Synthese (2026) 208:97
https://doi.org/10.1007/s11229-026-05755-y

ORIGINAL RESEARCH

Generative AI is untrustworthy

Duncan Pritchard[1]

Received: 17 October 2025 / Accepted: 27 July 2026

Abstract
Our reliance on the new wave of generative AI has increased exponentially in the last few years, with little sign of abating. This raises the natural question of whether we should trust this technology as a source of information. Drawing on the epistemology of trust, it is argued that generative AI does not meet the conditions for being a trustworthy source of information. In particular, it is argued that generative AI is untrustworthy because it is an unsafe source of information. This, in turn, entails that trusting generative AI as a source of information is not a route to knowledge.

Keywords AI · Epistemology · Epistemology of AI · Reliability · Safety · Trust · Trustworthiness

1 Introductory remarks

The last few years have seen the development of a new 'second wave' of generative AI that appears to be tremendously useful for a wide range of applications. Indeed, its potential is such that whole professions are now under threat from this new technology. The undoubted effectiveness of generative AI for many cognitive tasks raises the question of just how much we should trust it as a source of information. I will be arguing that generative AI is not a trustworthy source of information. Relatedly, I will be claiming that trusting generative AI is not a route to knowledge. Importantly, both of these claims are meant to be compatible with the general reliability of generative AI, in the specific sense that as a source of information it produces accurate results with a high degree of probability.

✉ Duncan Pritchard
dhpritch@uci.edu

[1] University of California, Irvine, USA

Published online: 07 August 2026

2 Two varieties of trustworthiness

In the literature on trust it is common to distinguish between two varieties. The first variety specifically relates to trusting a person. A wife trusts her husband to be faithful. A customer trusts their tax advisor to do a thorough and conscientious job preparing their taxes. A tourist trusts the person in the information booth to give them accurate directions to their destination. And so on. Call this *interpersonal trust.*

The second variety relates to the trust of objects, such as devices, apparatus, instruments, and so forth. A skydiver trusts their parachute to open and deliver them safely to earth. A driver trusts their car to start on the frosty morning when they need to get to work. A homeowner trusts the A/C to burst into action when the summer heatwave arrives. A golfer trusts the weather app on their phone to accurately predict the conditions on the course tomorrow morning. And so on. Call this *object trust.*

In order to settle the question of whether generative AI is trustworthy as a source of information we first need to determine what kind of trust is involved. One might think that since generative AI is technology, then the type of trust involved is straightforwardly object trust. But perhaps that is too quick. After all, generative AI is at the same time technology that is explicitly designed to engage with us in a way that invites us to treat to it as another person. For example, it replies to our questions with answers that are to a large extent indistinguishable from the answers a knowledgeable human being would give. Whether we ought to impersonally trust AI or not, the fact remains that our manner of trusting it is often of this nature. Accordingly, there is at least a question to be asked about whether we should treat our trust of generative AI in interpersonal terms.

The distinction between these two varieties of trust is important because there is more involved in being trustworthy when it comes to interpersonal trust as opposed to merely object trust. What is common to both forms of trust is reliance. When we trust an object we are relying on it to generate an outcome, as when we trust a parachute to open. Similarly, trusting a person involves relying on this person to generate an outcome, such as when one trusts a tax advisor to do the good job that they promised when it comes to preparing one's tax returns.

Interpersonal trust, however, demands more than just reliance. This is because one can rely on a person that one doesn't trust. Perhaps one knows that the tax advisor is a slapdash kind of person, but that he can be relied upon to do a good job with one's taxes nonetheless because his work is regularly audited by his boss. One thus has good reasons to rely on him even though one doesn't trust him because one knows that he isn't trustworthy in this regard.[1] The point is that when we trust someone we are doing more than just relying on them. This is reflected in our very different responses when a person we trusted, as opposed to an object, turns out to be untrustworthy. If a person you trusted lets you down, then they have betrayed your trust. In

[1] Similarly, non-reliance does not entail distrust. That I don't rely on someone to do something, such as collect my mail when I'm away, doesn't entail that I distrust them in this regard. For a useful discussion of the normative dimension to interpersonal trust and distrust, see Hawley (2014).

contrast, if a device that you trusted lets you down, then it hasn't betrayed your trust. It has just disappointed your expectations.[2]

In a comprehensive spirit, I will be considering whether generative AI is trustworthy as a source of information in both senses of trust. I will be claiming that generative AI is not the kind of thing that could be worthy of interpersonal trust. That leaves the possibility that it might be worthy of object trust. Unfortunately, I will be claiming that generative AI is not trustworthy in this sense either. This is because one cannot rely upon it to produce the relevant outcomes. Since both forms of trust involve reliance, this second point reinforces the first, in that it provides additional support for the claim that generative AI is not worthy of interpersonal trust.

