

Risk and Value in Trusting AI

Drew Chambers

Under review

Do not cite

1 Introduction

Here's a statement from Richard Susskind's book *How to Think About AI*:

AI is being bundled into and integrated with everyday software. [...] In practice, this means that most of us will soon have at our fingertips our very own personal assistants and research assistants, helping us to organize our lives and keeping us fully informed. Likewise, we will have our own personal tutors, ready to train us on all sorts of subjects. (2025, p. 10)

At the moment, Susskind's sentiment is ubiquitous. It is a sentiment underpinned by values like democracy and efficiency. More can be done with AI than without AI and affordances that were previously accessible only by some are now within the grasp of many, or so the sentiment goes. Until now, our assistants and tutors have been people, and it is in part their agency that enables them to serve in those capacities. On the other hand, when enthusiastic users of AI systems are met with criticism for using AI, a common refrain in response is that AI is "just a tool." In relations with people and tools, we require some sort of trusting attitude that lies on a spectrum from mere reliance to intimate dependence. If we relate AI to us as either a tutor or a tool, we stand in a trusting relationship with it. The two halves of that statement are each important in their own right. Is AI a tutor, a tool, or something else? Whatever AI is, is a trusting relationship the right kind to have with it?

At some level, this paper asks whether we should trust chatbots, but what will become clear is that focusing on reliability and the conditions for trust obscures the heart of the matter which is whether chatbots can give us what we want—or at least should want—when it comes to aims of inquiry and expression. Inquiry and expression are not the only use cases for AI just as teaching and assisting are not the only use cases for people. With every new model, more and more users are turning to chatbots for companionship, therapy, and love. For now, I set aside these instances in which AI is thought to be the friend, caregiver, or partner at our fingertips. Each of those roles entails inquiry and expression in its own right, but there's an intuitive difference between asking a chatbot to therapize you and asking a chatbot to give you feedback on a paper.

I take the position that trust is not warranted when it comes to chatbots. An orientation of reliance is risky, but perhaps acceptably so depending on the use case. The important contribution will be to demonstrate that there is risk not so easily obviated by being in the right epistemic relation with a chatbot and its outputs. Further, there is a larger value-gap between a chatbot and its non-AI counterparts than reliabilist models locate.

The plan to show just how wide the risk and value gaps are when using chatbots for expression and inquiry is as follows. First, I'll briefly examine what it means to trust something, looking at some distinctions between trusting people and trusting objects. This allows me to set the stage for the primary inquiry, which is whether chatbots are the sorts of things we can or should trust. Second, I'll argue against the idea that chatbots are just tools, a view I take to be the more palatable between it and thinking of chatbots as testifiers. Next, I'll argue against the more extreme view, that we should treat chatbot outputs as testimony. These moves will leave us in a conundrum: if chatbots aren't tools and aren't testifiers, what are we doing when we rely on them? The dominant answer in the literature comes from what I call reliabilist accounts. The spirit of these accounts is that so long as we hold the right beliefs about the status and reliability of chatbot outputs and are in a position to ensure accuracy by appealing to some reason other than the output itself, we can use chatbots responsibly without trusting them. I argue that these accounts get the epistemology largely right but nonetheless still miss two things: a value gap and a risk I call agential naivety. The value gap ignored by the reliabilist accounts, which themselves focus on the capacities of chatbots to help us acquire knowledge, is rooted in the particularity that we seek in communication and knowledge endeavors with other humans. Agential naivety grows out of this value gap by homing in on a mistaken belief that having the right beliefs about a chatbot will insulate us from being shaped by it. I conclude that we should trust chatbots only in strictly tool-like uses, where we have the expertise to check what they give us, all the while recognizing that any use puts us at risk of being shaped in ways that leave us without the values we otherwise would pursue absent a chatbot.

2 Trust

We can distinguish between two accounts of trust; broadly speaking, there have been agential accounts and non-agential accounts.¹ It is widely thought that the contemporary agential accounts have their origins in Annette Baier's "Trust and Antitrust," which identified trusting relationships as ones of vulnerability in which the truster trusts not just in the capacities of the trusted but specifically in their goodwill (1986, p. 234). Baier's

¹ For an overview of these accounts specifically in relation to AI, see Viehoff (2023).

account, while influential, has met various convincing counterexamples—such as those from Onora O’Neill (2002) and Katherine Hawley (2014)—which have led to a family of accounts that move beyond a goodwill specification but nonetheless define the trusting relationship as one requiring agential capacities on the part of the trusted. On these accounts, one trusts in another’s cooperation (Simpson, 2012), sentiment (Jones, 1996), sincerity (Holton, 1994), belief (Faulkner, 2007) or something else that would be possible only in conditions of the trusted being an agent.

