Research Article

Duncan Pritchard*

AI and Bullshit

https://doi.org/10.1515/humaff-2026-0061
Received April 8, 2026; accepted August 7, 2026

Abstract: A common charge levelled at the current wave of generative AI is that the informational outputs that it produces are bullshit. Drawing on the philosophical literature on bullshit, an examination is offered of the analogies and disanalogies between AI and bullshit. It is argued that there is a fundamental disanalogy between AI and bullshitting which ensures that, strictly speaking, AI's informational outputs are not bullshit. Nonetheless, it is also argued that there are significant parallels between the nature and source of the epistemic deficiencies of both AI and bullshitting. In particular, it is argued that both AI and bullshitters are untrustworthy informants, and for similar reasons. We are thus in a position to explain why it is natural to describe AI's informational outputs as bullshit, as it signals their distinctive kind of epistemic deficiencies, which they share with bullshit.

Keywords: AI; bullshit; epistemology; epistemology of AI; trust; trustworthiness

1 Introductory Remarks

In a relatively short space of time, AI has become a pervasive feature of modern life. This is largely due to the surprising effectiveness of the new 'second wave' of generative AI tools, powered by large language models (LLMs), that are currently all the rage.[1] Our reliance on this technology is now increasingly widespread, to the extent that whole professions are being endangered by the AI revolution. But how much trust should we put in AI as a source of information?

1 Although I suspect that many of points raised will equally apply to future iterations of AI – including the much-heralded 'third wave' of agentic AI – henceforth our focus will only be on the current wave of generative AI. For further discussion of the three waves of AI (predictive, generative, agentic), see Bruno, Canina & Biskjaer (2025).

***Corresponding author: Duncan Pritchard**, University of California Irvine, Irvine, USA,
E-mail: dhpritch@uci.edu. https://orcid.org/0000-0002-5997-0752

Tellingly, recent commentators have drawn an analogy between the outputs of AI and bullshit.[2] Drawing on the philosophy of bullshit, I will be examining how AI relates to bullshit. In particular, my focus will be on generative AI used as a source of information, such as in LLM chatbots like ChatGPT and Claude.[3] Is it accurate to accuse this particular form of AI, currently so popular, of being a bullshitter – i.e., a producer a bullshit?

Ultimately, I will be arguing that there is an important disanalogy between AI and bullshitting, to the extent that, strictly speaking, it is a mistake to describe AI as a bullshitter. But I will also be contending that there are some strong epistemic parallels between AI and bullshitting, which accounts for the temptation to describe the informational outputs of the former as bullshit. As I explain, these parallels relate to how AI resembles the bullshitter in virtue of how it is in fundamental respects – and for similar reasons – untrustworthy as a source of information. Once we understand these parallels, then although what AI is producing is not literally bullshit, there are enough important overlaps between them to explain why commentators would describe AI in these terms.

2 Bullshitting and Trustworthiness

In his seminal treatment of the topic, Frankfurt (2005) argued that in order to understand bullshit – and, in particular, to explain what is epistemically problematic about it – we need to appreciate the negative epistemic properties of the producer of bullshit – i.e., the *bullshitter*. Frankfurt argued that the defining characteristic of the bullshitter – and which makes the bullshit they produce so pernicious – is that they don't care about the truth. Given their lack of concern for the truth, the bullshitter's assertions – i.e., their bullshit – are not aimed at the truth but are rather motivated by other considerations, such as a desire to persuade. This means that bullshit is epistemically problematic even if it happens to be true.

