

Can We Epistemically Trust Conversational AI Testimony?

Parker DuVall

Purdue University
duvall30@purdue.edu

ABSTRACT

Why do we distrust conversational AI (CAI) testifiers? That is to say, why do many have the intuition that CAI are less epistemically trustworthy than a standard human testifier? While it seems like we have such a pre-theoretical intuition, we are often lacking a philosophical defense of this attitude. In this paper, I argue that testimony is more trustworthy if it is based in a first-person experience of consciousness. I start by laying out an account of CAI testimony, including considering what it would mean for a CAI system to have knowledge. I then use a number of pragmatic encroachment cases to show that when the stakes are high and we have to choose between two instances of testimony to ground our belief about *p*, we always prefer the piece of testimony from the speaker with a first-person experience of *p*. Since CAI systems cannot, and possibly never will, have a first-person experience of consciousness (i.e., there is nothing it is “like” to be a CAI), we will always view CAI testimony as less trustworthy. Finally, I consider and respond to a number of objections to my arguments.

Introduction

One criticism of conversational artificial intelligence (CAI) is that these systems are not the kinds of entities that can be the proper objects of trust relations and thus should not have important or large-scale uses. Such critics may see value in having a CAI create a funny image, but may be opposed to a CAI system, for instance, teaching schoolchildren or managing large scale finances. If we believe that CAI is not the kind of technology that can be trusted, in principle, in these kinds of important and serious use-cases, but is nevertheless being implemented in them, we may develop a distrusting attitude towards CAI. In some cases, critics of the technology hold the intuition that CAI is epistemically untrustworthy, but this intuition is often philosophically unsubstantiated. To give an argument in favor of this intuition, in this paper, I argue that because current CAI testimony cannot have phenomenologically based justification, it is less epistemically reliable and thus can be deemed less epistemically trustworthy. That is, because a CAI cannot have a first-person experience of consciousness, its testimony to the truth of a proposition can be understood as less conducive to the formation of true beliefs in the hearer of the testimony.

In the next section, I outline the account of testimony that I will use going forward in this paper. I start by making some preliminary claims on testimony and current debates. Then, I outline what it would mean for a CAI system to “know”, including Ocean Cangelosi’s (2024) argument for non-phenomenologically based knowledge (or, knowledge lacking a first-person experience). Next, I explore Ori Freiman’s (2024) account of CAI-specific testimony. I argue that CAI can, via Freiman’s account of testimony and Cangelosi’s account of non-phenomenological knowledge, be understood as both a generator and a transmitter of knowledge, setting the groundwork for section 3.

In section 3, I start by considering what it would be like for a human to reason like a current CAI. From this, we see that the common reasoning processes used in many CAI models are too different from those of a human, so we instead turn our focus to higher-end reasoning models. Then, I use four high-stakes cases of knowledge attribution in which we would be compelled to believe the testifier with phenomenologically justified belief over the testifier without it. For instance, in the first case, our intuition tells us to base our belief on the testimony of the passerby who had vivid and direct visual perception, rather than the passerby who had peripheral visual perception. These cases show that the kind of testimony that currently only humans can give can be understood broadly as more epistemically trustworthy than the only kind of testimony that current CAI systems can give, even when the cases control for distrusting attitudes towards CAI.

In section 4, I explore and respond to two additional objections. First, one could claim that intuition about trust is too variable of a notion to use for the arguments here, as intuitions about what testifier is more epistemically trustworthy (i.e., who can reliably lead us to true beliefs) can be tainted by non-epistemic attitudes (i.e., any attitude not related to the ability of the testifier to reliably lead us to true beliefs). For instance, the intuition of many living in 17th-century England was to deem “gentlemen” more epistemically trustworthy, simply in virtue of their status as a gentleman (Shapin 2011). My response aims to reassert that the distrust towards the testifier in the presented cases tracks the reliability of the source of justification for the testifier’s belief (e.g., direct perception vs. inference). That is to say, the hearer’s distrust is aimed at a specific proposition put forth by the testifier. The hearer will distrust this testimony if they judge the testifier to be an unreliable source of testimony, making it unlikely that their statement will generate a true belief. In this way, the kind of trust used here is a uniquely epistemic one (i.e., a hearer epistemically trusts a testifier whom they deem to be able to reliably lead them to true beliefs). Moreover, the cases presented account for possible non-epistemic distrusting attitudes towards CAI (e.g., a predisposition to think CAI systems are untrustworthy, epistemically or otherwise, in virtue of a negative opinion of their proprietor(s)) and still show that the testimony from these systems is less trustworthy.

