Issue.

William, Ivan, Eko Hari Rachmawanto, Heru Agus Santoso, and Christy Atika Sari. 2019. 'Face Recognition Using Facenet (Survey, Performance Test, and Comparison)'. In *2019 Fourth International Conference on Informatics and Computing (ICIC)*, 1–6. IEEE.