We can get a better handle on the nature of bullshitting by contrasting it with lying. The liar cares about the truth, it's just that they want to keep you from knowing what the truth is. Indeed, successful lying requires a respect for the truth, as this is needed to avoid being found out. This is where the contrast to the bullshitter lies, as they simply don't care about the truth. When a bullshitter makes their assertions, they are guided by such considerations as wanting to persuade, to be liked, to sound clever, to avoid conflict, and so on. The bullshitter might assert claims with an air of

2 See, for example, Hannigan, McCarthy & Spicer (2024) and Hicks, Humphries & Slater (2024).

3 Henceforth, when I write about 'AI' without qualification, the reader should take it that it is specifically this form of AI that I have in mind.

complete confidence, but if so that won't be because they are confident in the truth of what they assert, but will rather reflect other motivations for their assertion, such as the desire to impress. The truth, even to the extent of hiding the truth as the liar attempts, is at most a secondary consideration for the bullshitter, if it carries any weight at all.

Frankfurt argued that bullshitting has a pernicious effect on our epistemic culture, due to how it undermines the value of truth. One way of drawing out why bullshitting is epistemically problematic is by asking whether we should trust the bullshitter as a source of information. Notice that we can split this question into two sub-questions. The first specifically relates to *interpersonal trust*. Is the bullshitter a trustworthy person? In contrast to interpersonal trust, which as the name suggests is specifically concerned with trust in persons, there is a more general kind of *reliance trust* that is applicable to both people and objects. Can we appropriately rely on the information that the bullshitter tells us?

While interpersonal trustworthiness entails reliance trustworthiness, the converse doesn't hold. This is because someone can be an accurate source of information even if they are not a trustworthy person. Perhaps they would tell you falsehoods if thought they could get away with it, but they realize that they can't. Interpersonal trust is thus the more demanding notion. If a device that you trusted lets you down – your alarm clock doesn't go off when you expected it to, say – then it hasn't betrayed your trust. It has just disappointed you, that's all. In contrast, if a person you trusted lets you down, then they have betrayed your trust. You won't just be disappointed, but more specifically you will be disappointed in them.[4]

There is a considerable philosophical literature articulating exactly what is involved in interpersonal trust over and above mere reliance.[5] But one factor that this literature converges on is that it has something essentially to do with honesty, at least insofar as we are concerned with interpersonal trust with regard to informants at any rate.[6] A liar is not a trustworthy informant, even if they happen to tell you the truth.

Here is a relatively minimal condition that we can put on interpersonal trust that would, in part at any rate, capture this intuition: that the interpersonal trustworthiness of an informant entails that they are generally disposed to only assert

4 For an influential discussion of how persons can betray our trust while devices that we rely upon can merely disappoint us, see Baier (1986).

5 See, for example, Baier (1986); Jones (1996); Potter (2002), and Hawley (2014, 2019). For some recent surveys of this literature, see Hawley (2012) and McLeod (2020).

6 For example, a popular view is that what is required for interpersonal trust is that the person trusted can be relied upon in virtue of their goodwill towards the person trusting them. As applied to an informant, that would require something like honest dealings with the recipients of their testimony. See especially Baier (1986). For related views, see Jones (1996) and Potter (2002).

information that they believe is true. Call this the *general sincerity condition* on interpersonal trustworthiness of an informant. I say that this is a minimal condition because we would need a much stronger condition to capture what is involved in interpersonal trustworthiness of an informant. Indeed, as it stands the general sincerity condition doesn't exclude liars as being untrustworthy so long as they keep their lies to a minimum, as most liars do. In the normal run of things, after all, liars will often assert what they believe to be true. It is just that, when the opportunity comes along to make use of deception in order to further their own interests, then they will switch tactics and assert falsehoods – that's what makes them liars. The point is that, for most people at any rate, what makes you a liar is not that you tell lies all the time, but that you are disposed to tell lies on the occasions, which might be relatively rare, where lying suits your interests. But liars seem to be paradigmatically untrustworthy informants.[7]

So our condition is too weak to be a general constraint on what differentiates interpersonal trust from mere reliance. Nonetheless, it is plausible that interpersonal trust would require that one at least satisfy this minimal condition. This point is significant because the bullshitter doesn't satisfy even this minimal condition on interpersonal trust. Remember that the bullshitter doesn't care about the truth. This means that they go around asserting things without being concerned about whether what they are saying is true. They are more interested in other considerations, such as whether what they assert will be thought impressive, persuasive, shocking, and so on. But that is just to say that the bullshitter is not the kind of person who is generally disposed to only assert what they believe to be true.