Note that in both cases the negative claims that I will be defending are specifically tied to the features of generative AI. The possibility is thus left open that future iterations of AI, which move beyond generative AI, might be object trustworthy or even interpersonally trustworthy. Significantly, understanding why generative AI fails to be trustworthy as a source of information clarifies how future iterations would need to be different from generative AI in order to count as trustworthy in one or both of the ways discussed.

3 Generative AI and interpersonal trustworthiness

Let's start with our interpersonal trust of generative AI as a source of information. As noted above, as a descriptive matter at least, we often do interpersonally trust generative AI, given that it delivers information in a way that is just like a human response. That invites the normative question of whether we ought to trust generative AI in this manner.

We can understand this normative question in one of two ways. The first is whether generative AI meets the conditions associated with interpersonal trust—i.e., whether it is trustworthy by this measure. This version of the normative question assumes that generative AI is the kind of thing that can appropriately be treated as interpersonally trustworthy.

We can also ask whether generative AI is the kind of thing which is even in the market for interpersonal trust in the first place, which would be a very different way of understanding our normative question. If generative AI is not the kind of thing that can properly be subject to such trust, then it is trivially the case that the first construal of our normative question has a negative answer (i.e., that generative AI won't satisfy the conditions for interpersonal trustworthiness). For example, a thermometer cannot satisfy the conditions for interpersonal trust because it is not the kind of thing that it would be appropriate to interpersonally trust. Accordingly, if someone were to trust it in this manner, then they would simply be mistaken. I will be arguing for a negative answer to this second construal of our normative question, which therefore entails a negative answer to the first construal of our normative question as well.

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

We've already noted that the conditions for being impersonally trustworthy are more demanding than for object trust, on the grounds that interpersonal trust involves more than just reliance. There is a considerable literature that discusses what this extra dimension to interpersonal trust might involve.[3] For our purposes, however, we don't need to settle this general question here. Since we are concerned with the trustworthiness of generative AI as a source of information, we can focus our attention on the conditions for the interpersonal trustworthiness of an informant. Moreover, rather than articulating these conditions in full, we can ask instead whether there is a condition on this kind of interpersonal trust that generative AI would not satisfy. If that is the case, then that would suffice to show that generative AI is not interpersonally trustworthy in the specific sense that interests us.

Consider the following plausible condition on interpersonal trustworthiness with regard to an informant: that the informant needs to be presenting information that they believe to be true. An informant who is lying to you, for example, is clearly not trustworthy. Indeed, this is so even if, by some twist, they happened to tell you truths in error. If you found out that an informant you trusted lied to you, even if the information they presented fortuitously turned out to be accurate, then you would regard them as having betrayed your trust.

While liars are the most obvious case where this condition on the interpersonal trustworthiness of an informant is not met, there are other scenarios that are relevant in this regard. This is because there are cases where someone presents information that they believe to be false that don't involve lying, at least in a straightforward sense of lying at any rate. Consider, for example, Jennifer Lackey's example of the creationist schoolteacher.[4] This is someone who teaches evolution in school, and does it well, because that's her job as a science teacher. Crucially, however, she doesn't believe any the claims she teaches in this regard because she's a creationist (a fact that she keeps to herself, and especially from the students and the other teachers). Is she lying? Perhaps, though this is far from being straightforward. Ultimately, it depends on what account of lying one endorses.[5] For example, one might argue that what one is required to assert in a professional capacity as a teacher need not be what one believes, in which case we shouldn't automatically treat such assertions as expressions of belief.

One thing that sets this case apart from conventional cases of lying is that the information offered by the creationist schoolteacher is not only accurate but also generally thought to be accurate, including by the relevant experts in that domain (as even the creationist schoolteacher is aware). The creationist schoolteacher is thus at least an atypical liar on this score. Indeed, her pupils can and should rely on the information that she gives them. Whether or not we think the creationist schoolteacher is a liar, I don't think we would regard her as trustworthy. This is because she is merely asserting what she is required, as part of her job, to assert and not what she believes.

[3] For two very useful overviews of the philosophy of trust that cover the contrast between (interpersonal) trust and mere reliance, see Hawley (2012) and McLeod (2020).

[4] See Lackey (2008), though note that she uses this example to demonstrate a different point to the one we are making here.

[5] For a recent survey of philosophical work on lying, see Mahon (2015).