Second, I respond to the objection that the arguments presented here place too much emphasis on phenomenological experience as an epistemic justifier, and in doing so may threaten the justification of scientific discoveries that we do not have phenomenological experience of (e.g., unobservable theoretical entities). I respond by distinguishing between everyday testimonial exchanges and these kinds of scientific discoveries as instances where phenomenological experience is and is not possible respectively. As such, I claim that we should not fault an instance of knowledge that, in principle, cannot have any phenomenological experience behind it. Consequently, we can safely favor phenomenological experience in the kinds of testimonial exchanges presented here while keeping the justification of our scientific knowledge safe.

Building Up the Testimony

Given the broad literature on the epistemology of testimony,¹ this paper focuses on the basic debates and definitions, such as Lackey (1992; 2006) or Coady (1992), necessary to get the arguments presented here off the ground. While a comprehensive account of testimony and AI testimony is beyond the scope of this paper, it is necessary to clarify the central background assumptions that underpin the arguments presented here. Henceforth, the term testimony should be taken to mean an act of communication, usually but not always through speech, in which a speaker reasonably intends to convey information that *p* to a hearer.² This is a rough outline of testimony which glosses over many important debates in the literature, but will do us justice for getting the groundwork of this paper laid. This is importantly different from court testimony under a formal oath, which is often called “formal testimony” in the literature. With this in mind, we can easily imagine numerous cases of testimony. For instance, we can imagine being at a train station and stopping a passerby to ask, “When does the train to Baltimore leave?” We could also imagine learning a fact from an expert that we would never be able to figure out a priori and would likely never experience first-hand. For instance, the fact that the Burj Khalifa, a skyscraper in Dubai, is 2,722 feet tall.

One final piece is that there exists a debate over whether testimony can only transfer knowledge or if it can generate it as well.³ In the first case, the speaker has knowledge that *p* and transfers it when they testify that *p*. In the second case, the speaker does not have knowledge that *p* and thus when they testify that *p*, the knowledge that the hearer gains

¹ A non-extensive but influential list of articles in the literature includes Fricker (1987; 2006), Coady (1992), Faulkner (1998), Lackey (1999; 2006), and Elgin (2002).

² Based largely on Lackey’s (2006) rough approximation.

³ See Audi (1997), Adler (1994, 1996), Graham (2000), Owens (2000), Goldberg (2005), Lackey (1999, 2008), Pelling (2013), and Wright (2015).

is generated rather than transferred. Jennifer Lackey shows this feature of testimony with her “Creationist Teacher” case:

Suppose that a Catholic elementary school requires that all teachers include sections on evolutionary theory in their science classes and that the teachers conceal their own personal beliefs regarding this subject-matter. Mrs Smith, a teacher at the school in question, goes to the library, researches this literature from reliable sources, and on this basis develops a set of reliable lecture notes from which she will teach the material to her students. Despite this, however, Mrs Smith is herself a devout creationist and hence does not believe that evolutionary theory is true, but she none the less follows the requirement to teach the theory to her students. Now assuming that evolutionary theory is true, in this case it seems reasonable to assume that Mrs Smith's students can come to have knowledge via her testimony, despite the fact that she fails condition (ii) and hence does not have the knowledge in question herself. That is, it seems that she can give to her students what she does not have herself. For in spite of Mrs Smith's failure to believe and therewith to know the propositions she is reporting to her students about evolution, she is a reliable testifier for this information, and on the basis of her testimony it seems that the students in question can come to have knowledge of evolutionary theory. (Lackey 1999, 477)

Since we can conceive of CAI testimony in both cases (as we will see below), we need not take a hard stance on this debate here.

As technology advances, it may seem more reasonable to view CAI, or other artifacts, as testifiers in their own right. Nevertheless, mainstream conceptions of testimony tend to be wed to the idea that only persons can testify (Coady 1992, 268; Lackey 2008, 189). On the other hand, we do, at the very least, acquire beliefs on the basis of information we garner from artifacts. For instance, we can acquire knowledge from indicators, such as a fuel gauge, which can support the belief that my fuel levels are low.⁴ Moreover, we can form beliefs from artifacts that mediate testimony. For instance, if a friend sends a text to my phone, I can have my phone read the message aloud. Thus, even if the artifact is speaking (albeit in a metaphorical sense), it is merely mediating the testimony from my friend. In more recent times, though, it may appear that artifacts are testifying in some sense. Thus, while mainstream accounts deny that CAI systems can testify, it may be useful to have a notion of testimony that is specific to CAI systems. We can understand CAI systems as being able to testify both in the standard transitive sense as well in a generative manner as outlined by Lackey. Such a concept, however, requires that we first outline what it would mean for a CAI system to possess knowledge as well as what a CAI-specific account of testimony would entail.