In this regard, the bullshitter is worse than the typical liar. The typical liar, as we just noted, is disposed to only lie on the relatively small number of occasions where this is in their interests. Their default, like most other people, will be assert what they believe to be true. The bullshitter, in contrast, lacks this disposition. This is not because they have the opposing disposition to assert falsehoods all the time, but rather because for them the truth of what they assert is, at most, an afterthought. That the bullshitter is not generally disposed to only assert what they believe to be true explains why they are untrustworthy informants. Even if the bullshitter happens to regularly assert truths, this would suffice to make them trustworthy in this regard.

What about our less demanding notion of trustworthiness that was concerned with mere reliance? Should one rely on the information asserted by a bullshitter, and

7 Interestingly, one can also be a dishonest informant without telling explicit lies, as when one withholds important information and thereby misleads one's audience. Our condition wouldn't exclude such a person from being an untrustworthy person either, but arguably they are also untrustworthy.

thus trust them in this sense? In answering this question, we encounter an interesting point about what such reliance trust involves. One might have antecedently thought that all that is required for appropriate reliance on an informant is that what they assert is generally true. If that's right, then our question of whether one should rely on the word of a bullshitter collapses into the question of whether what they assert is generally true. If it is generally true, then one should rely on them; if it isn't, then one shouldn't. Knowing that someone is a bullshitter doesn't settle this question, as perhaps they happen to be bullshitters whose assertions generally converge with the truth. Although unlikely, there is nothing in the notion of a bullshitter that excludes this possibility.

That someone generally asserts true information is not enough to make it appropriate for others to rely on them as an informant, however. What is important is not only whether the informant is likely to assert something false but also whether they could easily assert something false. The crux of the matter is that if an informant could easily assert something false, then they are not reliance trustworthy – i.e., one shouldn't rely on their information – even if in general they mostly assert truths.

In order to see this point, it will be useful to set informants to one side and consider an example where we are reliance trusting a device rather than a person. Consider the example of trusting that one's parachute will open. The parachute opening when it is supposed to is clearly an outcome of great importance. No-one wants to be in the situation of hurtling towards the earth and discovering that their parachute has failed to open. That the parachute is likely to open, however, is not enough to ensure that it couldn't easily fail to open. One can imagine a design flaw in the parachute that means that although it usually opens – with a very high 99 % rate of success, say – it nonetheless could easily fail to open. Perhaps there is a part to the parachute that is especially weak and that if this fails, then that compromises the parachute as a whole. That the part doesn't fail 99 % of the time is nonetheless compatible with the fact that it could very easily fail. But if it could easily fail, then you shouldn't rely on the parachute opening, even despite the high likelihood of success. In this sense of being trustworthy, the parachute is untrustworthy.

Note that this point is entirely general, and certainly not specific to the dramatic outcome of a parachute failing to open. Consider your alarm clock. Suppose it works 99 % of the time, but that it is also prone to not go off when it is supposed to every now and again for no apparent reason. Like the parachute, it could easily let you down even though, from a purely statistical point of view, it is unlikely to do so. That it could easily let you down suffices to ensure that one should not rely upon it. If you need to be sure of getting up tomorrow – for that all important interview, say – then you shouldn't rely on this alarm clock. In short, it is not a trustworthy alarm clock.[8]

8 For further discussion of this point in the context of risk, see Pritchard (2015; 2026b).

Here's why this point is relevant to the bullshitter. While the bullshitter doesn't care about the truth, that doesn't mean that most of what they assert is false. That they don't care about the truth entails that when they make their assertions they are motivated by other considerations, such as appearing clever, being admired, winning an argument, and so on, and not by whether what they say is accurate. But that their assertions are being guided by these other considerations doesn't in itself entail that what they assert is false. We can thus at least envisage bullshitters who often assert truths.