Imagine that you are one of her pupils who trusted the testimony of your excellent science teacher. Perhaps you even abandoned the creationist beliefs of your religious upbringing as a result of her inspirational teaching of evolution (she is, remember, an excellent teacher). Subsequently you discover that she didn't believe what she told you and was merely communicating the official line on evolution as dictated by her job. Wouldn't you think that your trust had been betrayed?

Notice that what we are proposing is a very minimal condition on interpersonal trust of an informant. We would normally require a lot more of someone who was trustworthy in this sense. For example, one can actively mislead someone without asserting falsehoods, such as by failing to disclose relevant information. A car salesman who realizes that the customer is radically mistaken about the condition of the car they are purchasing might not assert any falsehood as part of the sale. It may not be generally prudent to interpersonally trust the word of car salespeople, but if you do, and you subsequently discover that your evident misapprehension about the condition of the car you're buying wasn't corrected, then you will surely feel that your trust was betrayed. Points like this suggest a more demanding condition on interpersonal trustworthiness of an informant. For example, one might require that the informant should be an honest source of information, which would entail highlighting relevant facts that the recipient of that information might have overlooked.[6] If that's right, then our more minimal condition on interpersonal trust ought to be relatively uncontroversial.[7]

The relevance of this condition for our use of generative AI is that there seems to be no credible way of construing what the AI is doing as presenting information that it believes to be true. The issue here isn't that AI is presenting information that it *doesn't* believe to be true, but rather that it is not the kind of thing that can present information in this manner at all. Generative AI is effective because it has been trained on massive datasets of our information. This enables it to learn how to effectively *mimic* human thought and reasoning, since it has been trained on many instances of such thought and reasoning. 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 the outputs that generative AI delivers. Indeed, once we appreciate how generative AI is producing its outputs,

[6] The standard accounts of interpersonal trustworthiness, which are not specific to interpersonal trust of a person *qua* informant, tend to go down these lines. For example, that what is required 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). See also endnote 7.

[7] Even an account of interpersonal trustworthiness that builds in very little about the motives of the person being trusted would seem to entail our minimal condition on interpersonal trustworthiness of an informant. For example, Hawley's (2014, 2019) influential account of interpersonal trust requires that one believe that the person one is relying upon has the relevant commitments. In the case of an informant, that would plausibly require that one regards one's informant as presenting what they believe to be true (i.e., such that they are committed to the truth of what they are asserting), such that if they aren't doing this, then they fail to be trustworthy.

then it becomes clear that it doesn't even understand the language that it is using to convey this information.[8]

This is why generative AI isn't even in the market for interpersonal trust. It's not just that it fails to satisfy the condition we have laid down on interpersonal trustworthiness, but that it's not the kind of thing that could ever satisfy such a condition. If generative AI doesn't even understand the information that it is presenting, then it cannot in any plausible sense believe it to be true (or, for that matter, believe it to be false). This is why generative AI, for all the sophistication of its informational outputs, is nonetheless more akin, in this regard at least, to a thermometer than a liar (who obviously can present information that they believe to be false).[9]

It follows that we are simply mistaken in interpersonally trusting generative AI. We have succumbed to 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.[10] Notice that this point is specific to generative AI, rather than being a general claim about AI. It is because of the particular way that generative AI functions that it is not in the market for interpersonal trust. That leaves it open that future iterations of AI—such as the much heralded 'third wave' of agentic AI that (it is claimed) will supersede 'second wave' generative AI—might satisfy the conditions for interpersonal trust.[11] Our discussion clarifies what would need to be different about this future AI in order for it to be apt for this kind of trust. This is that in order to be the kind of thing that could even potentially meet our belief condition on trustworthiness, it would need to genuinely understand the informational outputs that it is producing and not merely be mimicking our cognitive behavior.

4 Trustworthiness, safety and knowledge

We noted above that both object and interpersonal entails reliance, where the former kind of trust seems to largely consist of such reliance. For both kinds of trusting, our assessment of whether they are trustworthy is centrally concerned with the conditions under which such reliance is appropriate. For example, if I am trusting the parachute to open, then I am relying on it to perform this function. How must the parachute

[8] For a fuller defence of this point, see Mitchell (2019) and Chomsky et al. (2023). For a recent influential discussion of the idea that AI reasoning is merely apparent, see Shojaee et al. (2025).

[9] Inevitably, while there is a general consensus that generative AI lacks genuine understanding of the informational outputs that it presents, there is some dissent on this score. For a very helpful survey of the recent literature on this issue, see Mitchell and Krakauer (2023).