⁴ Thank you to an anonymous reviewer for this example.

AI Knowledge

The biggest obstacle to a CAI system having knowledge would be its lack of capacity for phenomenological experience (henceforth “PE”). If we consider the common conception of knowledge as a justified, true, and non-lucky belief, it seems *prima facie* possible that these are all possible without PE. A belief being true, at least on an externalist account, does not have any requirement for PE. A belief’s truth being non-lucky does not seem to require any kind of PE from the believer either. Belief and justification then would be the final two obstacles. These, similarly, may not strictly require (though will often feature) PE. For instance, I can hold beliefs that I am not currently attending to or have never actually attended to. We can imagine learning someone’s name is “Sam” and never consciously thinking “This person’s name is Sam” and yet, we do hold such a belief and could come to attend to it if needed. Justification may be said to require PE, but assuming we believe that beliefs can justify other beliefs, this need not require PE either. The CAI testifiers being able to know, however, will be important for the later parts of this paper, so we will want to use a more laid-out account.

In her paper “Can AI Know?”, Ocean Cangelosi puts forth an account of non-phenomenological propositional knowledge. Cangelosi argues that not only do we have many beliefs that have no phenomenological qualities (or, a salient “what-it’s-likeness”), but belief acquisition does not seem to require PE either (Cangelosi 2024, 81). To the first point, Cangelosi uses the example of someone entering a car and reaching for the driver’s side door, thus possessing the belief that the door is on their left. They may never actually have the time to think “the door is on my left.” Furthermore, the action of reaching out and closing the door seemingly amounts to automatic behavior that lacks salient PE, especially when one is engaged in other activities like a phone call, conversation with a passenger, or setting up navigation on their phone (Cangelosi 2024, 81). It seems likely as well to say that we could acquire perceptual beliefs without PE. Cangelosi uses the example of running into a building on a rainy day. If we have the intention to enter a building for the purpose of seeking shelter from the rain, we have formed the belief that the building has a roof. If we then assume that we have no epistemic access to direct perceptual justification that there is a roof, we may still have various clues based on what the building looks like. However, these clues can be easily absorbed into other tasks; thus, our cues that there really is a roof lack this salient “what-it’s-likeness”. For instance, if someone is engaged in a phone call (as previously mentioned), or otherwise could be considered “mentally preoccupied”, it is not hard to imagine that the cues could lack a salient “what-it’s-likeness” and thus be absorbed. One further example of this is peripheral vision. As we are focused on something else, such as a phone call or our actual destination, the peripheral vision cues that we take in often lack the salient “what-it’s-

likeness” of PE. As Cangelosi puts it, “it is possible to tune out perception without tuning out perceptual beliefs” (Cangelosi 2024, 81).

We may say, then, that even if belief acquisition or possession does not necessitate PE, justification must require it. As some, such as experiential evidentialists, may argue, to be justified in a belief, you must possess not only sufficient evidence for the belief but the belief itself must also be based on said evidence (often called the basing relation) (Lyons 2009). As Cangelosi points out, however, this kind of connection does not require PE, as we already know that belief acquisition and possession do not require PE, and we have seen that we can form non-phenomenological beliefs on the basis of non-phenomenological cues (Cangelosi 2024, 79-81).

AI-Specific Testimony

In trying to understand what it would mean for a CAI system to testify, we may try to plug in a CAI system (i.e., an AI system powered by large language models designed for natural language conversation) into the role of testifier within an account of testimony such as that of Fricker (1995), Lackey (2008), or Graham (1997). Ori Freiman (2024), however, argues that our mainstream accounts of testimony are anthropocentric. Even the previous, albeit rough, account of testimony that was given previously in this paper uses the word “intends”, which may not apply to a non-phenomenologically capable artifact such as a CAI system. It would be helpful, then, to have a uniquely non-anthropocentric conception of AI testimony. Freiman’s account requires that the testimony be (a) delivered in propositions, (b) generated and delivered with no other human directly involved, and (c) contain an output that is perceived and phenomenally similar to that of a human (Freiman 2024, 13).

The first requirement, propositional content as the sole delivery method of the testimony, is rather straightforward. However, there are two strengths of demand that this could require. The stronger demand requires the propositions to be non-inferential propositions, such as a robot making a “stop” hand gesture, causing the hearer/viewer to infer a proposition akin to “You need to stop!”, while the weaker demand allows for such propositions. We can use the stronger demand, although it does not make a difference for the arguments presented here.