What's significant about the bullshitter, however, is that since they don't care about the truth, then if they do regularly assert the truth they could nonetheless very easily have asserted falsehoods instead. Since their assertions are not guided by the truth, then there is nothing preventing them from asserting a falsehood if that serves their interests better – if, for example, they think that asserting a falsehood will impress their audience, will get them the job, will extract them from a tricky situation, and so on. This is an important difference between a bullshitter who happens to mostly assert truths and an honest person who is guided by the truth in what they assert and who, as a result, also mostly asserts truths. While both this particular bullshitter and the honest person mostly assert truths, only the honest person is such that they couldn't have easily asserted falsehoods. That's what caring about the truth means. It's not just good fortune that an honest person's assertions line up with the truth, but rather reflects the fact that there is a systematic (albeit, obviously, defeasible) connection between what they assert and the truth. In contrast, even if the bullshitter happens to assert truths, the fact remains that they could easily have asserted falsehoods instead. That's a consequence of them failing to care about the truth.

We have thus identified two ways in which the bullshitter is untrustworthy as an informant. The first specifically relates to the more demanding notion of interpersonal trust. The bullshitter isn't interpersonally trustworthy because their assertions are not sincere expressions of what they believe to be true but rather fundamentally disconnected from a concern for the truth. The second relates to the more general notion of reliance trust that applies to both persons and objects. The bullshitter isn't trustworthy in this sense because even if they happen to assert mostly truths, they could nonetheless have easily asserted falsehoods. Bullshitters are thus untrustworthy as informants in both of the senses of untrustworthiness that we have identified. It is no wonder then that bullshitting is regarded as being epistemically problematic. Moreover, notice that the bullshitter's untrustworthiness in both cases is independent of whether what they assert happens to be mostly true. As we will see, this point will be significant once we examine the relationship between AI and bullshit.

3 AI Is Not a Bullshitter

Now that we have set out what bullshit is and why it is epistemically problematic, we are in a position to evaluate the claim that the informational outputs that AI produces are bullshit. I will be arguing for two claims. The first is that, strictly speaking, AI cannot produce bullshit because it cannot satisfy the criteria for being a bullshitter. The second is that there are, nonetheless, important overlaps between AI and bullshitting that ensure that AI's informational outputs are in important respects akin to bullshit. In particular, like the bullshitter, AI is an untrustworthy informant, and for similar reasons.

We noted above that what is key to the bullshitter is their lack of concern for the truth. I think one of the main motivations for describing what AI is producing as bullshit relates to this aspect of the bullshitter, given that we find AI exhibiting the same lack of concern for the truth. Indeed, this is a product of AI's design. AI is effective because it has been trained on massive datasets of our information. This entails that the outputs that AI produces have a fundamentally different orientation when compared to our cognition. Our cognitive lives take place in a relationship to a world that is external to us, a world that is both physical and social. We respond to these external inputs in forming our thoughts, beliefs, doubts, and so forth that are the constituents of our assertions. This brings a truth-orientation dimension to our cognitive lives. In contrast, AI lacks external inputs of this kind, as it is trained only on our responses to these external inputs. This is the sense in which at a fundamental level it is disconnected from how things are. This is why its outputs are not responsive to the world around it but rather reflections of our own judgements about the world.[9]

Its training on these massive datasets of our information enables AI to learn how to effectively *mimic* the human thought and reasoning that is found in these datasets. But being an effective mimic of human thought and reasoning is not the same thing as actual thinking and reasoning, even if the difference won't always be evident from looking at AI's outputs. In fact, the AI isn't engaging in any actual reasoning in generating its outputs. Indeed, not only does AI not reason its way to the outputs it produces, but there is also no reason to think that it understands any of its outputs either. It is merely a statistical engine, predicting what the right output should be based on the massive data sets it has been trained on.[10]

---

9 For further discussion of this point, see Wooldridge (2022).

10 For a fuller defence of this point, see Mitchell (2019) and Chomsky, Roberts & Watumul (2023). For a recent influential discussion of the idea that AI reasoning is merely apparent, see Shojaee et al. (2025). For a helpful survey of the recent literature on this topic, see Mitchell & Krakauer (2023).