[10] For a recent discussion of the ELIZA effect, including the early chatbot that coined this name, see Natale (2021, ch. 3).

[11] Personally, I am doubtful of this, for the simple reason that agentic AI is usually described as being an extension of generative AI. Accordingly, the issues that we have noted for generative AI might reasonably be expected to apply to this future iteration of it too. But it goes beyond the scope of this paper to argue for this claim here. For further discussion of the three waves of AI (predictive, generative, agentic), see Bruno et al. (2025).

perform in order for this reliance to be appropriate, such that the parachute is deemed trustworthy?

I suggest that what is required here is *safety*, in the sense that the outcome at issue should safely obtain. An outcome is safe when it could not easily fail to occur. In contrast, an outcome is *fragile*, and so *unsafe*, when it could easily fail to occur. To illustrate, suppose I am dangling off the side of a penthouse apartment right now, as some thrill seekers like to do. In these conditions I could easily fall to my death. All it would take is for me to lose my grip and I would be toast. The outcome of me avoiding such a fatal fall is thus unsafe. In contrast, the same outcome is safe if I am sitting in the lounge of my apartment watching TV. In such conditions, it could not easily occur that I fall to my death.

If trustworthiness demands safety, then the outcome associated with that trust could not easily fail to obtain. In particular, trustworthiness excludes fragility with regard to that outcome. Think about our parachutist. It is of the utmost importance to them, when trusting their parachute, that the outcome of it opening when required is safe. In contrast, if it could easily fail to open—if it's failing to open is a fragile possibility—then that would suffice to ensure that the parachute is not to be trusted, as it could easily send them to their death.

Or think about the driver's trust that their car will start, even on a wintery morning when they are depending on it. This trust entails that this outcome will safely obtain. This means that it couldn't easily be the case that the car fails to start. If the car is constituted such that its starting is a fragile matter—for example, if it barely manages to splutter into life each time the ignition key is turned—then it is not a trustworthy car.

One crucial thing to notice about the claim that trustworthiness entails safety is that safety is not the same as reliability, at least where the latter is straightforwardly understood in purely probabilistic terms.[12] So construed, a device is reliable provided that it can relied upon to generate the target outcome with a sufficiently high degree of probability. For example, if a parachute opens 99% of the time, then it would count as reliable in this specific sense.

Even this high degree of reliability would not be enough to ensure that the parachute is safe, however. This is because even a parachute with a 99% success rate of opening might nonetheless be so constructed that it could very easily fail to open. Imagine, for example, a design flaw in the parachute which means that it nearly fails to open each time one pulls the cord. That the probability that it would fail to open is only around 1% does not exclude the possibility that its opening is a fragile occurrence.

The same goes for the car that almost doesn't start each morning. It could well be that in general it does start, and hence that it is to this extent reliable, in the sense that there is a high probability that it will start. But if it only barely manages to start in each case, then it follows that it could very easily fail to start. And that means that

[12] The qualification is important, since reliability is not always understood in this purely statistical sense, especially in the contemporary epistemological literature. Indeed, it has been argued that there are some interpretations of reliability—i.e., where it is not given a purely statistical reading—that can entail safety. See, for example, Palermos (2015). While I'm skeptical about such claims, they are not relevant to the narrower interpretation of reliability that we are employing.

the reliability of the car is consistent with it being untrustworthy because its starting is unsafe.

One question we might ask at this juncture is why we should care about safety if the target outcome can be reliably counted on to occur? If the parachute is 99% likely to open, then why isn't that enough to ensure that it is trustworthy? Once we realize that reliability and safety can come apart in the way just described, however, then I think it becomes very clear why would care about the absence of the latter even in the presence of the former. A reliable parachute that is nonetheless unsafe is not to be trusted because the lack of safety means that there is an intolerable level of *risk* in play. Would you trust a parachute that could easily fail to open? Being assured that *in general* it will open would not be enough, even if the percentages are quite favorable. Indeed, that's why, for example, parachutes routinely have both a primary and a secondary chute. This is not because the reliability of the primary chute opening is low, but because having two chutes ensures that it could not easily occur that your chute doesn't open (as if the first failed, then the second would be triggered).

The point is that our assessments of risk are primarily concerned with how easily the target unwanted event, such as a parachute not opening, could occur. Trustworthiness demands that there is a low level of risk that the unwanted outcome fails to occur. If this event is unsafe, then there is a high level of risk in play, even when it might be the case that it is relatively unlikely that this event would occur.[13] While reliability might in general ensure that this is so, in the cases where reliability and safety come apart our judgement about the level of risk will be sensitive to the fragility in play. That trustworthiness demands safety thus entails that even if the outcome associated with that trusting has a high probability of occurring, that doesn't make such trusting appropriate if that outcome is nonetheless fragile.