It is not enough for AI testimony to merely be propositional. Instead, the propositions delivered must be generated and delivered without human involvement (Freiman 2024, 14). In doing so, Freiman safeguards against two cases: human-generated propositions from artificial speakers and devices that merely mediate human testimony (Freiman 2024, 14). In the first case, we could imagine announcements of train times at a train station or a defibrillator that audibly gives the vocal order of “attach the sticky pads to the patient’s skin” (Freiman 2024, 14). In the second case, we could imagine a phone that

reads aloud a text sent to you that reads “Coming home soon!”. The phone is not testifying that it is coming soon, obviously. Rather, it is mediating the testimony of the human sender. No cases of these types meet our intuitions for what testimony is or should be considered. For instance, we would not say that someone being forced to say that p is true is testifying, at least in the proper and intuitive sense, to the truth of p .

The final requirement for AI testimony is that the output (i.e., the testimony) be phenomenologically similar to that of a human, or at least be perceived as such (Freiman 2024, 16). For instance, a text from a CAI system is going to be phenomenologically similar to a text from a human being, so it could qualify as AI testimony. That is, the first-person experience of receiving a text from a human is the same as receiving a text from a CAI.

With these two accounts of testimony for a CAI system in mind, we can start to build cases of CAI systems having less trustworthy testimony than that of humans. We can either view AI testimony as the possible transmission of non-phenomenological knowledge or generation of knowledge. The cases below, however, have been formatted with the assumption of granting Cangelosi’s (2024) arguments.

Reasoning and Testimony Cases

One reason to point to in order to illustrate why individuals may be distrusting of CAIs is the reasoning gap between humans and CAI systems. For instance, imagine if a person somehow froze time for themselves and read every available piece of text on the internet. Upon “un-freezing” time, you ask them a question about proposition p , and they respond “Well, based on your question, I can predict one word at a time with my extreme set of data. The answer to your question should be q ”. We probably would not be quick to trust this person. What we want, then, is a testifier who can reflect on the evidence available, and can reason much in the ways that humans do. While this is not characteristic of all CAI systems, there exists a class of reasoning models that seem to be able to reflect on how they come to certain conclusions and outputs. For instance, AI companies such as DeepSeek or Open AI have developed CAI systems that show their reasoning in their response by referring back to what they “know”, what they have been told, and what their training contains (OpenAI 2024). For instance, such a model, when presented with the question “Who is the first president of the United States?”, may provide a reasoning pathway that shows it “thinking” that, “Okay, I know that I have to try to present the truth, so the answer is George Washington. But, I also have been trained to give important details when possible, so perhaps I should also add that George Washington was president of the United States from 1789 all the way until 1797”. So, we can stipulate that for the following cases, all mentioned CAI systems are capable of such levels of reasoning. Thus,

these reasoning language models are at least *prima facie* more reliable or trustworthy than traditional large language models.

Two final preliminaries are in order. First, the following are all cases of “pragmatic encroachment”, or instances in which the practical stakes “encroach” on the epistemic. This means that these are all cases in which the high stakes of the specific situation seems to make it extra important to “get it right”. As a result, we often find our intuitions leading us towards more reliable pieces of justification. That is, we often require more reliable evidence in defense of *p*’s truth. As an example, consider the two cases that Stanley (2005) presents in his book *Knowledge and Practical Interests*:

Low Stakes. Hannah and her wife Sarah are driving home on a Friday afternoon. They plan to stop at the bank on the way home to deposit their paychecks. It is not important that they do so, as they have no impending bills. But as they drive past the bank, they notice that the lines inside are very long, as they often are on Friday afternoons. Realizing that it wasn’t very important that their paychecks are deposited right away, Hannah says, “I know the bank will be open tomorrow, since I was there just two weeks ago on Saturday morning. So we can deposit our paychecks tomorrow morning”.

High Stakes. Hannah and her wife Sarah are driving home on a Friday afternoon. They plan to stop at the bank on the way home to deposit their paychecks. Since they have an impending bill coming due, and very little in their account, it is very important that they deposit their paychecks by Saturday. Hannah notes that she was at the bank two weeks before on a Saturday morning, and it was open. But, as Sarah points out, banks do change their hours. Hannah says, “I guess you’re right. I don’t know that the bank will be open tomorrow”. (Stanley 2005, 3–4)

Both cases are evidentially similar, but we nevertheless tend to agree with a knowledge attribution in the Low Stakes case but agree in withholding knowledge attribution in the High Stakes case. While an extensive overview of pragmatic encroachment theory is not necessary for our purposes, the broad moral of the arguments presented in its favor is that the more important it is that we get *p* right, the harder it is to know that *p*.