Once one grasps that this is how AI works, then it becomes clear why its informational outputs are not truth-directed in the way that our assertions characteristically are. AI outputs cannot be sincere expressions of beliefs because AI is simply not the kind of thing that could have beliefs, much less be sincere in what it believes. To be a believer is to think, to reason, to understand, and AI is none of these things. This point gets obscured by how much of an effective mimic it is of our thought and reasoning, to the extent that we are apt to attribute thought, reason and understanding to it – and thus beliefs too – but this is based on a misunderstanding about what AI is actually doing when it generates its outputs.[11]

The significance of this point is that although AI can seem very alike the bullshitter in terms of its lack of concern for the truth, there is nonetheless a very important disanalogy between the two. When making assertions, the bullshitter exhibits a motivational structure in which the truth does not play a significant role. This is the sense in which the bullshitter lacks a concern for the truth: they are indifferent to the truth because they are motivated to a greater extent by other considerations. In contrast, AI lacks any kind of motivational structure at all. Take the claim that AI lacks a concern for the truth. While this is technically accurate – it's true that AI doesn't have a truth-directed motivational state – this is also in a certain sense highly misleading, since it fails to distinguish AI, which lacks any kind of motivational state, from a bullshitter whose motivational states don't appropriately prize the truth.

This distinction is important, since what is crucial to the bullshitter is that they are motivated as informants by considerations other than the truth. AI, in contrast, is unlike a bullshitter on just this front. It's not that it lacks an appropriate concern for the truth because it is motivated by other factors, but because it doesn't have any motivational states at all. This is the sense in which AI is not really even indifferent to the truth, as that suggests that truth is a low-ranking consideration in its motivational structure as opposed to there being no motivations at all.[12] AI merely seems to have motivational states because of how it is trained to mimic our responses. This is why, strictly speaking, AI is not a bullshitter and hence what it produces is not bullshit.

---

11 This is what computer scientists call the 'ELIZA effect', whereby we are misled by the apparently humanlike outputs of the generative AI into attributing to it the kind of thought, reasoning and understanding that is implicit in the information that we present to others. For a recent discussion of this effect, including the early chatbot that coined this name, see Natale (2021, ch. 3).

12 This is why I have some reservations about how AI is often described in the literature as being indifferent to the truth. See, for example, Fredrikzon's (2025) discussion of AI's "epistemic indifference." The bullshitter is indifferent to the truth. In contrast, AI doesn't even rise to the level of indifference.

4 AI Informational Outputs and Bullshit

It would be a mistake, however, to leave that matter there. While AI is not, strictly speaking, a bullshitter, and hence the informational outputs it produces are not, strictly speaking, bullshit, there clearly are significant parallels between bullshitting and AI. When commentators compare AI's informational outputs to bullshit, it is these parallels that are driving the comparison. In particular, to describe AI's informational outputs as bullshit is to suggest that they are epistemically deficient in ways that are analogous to the epistemic deficiencies of bullshit. We can bring this point into sharp relief by highlighting how AI is untrustworthy in a way that parallels the untrustworthiness of bullshitting, and for similar reasons.

Let's start with the interpersonal conception of trustworthiness. We noted above that the bullshitter fails the minimal general sincerity condition that applies to this notion of trustworthiness. Given the foregoing, it should be clear that AI is unable to satisfy this condition too. In particular, AI doesn't assert what it believes to be true for the simple reason that it doesn't have any beliefs. Indeed, I would suggest that something that is unable to have beliefs is not able to make assertions either. But even if we relax the minimal condition on interpersonal trustworthiness so that its focus is informational outputs rather than assertions, then the point remains that AI doesn't believe these informational outputs to be true. This is thus one sense in which AI is like the bullshitter, in that neither of them has a concern for the truth that would be characteristic of a trustworthy informant.