We can apply this point about trustworthiness requiring safety to the specific case of trusting a source of information. When one is trusting someone or something as a source of information, one is usually hoping to acquire knowledge as a result. Crucially, however, knowledge entails safety. This means that an information source that is untrustworthy because it is unsafe won't be a source of knowledge. In particular, trusting that source of information will not be a route to knowledge.

Knowledge entails safety in the sense that, if one has knowledge, then one is forming one's belief in such a way that it couldn't be easily false.[14] Put another way, knowledge excludes what is known as *veritic luck*, which is when one's belief, so formed, could have easily been false.[15] The problem with veritic luck, as I have argued elsewhere, is that it exposes one's belief to a high degree of epistemic risk (i.e., the epistemic risk of one's belief being false).[16]

[13] Cases like this suggest that there is a modal dimension to risk. For two recent accounts of risk that are sensitive to this feature, see Pritchard (2015, 2026, *passim*) and Ebert et al. (2020).

[14] For some of the key defences of the safety condition in epistemology, including the idea that this is necessary for knowledge, see Sainsbury (1997), Sosa (1999), Williamson (2001, *passim*), and Pritchard (2005, *passim*; 2007, 2012), and Pritchard (2005, *passim*; 2007, 2012).

[15] See Pritchard (2005, *passim*). For a recent discussion of the idea that knowledge excludes veritic luck, see the exchange between Hetherington (2024a, b; Pritchard (2024a, b).

[16] See, especially, Pritchard (2016, 2020).

Just as reliability, understood purely statistically, is compatible with fragility (lack of safety), so a true belief could be formed via a reliable process—i.e., one that results in a true belief with a sufficiently high degree of probability—and yet fail to be knowledge because it isn't safe. For example, suppose you form your belief that your national lottery ticket is a loser just by reflecting on the long odds involved in your ticket being a winner. If one repeated that exercise multiple times, one would likely generate a true belief on each occasion, given the massive odds against winning a national lottery. Even so, you don't know that your ticket is a winner by forming your belief in this fashion, even if your belief is true and the basis for your belief is reliable. This is because you could very easily be wrong. In the scenario in which your lottery ticket is a winner—a scenario that could easily occur (only a few colored balls need to fall in a slightly different configuration)—then you will have a false belief.

This is why one can't come to know that one has lost the lottery by simply reflecting on the long odds involved, but one can come to know this by reading the lottery result in a respected newspaper. While the former could easily result in a false belief, the latter could not. Respected newspapers go to significant lengths to make sure that they print correct national lottery results, and hence it couldn't easily be the case that one ends up with a false belief that one has lost the lottery via this route.[17] If one wants a route to knowledge, then one needs to employ sources that generate safely true beliefs and not merely reliably true beliefs.

We have seen that both object and interpersonal trust involve reliance. Such reliance is only appropriate provided that the target outcome is safe. In terms of trusting a source of informant, this means that the source could not easily be in error. If this is an unsafe source of information, then it would not be appropriate to rely on it. Accordingly, it would not be trustworthy. Since knowledge also entails safety, sources of information that are untrustworthy because they are unsafe are also not sources of knowledge. In particular, if one's basis for one's belief is simply the information source—i.e., one is trusting that information source—then this cannot be a route to knowledge since the basis is unsafe.

5 Generative AI, while reliable, is unsafe

The relevance of these points about trustworthiness, safety and knowledge is that I will be maintaining that generative AI is an unsafe source of information. If that's right, then it cannot be a source of knowledge. More precisely, trusting generative AI as a source of information cannot be a route to knowledge.

The unsafety of generative AI is well documented, albeit not usually in these specific terms. Consider the fact that generative AI is prone to generating information outputs as factual that are completely dislocated from reality. For example, in answering a question, the generative AI may fabricate information, even to the extent of cre-

[17] If you are still unconvinced, imagine that someone you know tears up your lottery ticket on the grounds that they think it is a losing ticket. Instead of reading the results in a respected newspaper, however, their belief that it is a loser is solely the result of them reflecting on the long odds involved. Wouldn't you be appalled to hear this? As you would no doubt put it, they didn't know it was a losing ticket.

ating non-existent sources in support of the text being provided as a response to the prompt. These errors are commonly known as *hallucinations*, in that the generative AI seems to be in the grip of some sort of delusion when it produces these outputs.[18]

Interestingly, this terminology is at least sub-optimal given the nature of the phenomenon being described.[19] This is because this description seems to imply that the generative AI is itself at least temporarily convinced of the correctness of the false information that it is providing. And yet when the false information is flagged by the user, the generative AI usually immediately retracts it and acknowledges that it is false. That suggests that far from being an hallucination the faulty response instead reveals that the generative AI is employing a defective conception of what constitutes a good answer to a prompt. In particular, instead of treating the accuracy of the information being offered as a necessary ingredient of a good answer, it has instead prioritized other factors.