The second preliminary concerns the use of the term “trust”. While some, such as Pepijn Al (2022), have argued that “trust” as it is commonly used cannot be properly applied to artificial intelligence and others, such as Ori Freiman (2022) or Baron (2025), have argued that trusts of certain kinds can be applied to artificial intelligence, the kind of trust used here is an epistemic trust. Epistemic trust, as it is considered here, refers to the ability of an individual to reliably bring another to true beliefs. So, if someone is an expert on a subject, their testimony would be trustworthy, in this sense, with regards to that specific subject. Henceforth, uses of the word “trust” refer to this kind of uniquely epistemic trust.

Phenomenological vs. Non-Phenomenological Testimony Cases

To show that phenomenologically based testimony (henceforth PBT) is at least *prima facie* more reliable than non-phenomenologically based testimony (henceforth NPBT) consider the following case:

Case 1:⁵ Rushed and in a panic, you cannot seem to locate the train station's train schedule. In an attempt to figure out when the train leaves, you stop two individuals, both of which present you with no epistemic defeaters such that you should automatically not trust them. To each one you ask "What time does the train to Baltimore leave?". The first passerby confidently says "I just looked at the train schedule; the train to Baltimore leaves at 5". The second passerby confidently says "I saw the train schedule out of the corner of my eye while walking; the train to Baltimore leaves at 4".

It seems that, all else being equal, we would prefer the testimony of the first passerby. This is because we deem the belief that the second passerby testifies to be less reliable than that of the first passerby. The difference between these two beliefs is that the first is justified with PE, while the second is not. Using Cangleosi's example of peripheral vision as a possible non-phenomenological belief-forming process, we can imagine that the second passerby's belief could have been formed without any phenomenological experience behind it. The non-phenomenologically justified belief is seemingly less justified in part because, by its very nature, it can be absorbed into other tasks and thus less trustworthy. We can see a similar effect if we insert a CAI system into this case.

Case 2: You are the parent of a child that has been kidnapped. You rush to the train station with the knowledge that if you miss the only train to Baltimore, your child will die. Rushed and in a panic, you cannot seem to locate the train station's train schedule. In an attempt to figure out when the train leaves, you stop a passerby and ask "What time does the train to Baltimore leave?". The passerby confidently says "I just looked at the train schedule; the train to Baltimore leaves at 5". You notice that there are no other passerby around. Looking to be sure you have the schedule correct, you prompt a CAI system "What time does the train to Baltimore leave?". The chatbot responds with an answer of "The train to Baltimore leaves at 6".

In this case, we may still stake our belief, and thus attribute knowledge, to the first passerby. The CAI is generally reliable, but we know that the CAI is essentially predicting the words to respond back to me, and very well could be wrong. Furthermore, we may be worried about the CAI system pulling inputs from different train schedules that do not match today's. Of course, distrusting attitudes against CAI are not entirely uncommon, so we should imagine a case where the CAI seems to have the upper hand.

⁵ While not strictly necessary, I use trains here as an homage to existing pragmatic encroachment literature.

Case 3: You are the parent of a child that has been kidnapped. You rush to the train station with the knowledge that if you miss the only train to Baltimore, your child will die. Rushed and in a panic, you cannot seem to locate the train station's train schedule. You come across two passerby. The first is a human who, when asked "When does the train to Baltimore leave?", says "I saw the train schedule out of the corner of my eye while walking; the train to Baltimore leaves at 4". The second passerby is an incredibly advanced CAI system who, to your knowledge, is able to take in perceptual evidence, reflect on it, and weigh it against other pieces of evidence it may have. This CAI system is, essentially, entirely equal in intellect, cognitive abilities, and perceptual abilities to a human. When you ask this CAI passerby "What time does the train to Baltimore leave?", the CAI passerby confidently says "I just looked at the train schedule; the train to Baltimore leaves at 5."

Keeping in mind that this technology does not (and perhaps will never) exist, we can imagine that such a CAI system would, in this instance, be a more epistemically trustworthy testifier. We would not only be compelled to say that the CAI passerby knows, but it seems that we would also be in violation of our epistemic obligations to base our beliefs on the best evidence available if we said that the human passerby knows but the CAI passerby does not. "Violation of our epistemic obligations" simply means that we are presented with two possible justifications and choose the *prima facie* weaker one due to non-epistemic factors. For instance, if a misogynist man was dying and decided to ignore the advice of his female doctor in favor of the advice of his male neighbor who has no formal medical experience or training, we would say that he violated his epistemic obligations to base his belief on the best evidence available. Assuming the CAI system is capable of phenomenological experience, its testimony about the train schedule is more trustworthy than the human testifier. This is the same dynamic we saw in case 2, in which the human with PBT is deemed to be more epistemically trustworthy as a testifier than the CAI system with NPBT. Finally, we can imagine a case in which we are forced to base our belief on either the testimony of the CAI system from Case 2 or Case 3.