There is, however, a crucial difference between the bullshitter and AI on this front. We've alluded to this difference already in how we've described the former as failing the minimal condition on interpersonal trustworthiness and the latter as being unable to satisfy it. The bullshitter doesn't fail this condition because they lack beliefs, but rather because their indifference to the truth means that their assertions are not guided by a concern for the truth. In contrast, AI is simply not the kind of thing that could have beliefs, which is why it is unable to satisfy this condition. Put another way, we could imagine a bullshitter changing their ways and discovering a concern for the truth. As a result, they would become someone who satisfies this condition and so is a trustworthy informant. But this is not a possibility for AI.

The disanalogy between AI and bullshitters is a reason for thinking that we shouldn't even be evaluating AI for interpersonal trustworthiness in the first place. We apply the more demanding conditions for interpersonal trustworthiness to people because we attribute to them rich cognitive lives involving beliefs, motivations, reasoning and so forth. It is only if we make these attributions that it makes sense to assess them for this kind of trustworthiness. The point of the foregoing is that we shouldn't be led astray by the similarities between AI's outputs and our own

thought and reasoning into attributing the rich cognitive lives that we exhibit to AI. The upshot is that although AI is unable to satisfy the test for interpersonal trustworthiness, we should never have subject it to such a test, as it is not the kind of thing that could coherently be interpersonally trustworthy. Accordingly, we should only be evaluating AI for the more general notion of reliance trustworthiness.[13]

This brings us to the second way in which bullshitters are untrustworthy that we noted above. This is that we cannot appropriately rely on them as informants because even if they happen to mostly assert truths, they could nonetheless easily assert falsehoods. This point has a straightforward application to AI. For although AI informational outputs are generally accurate, they are also prone to error in the specific sense that they could easily be false. Moreover, this feature of AI's informational outputs is a manifestation of the point noted earlier about how there is a fundamental sense in which AI is disconnected from the truth. This further reinforces the analogies between AI and bullshitting, given that what ensures that the bullshitter isn't reliance trustworthy is their lack of concern for the truth.

The reason why AI's informational outputs could easily be false, even though they are generally accurate, is because it is notoriously disposed to represent outputs as factual that are completely dislocated from reality. These confabulations are known as *hallucinations*, given that they seem as if the AI is in the grip of some sort of delusion when it produces these outputs, as if it is under the impression that the false outputs are accurate.[14] The reality, however, is that these false outputs reveal a fundamental disconnect with the truth. For example, when the error is pointed out, the AI will readily admit that the output it delivered was false. This suggests that the problem isn't a misapprehension about the truth, much less active deceit (which would in any case presuppose motivations that AI lacks), but rather an undue level of willingness to confidently offer outputs that are false.[15] This should remind us of the bullshitter's propensity to confidently make assertions regardless of whether what is asserted is true.

We've already noted AI's disconnect from the truth, of course, but what is significant about this phenomenon is how this disconnect generates a certain kind of false informational output. In particular, notice that the problem posed by

13 See Ryan (2020) for further defense of this claim. For an opposing view, see Simion & Kelp (2023).

14 For further discussion, see Sun et al. (2024) and Kalai et al. (2025). Some commentators have suggested that 'confabulation' is a better term to use. See, for example, Smith, Greaves & Panch (2023) and Wiest & Turnbull (2025). See also the related problem associated with misdescribing what is in an image based on small changes in pixelization that are indistinguishable to the human eye. For discussion, see Su, Vargas & Sakurai (2019).