The hallucinations of generative AI remind us that even though it is generally very good at generating informational outputs that look like the kind of responses that an expert human might offer, the reality is that generative AI is not producing these answers in anything like the way that we do. There is a particular kind of generative AI error that makes this point especially vivid. Generative AI has become very adept at giving natural language descriptions of images. It is so effective in this regard that it is tempting to describe what it is doing as actually *seeing* the images and using what it sees to describe their contents. Of course, that's not what it is doing at all. Remember that it is trained on large data sets to be able to accurately predict what we would describe as being in these images, which is a world apart from seeing what it is in the images. This much becomes evident in a telling class of cases where generative AI gets the result wrong.

Generative AI can be very sensitive to subtle changes in the pixelation of an image. This means that when faced with two images that are indistinguishable to the human eye, it can accurately describe what is in one of the images but might have no clue—or generate a completely inaccurate response—with regard to the other image.[20] As this makes clear, generative AI isn't seeing what is in the images at all. It seems that way because its imitation of what we do when we describe the image is so effective, but when we notice the kind of errors the generative AI is inclined to make, then it becomes obvious that it is doing nothing more than being a successful mimic. The same is true for other kinds of generative AI hallucination. When generative AI formulates a sophisticated and comprehensive answer to one of our prompts, it may seem as if this output is the result of thinking, reasoning, remembering and so forth just as it would be in a comparable case of a human response. What the hallucinations remind us, however, is that in fact the underlying mechanisms leading to the generative AI output are completely different.

In particular, hallucinations reveal a crucial difference between how we operate as sources of information and how generative AI operates. Even if we restrict our

[18] For further discussion, see Sun et al. (2024) and Kalai et al. (2025).

[19] Some commentators have suggested that 'confabulation' is a better term to use. See, for example, Smith et al. (2023) and Wiest and Turnbull (2025).

[20] For further discussion of this problem, see Su et al. (2019).

attention to experts offering information within the domain of their expertise, there is still scope for someone to fail to be an epistemically good source of information. Even experts are fallible, after all, and so might get confused or mispresent crucial information. Moreover, experts can be prone to deception too. These errors are very different to the hallucinations of generative AI, however. For what hallucinations reveal is rather, as Johan Fredrikzon (2025) has termed it, "epistemic indifference." This is different from the errors that arise through fallibility or deception. In the former case, the informant is at least attempting to convey accurate information. In the latter case, the informant is conveying what they believe is inaccurate information.[21] Neither applies to hallucinations. The generative AI isn't offering the false information because it is confused, for example, as its ready willingness to concede that the information is false when challenged indicates. Relatedly, it isn't engaged in deceit either. Even if one were willing to grant that generative AI had the kind of sophisticated psychological states required for deception, such as the required motivations, the point would remain that its outputs do not bear the hallmarks of deceit.[22] For example, if deception is the aim, then why wouldn't the generative AI double-down on the faulty information when challenged rather than benignly conceding that it is false?[23]

In contrast, when generative AI hallucinates this reveals that it is on a fundamental level unconcerned about the truth of what it is presenting. It defaults to true information because it has been trained to effectively mimic how we would expertly respond to the cognitive task in hand. But it turns out that this is compatible with generating outputs that can, at times, be quite dramatically false. Once we put these two points together, then I think it becomes clearer why the informational outputs of generative AI, while generally reliable, are nonetheless unsafe. The reliability of the informational outputs is sustained by the fact that generative AI clearly treats an accurate answer as the default response. But its epistemic indifference means that it will also easily generate false answers too, in surprising ways. The reliability of generative AI's informational outputs can coexist with their fragility.

It follows from the foregoing that generative AI is untrustworthy because its informational outputs, while reliable, are not safe. Our primary target in making this claim is our object trust of generative AI, as technology that is used as an information source. Recall, however, that we have argued that both object and interpersonal trust involves reliance, and that appropriate reliance requires the safety of the target outcomes. That means that our claim that generative AI is untrustworthy because we

[21] At least usually, anyway. There are cases, for example, where someone presents information that is strictly speaking accurate in a fashion that is engineered to mislead. Plausibly, this is nonetheless a form of deception, given the intent involved.