Case 4: You are the parent of a child that has been kidnapped. You rush to the train station with the knowledge that if you miss the only train to Baltimore, your child will die. Rushed and in a panic, you cannot seem to locate the train station's train schedule. You come across an incredibly advanced CAI system who, to your knowledge, is able to take in perceptual evidence, reflect on it, and weigh it against other pieces of evidence it may have. This CAI system is, essentially, entirely equal in intellect, cognitive abilities, and perceptual abilities to a human. When you ask this CAI passerby "What time does the train to Baltimore leave?", the CAI passerby confidently says "I just looked at the train schedule; the train to Baltimore leaves at 5". Looking to be sure you have the schedule correct, you pull out your phone and prompt an CAI system, which does not have the capacity

for phenomenological experience, “What time does the train to Baltimore leave?”. The chatbot responds “The train to Baltimore leaves at 6”.

It would appear that even if we are inclined to say “Well, I would hope for a third testifier to base my belief on instead of two bots”, when we are truly out of other options, we would prefer the testimony of the advanced CAI system. While there could be sentiments of CAI distrust in other cases, thus swaying us towards the human testifier, this example shows that the PBT is always preferable, *ceteris paribus*.

It seems that in a high-stakes knowledge attribution case, we are more inclined to say that the entity with PBT does know, as opposed to the entity with NPBT. While the entity with NPBT may end up actually knowing, why do we trust the individual with PBT more? In the second case, we may find ourselves trusting the human more simply because of a predisposition to distrust CAI systems. That being said, this disposition is not available as an explanation in cases one, three, and four. Specifically, in case three, not trusting the CAI system would amount to us violating our epistemic obligations, rather than tracking the quality of the testimony. It seems that the entity with PBT is more justified in their assertion, and thus it seems more likely that their belief will amount to knowledge. The entity without PBT, on the other hand, seems to have justification that is on, as it were, shaky ground. As a result, PE appears to make testimony more trustworthy, *ceteris paribus*.

Our intuitions that the NPBT is less reliable allow us to examine the second case a bit closer. This case, in which the human testimony is phenomenologically based and the CAI testimony is not, models the current state of CAI systems. These CAI systems, such as Deepseek’s R1 or OpenAI’s ChatGPT-4o, are without a doubt incredibly complex and do seem to have use cases. With that being said, if our intuitions about the value of the testimony of these systems are correct, we have at least one reason to think that these use cases should not be serious ones. That is to say, the testimony of these systems should not be used for teaching future doctors, lawyers, or engineers, advising federal spending, or any case in which the failure of the testimony results in widespread damage or distrust in our institutions.

Possible Objections and Responses

One could object on the grounds that the arguments presented in this paper are largely based on an idea of trustworthy testimony, and trust is a fickle and often subjective matter. For instance, if one were deemed a “gentleman” in 17th-century England, many were quick to trust such a person in testimonial exchanges as a gentleman’s “word was his bond” (Shapin 2011, 65). Trust, however, may not be actually “tracking” any kind of quality or value in an instance of testimony. Perhaps, then, we should be wary of writing

off CAI based on these examples. After all, there seems to be tremendous upside to the technology, and writing it off too early could stunt or completely stop its growth, thus limiting us from reaping the benefits these CAI systems could bring us in the near or distant future.

Trust and distrust often play a role in testimonial exchanges. With that being said, it is not something that the cases or arguments presented above cannot account for. If we are worried about a case in which the hearer has an underlying disposition to distrust CAI, we can imagine a case where the hearer is ambivalent, overall trustworthy of CAI, or cannot tell the CAI from the human. For example, imagine the previous train cases except there is no passerby to ask when the train to Baltimore leaves. Rather than take a gamble, you decide to text two friends who may know the answer. The first friend responds that the train should be leaving at 5. The second friend responds that the train should be leaving at 4. Now assume that you know that the first friend regularly takes that train to Baltimore, while the second friend, who also regularly takes the train to Baltimore, has a habit of having their favorite CAI system answer questions he is too lazy to answer himself. It would appear that we would be inclined to trust the testimony of the first friend, partially because the first friend has the experience to back up his belief and we have good reason to think the testimony is solely his. While we have no epistemic access to whether or not the testimony from the second friend was truly theirs or that of an CAI system, the fact that the testimony could be coming from a testifier incapable of PE is enough to sway us in favor of the testimony of the first friend. If we are worried that the mere mention of CAI systems is enough to activate distrust and sway our intuitions, then we need not look any further than the first train case to see that PBT is epistemically favored. What the testimony cases are tracking is our trust in the various instances of testimony, which we can cash out in terms of reliability. We believe that the testifier with the PBT is more trustworthy, as we believe it a more reliable source of evidence for our beliefs.