15 The issue of whether AI is engaged in deception gets muddled in the contemporary literature by the fact that many commentators explicitly employ the notion of deception in a non-standard way that doesn't involve any intent to deceive. See, for example, Simon (2025, §3).

hallucination is not that it entails that AI's informational outputs are generally false. On the contrary, from a purely statistical point of view at least, AI's informational outputs are usually likely to be true. The difficulty is rather that given how AI is generating its informational outputs it could very easily present a false output rather than a true one. As we noted above, however, this is sufficient to make it such that one ought not to rely on an informant. AI, *qua* informant, is thus not reliance trustworthy.

Again, there is a parallel between the untrustworthiness of the bullshitter and AI on this specific front. In both cases, the information that they convey is not to be relied upon because it could easily be false. Similarly, in both cases that this information could easily be false is compatible with them being sources of information that generally deliver accurate information. Moreover, it is also significant that for both bullshitting and AI the reason why they could easily deliver false information is because of a lack of concern for the truth.[16]

5 Concluding Remarks

We've argued that there is a fundamental disanalogy between AI and bullshitting. Strictly speaking, AI is not (even) a bullshitter, as it lacks the requisite motivational states, and hence the informational outputs that it produces are not bullshit. Nonetheless, we've also argued that there are significant overlaps between AI and bullshitting. In particular, both are epistemically deficient in similar ways, and for similar reasons. We have brought this point out by considering how the bullshitter is untrustworthy, both in the interpersonal and the reliance sense of trustworthiness. The same is also true of AI (though, as we've noted, AI shouldn't even be evaluated for interpersonal trustworthiness). Moreover, the source of these epistemic deficiencies is the same in both cases, in that it arises from how AI, like the bullshitter, lacks a concern for the truth. Accordingly, even though we have ultimately argued against treating AI as a bullshitter, we have also explained why it is so natural to describe AI and its informational outputs in these terms. AI might not literally be a bullshitter, but the nature and source of its epistemic deficiencies means that what it produces is very much akin to the bullshit produced by bullshitters.

Acknowledgments: I am grateful to Kieran Brayford and to two anonymous reviewers for Human Affairs: Postdisciplinary Humanities & Social Sciences Quarterly for their comments on an earlier version of this manuscript. This research was supported by the John Templeton Foundation ('Cultivating Intellectual

16 For further discussion of whether AI is trustworthy as a source of information, see Simion & Kelp (2023) and Pritchard (2026a).

Character in the AI Age', #63365) and by the Czech Science Foundation ('Vicious Epistemic Cultures', #24-11282S, PI: Laura Candiotto, University of Pardubice). The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the John Templeton Foundation or the Czech Science Foundation. This essay was written while a Senior Research Associate of the African Centre for Epistemology and Philosophy of Science at the University of Johannesburg.
Research ethics: Not applicable.
Informed consent: Not applicable.
Author contributions: The author has accepted responsibility for the entire content of this manuscript and approved its submission.
Use of Large Language Models, AI and Machine Learning Tools: None declared.
Conflict of interest: The author states no conflict of interest.
Research funding: John Templeton Foundation ('Cultivating Intellectual Character in the AI Age', #63365).
Data availability: Not applicable.