[22] The issue of whether AI is engaged in deception gets muddled in the contemporary literature as 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, Sect. 3).

[23] Of course, one could argue that generative AI is aiming to deceive, it is just that it is very bad at it. While that's a possibility, I think we can at least say that a much more plausible explanation of what is going on is that it isn't engaged in deceit at all, especially once one factors in the previous point that genuine deceit seems to require a sophisticated psychology that generative AI appears to lack.

can't appropriately rely on it is also a further reason to treat it as interpersonally untrustworthy.[24]

6 Concluding remarks

It's important to be clear about what has been argued for and what hasn't been argued for. Two points in particular are worthy of emphasis, especially in terms of what they suggest for our use of generative AI moving forward.

The first is that it has not been claimed that AI is untrustworthy *simpliciter*, but only that the current generative version of AI is untrustworthy. That leaves it open whether future iterations of AI might be trustworthy. Indeed, our discussion has the merit of highlighting what would need to change about AI in order to make it trustworthy. In particular, if AI developers can solve the issues that make generative AI an unsafe information source, then that would enable it to be at least object trustworthy.[25]

The second point to emphasize is that in maintaining that generative AI is an unsafe, and thus untrustworthy, source of information, and hence for this reason is not a source of knowledge, it is not being claimed that we cannot use generative AI to acquire knowledge. The point is only that *trusting* generative AI is not a source of knowledge. That is, if one's only basis for endorsing information is that it comes from generative AI, then this is not a route to knowledge, given that this basis is unsafe. That doesn't preclude one from acquiring knowledge via the use of generative AI so long as one combines it with independent epistemic support.

Putting these two points together offers us some concrete guidance about how we should be employing generative AI. The first point highlights the importance of being circumspect in our use of generative AI as a source of information. The second point highlights how such circumspection, properly employed, can help us to skillfully acquire knowledge via the use of generative AI even given its epistemic flaws. If one is alert to the possibility of hallucinations, for example, and so is primed to check information and scrutinize it for errors (i.e., as opposed to merely trusting it), then one can put oneself in an epistemic situation whereby one's basis for belief is safe, and hence a potential route to knowledge, even despite the untrustworthy nature of generative AI.

Acknowledgments I am grateful to two anonymous referees for *Synthese* for comprehensive comments on an earlier version of this paper. Thanks also to Mirko Farina. This research was supported by the John Templeton Foundation ('Cultivating Intellectual Character in the AI Age', #63365). This essay was written while a Senior Research Associate of the *African Centre for Epistemology and Philosophy of Science* at the University of Johannesburg.

Data availability Not applicable.

[24] As I've argued elsewhere, the untrustworthiness of AI as a source of information explains why it is often compared to bullshit. See Pritchard (forthcoming).

[25] For a recent discussion of the trustworthiness of AI that is geared towards offering a general framework for assessing AI trustworthiness, see Simion and Kelp (2023).

Declarations

AI use AI was not used at any point in the writing of this paper.

Competing interests The author confirms that they have no relevant financial or non-financial interests to disclose.

References

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

Bruno, C., Canina, M., & Biskjaer, M. M. (2025). The three waves paradigm: Tracing the evolution of artificial intelligence in design. In G. E. Corazza (Ed.), *The cyber-creativity process: How humans co-create with artificial intelligence*. Palgrave Macmillan.

Chomsky, N., Roberts, I., & Watumul, J. (2023). The false promise of ChatGPT. *The New York Times*. https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html

Ebert, P., Durbach, I., & Smith, M. (2020). Varieties of risk. *Philosophy and Phenomenological Research*, *101*, 432–455.

Fredrikzon, J. (2025). Rethinking error: Hallucinations and epistemological indifference. *Critical AI*. https://doi.org/10.1215/2834703X-11700255

Hawley, K. (2012). *Trust: A very short introduction*. Oxford University Press.

Hawley, K. (2014). Trust, distrust and commitment. *Noûs*, *48*, 1–20.

Hawley, K. (2019). *How to be trustworthy*. Oxford University Press.

Hetherington, S. (2024a). Knowledge can include luck. In B. Roeber, M. Steup, J. Turri & E. Sosa (Eds.), *Contemporary debates in epistemology* (3rd edn., pp. 151–159). Blackwell.

Hetherington, S. (2024b). On whether knowing can include luck: Asking the correct question. In B. Roeber, M. Steup, J. Turri, & E. Sosa (Eds.), *Contemporary debates in epistemology* (3rd edn., pp. 171–173). Blackwell.