Another objection may point out that the arguments presented here seem to put too much faith in PE as an epistemic justifier. As a result, from the views presented here, it may follow that many of our important scientific discoveries (such as dark matter), fundamental notions (such as electrons), or any kind of unobservables (such as the force of gravity) that we do not have PE of are “less justified”.⁶ Such an objection is not unfounded, but can nevertheless be avoided by making a distinction between the kinds of testimonial exchanges discussed here and the kinds of scientific endeavors and discoveries that may be threatened if this objection were to stand. This distinction occurs between knowledge that does have the possibility for phenomenological experience and

⁶ Thank you to an anonymous reviewer for this objection.

that which does not. Everyday testimonial exchanges will almost always fall in the former category, while many scientific discoveries fall into the latter category. Any kind of knowledge that falls into the latter category should not be faulted for not having any possibility for PE. For instance, it would be improper to say that the discovery of dark matter is less justified in virtue of the lack of PE, as dark matter cannot interact with light and thus cannot be seen directly. On the other hand, in everyday testimonial exchanges, such as the train cases presented above, it seems perfectly fine to regard them as more justified when they are accompanied by PE, as there is an actual possibility for PE. As such, we can safely favor PE in everyday testimonial exchanges while retaining the strong justification of our scientific discoveries.⁷

One final objection that one could take is that my arguments show nothing over and above what is shown by the current output errors (often called “hallucinations”) of CAI models (Erdem et al. 2025 and Chakraborty et al. 2025 both provide interesting overviews of the hallucination-rate problem). I would respond, however, that humans too lie about our sources, and yet the popular intuition still seems to be that humans are *prima facie* more reliable than AI systems. We could say that this intuition is tracking something and that it should not be brushed off entirely. Namely, it is tracking a genuine justification gap between testifiers.

References

- Adler, Jonathan E. “Testimony, trust, knowing.” *Journal of Philosophy*, vol. 91, no. 5, 1994, pp. 264–275, <https://doi.org/10.2307/2940754>.
- Adler, Jonathan E. “Transmitting knowledge.” *Noûs*, vol. 30, no. 1, Mar. 1996, p. 99, <https://doi.org/10.2307/2216305>.
- Al, Pepijn. “(E)-Trust and Its Function: Why We Shouldn’t Apply Trust and Trustworthiness to Human–AI Relations.” *Journal of Applied Philosophy*, vol. 40, no. 1, 24 Aug. 2022, pp. 95–108, <https://doi.org/10.1111/japp.12613>.
- Audi, Robert. “The Place of Testimony in the Fabric of Knowledge and Justification.” *American Philosophical Quarterly*, vol. 34, no. 4, 1997, pp. 405–22. JSTOR, <http://www.jstor.org/stable/20009910>. Accessed 14 Mar. 2025.
- Baron, Sam. “Trust, explainability and ai.” *Philosophy & Technology*, vol. 38, no. 1, 8 Jan. 2025, <https://doi.org/10.1007/s13347-024-00837-6>.
- Cangelosi, Ocean. “Can ai know?” *Philosophy & Technology*, vol. 37, no. 3, 1 July 2024, <https://doi.org/10.1007/s13347-024-00776-2>.
- Chakraborty, Neeloy, et al. “Hallucination detection in foundation models for decision-making: A flexible definition and review of the state of the art.” *ACM Computing Surveys*, vol. 57, no. 7, 5 Mar. 2025, pp. 1–35, <https://doi.org/10.1145/3716846>.
- Coady, C. A. J. *Testimony: A Philosophical Study*. Clarendon; Oxford University Press, 1992.

⁷ An earlier draft of this paper was presented at the University of South Carolina, and I want to thank the many attendees for their lively discussion, comments, and objections.