References

Baier, A. 1986. "Trust and Antitrust." *Ethics* 96 (2): 231–60.
Bruno, C., M. Canina, and M. M. Biskjaer. 2025. "The Three Waves Paradigm: Tracing the Evolution of Artificial Intelligence in Design." In *The Cyber-Creativity Process: How Humans Co-Create with Artificial Intelligence*, edited by G. E. Corazza, 129–62. London: Palgrave Macmillan.
Chomsky, N., I. Roberts, and J. Watumul. 2023. "The False Promise of ChatGPT." *The New York Times*. https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html (accessed January 1, 2026).
Frankfurt, H. 2005. *On Bullshit*. Princeton: Princeton University Press.
Fredrikzon, J. 2025. "Rethinking Error: "Hallucinations" and Epistemological Indifference." *Critical AI* 3 (1). https://doi.org/10.1215/2834703X-11700255.
Hannigan, T. R., I. P. McCarthy, and A. Spicer. 2024. "Beware of Botshit: How to Manage the Epistemic Risks of Generative Chatbots." *Business Horizons* 67 (5): 471–86.
Hawley, K. 2012. *Trust: A Very Short Introduction*. Oxford: Oxford University Press.
Hawley, K. 2014. "Trust, Distrust and Commitment." *Noûs* 48 (1): 1–20.
Hawley, K. 2019. *How to be Trustworthy*. Oxford: Oxford University Press.
Hicks, M. T., J. Humphries, and J. Slater. 2024. "ChatGPT is Bullshit." *Ethics and Information Technology* 26 (38).
Jones, K. 1996. "Trust as an Affective Attitude." *Ethics* 107 (1): 4–25.
Kalai, A. T., O. Nachum, S. S. Vempala, and E. Zhang. 2025. "Why Language Models Hallucinate." *arXiv*. https://doi.org/10.48550/arXiv.2509.04664.
McLeod, C. 2020. "Trust." In *Stanford Encyclopeadia of Philosophy*, edited by E. Zalta. https://plato.stanford.edu/entries/trust/ (accessed January 1, 2026).
Mitchell, M. 2019. "Artificial Intelligence Hits the Barrier of Meaning." *Information* 10 (2): 51.
Mitchell, M., and D. C. Krakauer. 2023. "The Debate over Understanding in Ai's Large Language Models." *Proceedings of the National Academy of Sciences* 28 (13): 120.

Natale, S. 2021. *Deceitful Media: Artificial Intelligence and Social Life After the Turing Test*. Oxford: Oxford University Press.

Potter, N. N. 2002. *How Can I be Trusted? A Virtue Theory of Trustworthiness*. Lanham: Rowman and Littlefield.

Pritchard, D. H. 2015. "Risk." *Metaphilosophy* 46: 436–61.

Pritchard, D. H. 2026a. "Generative AI is Untrustworthy." *Synthese* 208. https://doi.org/10.1007/s11229-026-05755-y.

Pritchard, D. H. 2026b. *Tempting Fate: Taking Risks, Pushing One's Luck, and Living an Authentic Life*. Princeton, NJ: Princeton University Press.

Ryan, M. 2020. "In AI We Trust: Ethics, Artificial Intelligence, and Reliability." *Science and Engineering Ethics* 26 (5): 2749–67.

Shojaee, P., I. Mirzadeh, K. Alizadeh, M. Horton, S. Bengio, and M. Farajtabar. 2025. "The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity." *Apple Machine Learning Research*. https://machinelearning.apple.com/research/illusion-of-thinking (accessed January 1, 2026).

Simion, M., and C. Kelp. 2023. "Trustworthy Artificial Intelligence." *Asian Journal of Philosophy* 2 (8): 8.

Simon, J. 2025. "Generative AI, Quadruple Deception and Trust." *Social Epistemology* 40 (1): 101–15.

Smith, A. L., F. Greaves, and T. Panch. 2023. "Hallucination or Confabulation? Neuroanatomy as Metaphor in Large Language Models." *PLOS Digital Health* 2 (11): e0000388.

Su, J., D. V. Vargas, and K. Sakurai. 2019. "One Pixel Attack for Fooling Deep Neural Networks." *arXiv*. https://arxiv.org/pdf/1710.08864 (accessed January 1, 2026).

Sun, Y., D. Sheng, Z. Zhou, and Y. Wu. 2024. "AI Hallucination: Towards a Comprehensive Classification of Distorted Information in Artificial Intelligence-Generated Content." *Humanities and Social Sciences Communications* 11 (1278): 1278.

Wiest, G., and O. H. Turnbull. 2025. "Faulty Artificial Intelligence, or the Sleep of Reason." *New England Journal of Medicine AI* 2 (11). https://doi.org/10.1056/AIp2500785.

Wooldridge, M. 2022. "What is Missing from Contemporary AI? the World." *Intelligent Computing* 2022. https://doi.org/10.34133/2022/9847630.