Jones, K. (1996). Trust as an affective attitude. *Ethics*, *107*, 4–25.

Kalai, A. T., Nachum, O., Vempala, S. S., & Zhang, E. (2025). Why language models hallucinate. *arXiv*. https://doi.org/10.48550/arXiv.2509.04664

Lackey, J. (2008). *Learning from words: Testimony as a source of knowledge*. Oxford University Press.

Mahon, J. E. (2015). The definition of lying and deception. In E. Zalta (Ed.), *Stanford encyclopedia of philosophy*. https://plato.stanford.edu/entries/lying-definition/

McLeod, C. (2020). Trust. In E. N. Zalta (Ed.), *Stanford encyclopedia of philosophy*. https://plato.stanford.edu/entries/trust/

Mitchell, M. (2019). Artificial intelligence hits the barrier of meaning. *Information*, *10*, 51.

Mitchell, M., & Krakauer, D. C. (2023). The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences*, *28*, 120.

Natale, S. (2021). *Deceitful media: Artificial intelligence and social life after the turing test*. Oxford University Press.

Palermos, S. O. (2015). Could reliability naturally imply safety? *European Journal of Philosophy*, *23*, 1192–1208.

Potter, N. N. (2002). *How can I be trusted? A virtue theory of trustworthiness*. Rowman & Littlefield.

Pritchard, D. H. (2005). *Epistemic luck*. Oxford University Press.

Pritchard, D. H. (2007). Anti-luck epistemology. *Synthese*, *158*, 277–297.
Pritchard, D. H. (2012). Anti-luck virtue epistemology. *Journal of Philosophy*, *109*, 247–279.
Pritchard, D. H. (2015). Risk. *Metaphilosophy*, *46*, 436–461.
Pritchard, D. H. (2016). Epistemic risk. *Journal of Philosophy*, *113*, 550–571.
Pritchard, D. H. (2020). Anti-risk virtue epistemology. In J. Greco, & C. Kelp (Eds.), *Virtue-theoretic epistemology: New methods and approaches*. Cambridge University Press.
Pritchard, D. H. (2024a). There cannot be lucky knowledge. In B. Roeber, M. Steup, J. Turri, & E. Sosa (Eds.), *Contemporary debates in epistemology* (3rd edn., pp. 159–168). Blackwell.
Pritchard, D. H. (2024b). Reply to Hetherington. In B. Roeber, M. Steup, J. Turri, & E. Sosa (Eds.), *Contemporary debates in epistemology* (3rd edn., pp. 171–173). Blackwell.
Pritchard, D. H. (2026). *Tempting fate: A philosophical guide to risk, luck, and a meaningful life*. Princeton University Press.
Pritchard, D. H. (Forthcoming). AI and bullshit. *Human Affairs: Postdisciplinary Humanities & Social Sciences Quarterly*.
Sainsbury, R. M. (1997). Easy possibilities. *Philosophy and Phenomenological Research*, *57*, 907–919.
Shojaee, P., Mirzadeh, I., Alizadeh, K., Horton, M., Bengio, S., & Farajtabar, M. (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
Simion, M., & Kelp, C. (2023). Trustworthy artificial intelligence. *Asian Journal of Philosophy*. https://doi.org/10.1007/s44204-023-00063-5
Simon, J. (2025). Generative AI, quadruple deception and trust. *Social Epistemology*. https://doi.org/10.1080/02691728.2025.2491087
Smith, A. L., Greaves, F., & Panch, T. (2023). Hallucination or confabulation? Neuroanatomy as metaphor in large language models. *PLOS Digital Health*. https://doi.org/10.1371/journal.pdig.0000388
Sosa, E. (1999). How to defeat opposition to Moore. *Philosophical Perspectives*, *13*, 141–154.
Su, J., Vargas, D. V., & Sakurai, K. (2019). One pixel attack for fooling deep neural networks. *arXiv*. https://arxiv.org/pdf/1710.08864
Sun, Y., Sheng, D., Zhou, Z., & Wu, Y. (2024). AI hallucination: Towards a comprehensive classification of distorted information in artificial intelligence-generated content. *Humanities and Social Sciences Communications*, *11*, 1278.
Wiest, G., & Turnbull, O. H. (2025). Faulty artificial intelligence, or the sleep of reason. *New England Journal of Medicine AI*. https://doi.org/10.1056/AIp2500785
Williamson, T. (2001). *Knowledge and its limits*. Oxford University Press.