- Elgin, Catherine. "Take it from me: The epistemological status of testimony." *Philosophy and Phenomenological Research*, vol. 65, no. 2, Sept. 2002, pp. 291–308, <https://doi.org/10.1111/j.1933-1592.2002.tb00203.x>.
- Erdem, Orhan, et al. "Hallucination in AI-generated financial literature reviews: Evaluating bibliographic accuracy." *International Journal of Data Science and Analytics*, 15 Feb. 2025, <https://doi.org/10.1007/s41060-025-00731-0>.
- Faulkner, Paul. "David Hume's reductionist epistemology of testimony." *Pacific Philosophical Quarterly*, vol. 79, no. 4, Dec. 1998, pp. 302–313, <https://doi.org/10.1111/1468-0114.00065>.
- Freiman, Ori. "AI-Testimony, Conversational AIS and Our Anthropocentric Theory of Testimony." *Social Epistemology*, vol. 38, no. 4, 6 Mar. 2024, pp. 476–490, <https://doi.org/10.1080/02691728.2024.2316622>.
- Freiman, Ori. "Making sense of the conceptual nonsense 'trustworthy AI.'" *AI and Ethics*, vol. 3, no. 4, 28 Nov. 2022, pp. 1351–1360, <https://doi.org/10.1007/s43681-022-00241-w>.
- Fricker, Elizabeth. "Critical notice." *Mind*, vol. 104, no. 414, 1995, pp. 393–411, <https://doi.org/10.1093/mind/104.414.393>.
- Fricker, Elizabeth. "Second-hand knowledge*." *Philosophy and Phenomenological Research*, vol. 73, no. 3, Nov. 2006, pp. 592–618, <https://doi.org/10.1111/j.1933-1592.2006.tb00550.x>.
- Fricker, Elizabeth. "The Epistemology of Testimony." *Aristotelian Society Supplementary Volume*, vol. 61, no. 1, 1 July 1987, pp. 57–106, <https://doi.org/10.1093/aristoteliansupp/61.1.57>.
- Goldberg, Sanford. "Testimonial knowledge through unsafe testimony." *Analysis*, vol. 65, no. 288, Oct. 2005, pp. 302–311, <https://doi.org/10.1111/j.1467-8284.2005.00570.x>.
- Graham, Peter J. "Transferring knowledge." *Noûs*, vol. 34, no. 1, Mar. 2000, pp. 131–152, <https://doi.org/10.1111/0029-4624.00205>.
- Graham, Peter J. "What is testimony?" *The Philosophical Quarterly*, vol. 47, no. 187, Apr. 1997, pp. 227–232, <https://doi.org/10.1111/1467-9213.00057>.
- Lackey, Jennifer. "It Takes Two to Tango: Beyond Reductionism and Non-Reductionism in the Epistemology of Testimony." *Epistemology of Testimony*, edited by Jennifer Lackey and Ernest Sosa, Oxford University Press, 2006, pp. 160–182.
- Lackey, Jennifer. *Learning from Words: Testimony as a Source of Knowledge*. Oxford University Press, 2008.
- Lackey, Jennifer. "Testimonial knowledge and transmission." *The Philosophical Quarterly*, vol. 49, no. 197, Oct. 1999, pp. 471–490, <https://doi.org/10.1111/1467-9213.00154>.
- Lyons, Jack C. *Perception and Basic Beliefs: Zombies, Modules, and the Problem of the External World*. Oxford University Press, 2009.
- OpenAI. "Introducing Openai O3 and O4-Mini | OpenAI." *Introducing OpenAI O3 and O4-Mini*, 16 Apr. 2024, openai.com/index/introducing-o3-and-o4-mini/.
- Owens, David. *Reason Without Freedom: The Problem of Epistemic Normativity*. Routledge, 2000.
- Owens, David. "Testimony and Assertion." *Philosophical Studies*, vol. 130, pp. 105–129, <https://doi.org/https://doi.org/10.1007/s11098-005-3237-x>.
- Pelling, Charlie. "Testimony, testimonial belief, and safety." *Philosophical Studies*, vol. 164, no. 1, 2013, pp. 205–217, <https://doi.org/10.1007/s11098-011-9849-4>.
- Shapin, Steven. *A Social History of Truth: Civility and Science in Seventeenth-Century England*. University of Chicago Press, 2011.
- Stanley, Jason. "Knowledge and Practical Interests, Selections." *Epistemology: An Anthology*, edited by Ernest Sosa et al., 2nd ed., Blackwell Publishing, 2008, pp. 721–741.
- Stanley, Jason. *Knowledge and Practical Interests*. Oxford University Press, 2005.
- Wright, Stephen. "Circular Testimony." *Philosophical Studies*, vol. 173, no. 8, 30 Oct. 2015, pp. 2029–2048, <https://doi.org/10.1007/s11098-015-0595-x>.