The agential accounts have to grapple with the fact that in normal use, it is natural to talk of trusting non-agents. Is it a conceptual mistake or just misspoken to say something like “I trust my calendar”? Do we really mean “I rely on my calendar to deliver a true proposition concerning the date”? Non-agential accounts attempt to explain how it can be appropriate to speak of trust, rather than mere reliance, when it comes to things without agency. Thi Nguyen (2023) is one such example of a non-agential account of trust.² Nguyen’s account jettisons Baier’s goodwill requirement but retains the normative diagnostic that underpins it, which is the sense of betrayal when our trust is violated. When we rely on something, we may be irritated when it doesn’t come through for us as we expect, but when we trust something, we have a stronger reaction, one characterized by feeling betrayed, hurt, dismayed, or angry. With this litmus, it becomes easier to see how talk of trusting objects and environments is sensible. Imagine how gutted it would feel to find out that the time clock at work had been running a constant one-hour slow every workday; it would be an utterly natural response to be more than just irritated at having to do some extra paperwork to remedy the accounting error. You trusted that the time clock would report your working hours accurately and learning that it did not evokes a sense of betrayal.

Trust, on Nguyen’s account, is to adopt an unquestioning attitude toward the thing trusted. When we trust something in this way, we suspend deliberation about the thing in the trusted matter. I trust the time clock will record my punch-in and punch-out accurately. I trust my email will be delivered to the intended recipient after I send it. I trust my mechanic of twenty years that my car does need a new spark plug. In all these cases, presuming there is actual trust, I do not continue deliberating about the matter at hand and take it that the matter is settled. There are three insights to be gleaned from this account, thinks Nguyen. First, “when we say we trust agents, often much of that trust is actually to be cashed out in terms of the unquestioning attitude sense of trust” rather than the terms of agency—a disposition or mental state, for instance—on the trusted person’s end. What matters in these cases of agent-to-agent trust is just that the truster can put an end to deliberating

² I rely on Nguyen’s account particularly for its parsimony in adjudicating cases of trust from non-trust and distrust.

about the matter at hand. Second, this account also demonstrates that the textbook cases of trust in the literature obscure how omnipresent trust as an unquestioning attitude is in our everyday lives: in Nguyen's words, "Trust is so common—it is such a background feature of our lives—that, much of the time, we don't even notice how much we trust" (2023, p. 222). Third, and perhaps most importantly, the explanation for why we trust and why trust is an indispensable background feature of our lives is that trust is a way to expand "our agency by integrating in bits of the external world" (2023, p. 217).

This third insight is what Nguyen terms the "integrative stance" of trust, and it explains how objects "can become phenomenally integrated into one's practical functioning and cognition" (2023, p. 230). Agential integration helps us to distinguish cases of relying on objects from trusting in objects. Certain things—devices, people, structures, systems—have "come to be functionally integrated with [us] to various extents, and the breakdown of that integration is a violation from the point of view of the integrative stance" (Nguyen, 2023, p. 231). If an unquestioning attitude toward things means to integrate them with one's own practical reasoning, then trust, in this sense, is to bring something into oneself, to work toward the seamless integration of something else with oneself. There are two reasons for trusting in such a way that can be brought into sight by thinking of both how we live and how we want to live. On the first consideration, call it non-deliberative efficiency, we simply do engage multiple things in our practical reasoning and agency; given this fact, trust allows this engagement to be more fluid. Without an unquestioning attitude, it is not the case that we simply would stop reasoning and acting on the basis of other things, rather we would just do so haltingly and without the gracefulness that trust in these things permits. The second consideration, call it distributed competence, has us consider our limitations. We have limited resources across the board—cognitively, bodily, and otherwise. Trust is a freeing up of our limited resources thereby enabling us to accomplish more than we could without it; it allows us to fold things into our agency, expanding the competencies on which we can draw to achieve ends.

Nguyen's is one account of how it is not simply metaphorical or mistaken to say we trust non-agents, but it can be difficult to see whether an account of trust can cover all cases of trust largely because the way we think of trust is slippery. For instance, in saying that I trust someone I can mean that I take it that they have the quality of being trustworthy. We could call this a virtue-sense of trust, and in this case, we are attributing a certain kind of character to what is trusted in which trustworthiness is bound up with other qualities like honesty. The table may be trustworthy in that it is reliable, but not trustworthy in that it possesses a certain kind of character. When we take this back to trust as an unquestioning attitude, it can be seen that there still exists that question of *what* receives an unquestioning attitude

on our part. Nguyen offers a variety of distinctions, such as trusting informational sources in which the unquestioning attitude means “not questioning X’s deliverances concerning Z” where Z is some domain. Trust occurs between oneself and another thing, which suggests that the object of trust simply is the other thing, but this is an unsatisfactory answer to the question of what is trusted, because even in cases where we mean trustworthiness as a function of character, we do not extend trust to something simpliciter. Our trust is always *qua* some aspect of the trusted, whether their character, their capacity to do something, the certainty that they *will* do something in particular, their knowledge in a particular domain, and so on. To wit, “I trust my doctor to perform their medical duties, and not [...] to successfully do modal logic or play jazz” (Nguyen, 2023, p. 226). This case demonstrates the way that the object of our trust is actually variegated: I trust that my doctor *can* perform their medical duties and that they *will* perform their medical duties.

What then *could* it mean to trust a chatbot? On the one hand, we could say that trusting a chatbot means to attribute to it the quality of trustworthiness as a description of its character. On the other hand, we could say that trusting a chatbot means to put aside questioning of its deliverances or capacities, whether wholesale or in some domain. The first instance would be to treat a chatbot as apt for agential trust, whereas the second instance would be to treat a chatbot as apt for object trust. Duncan Pritchard (2026) argues against a chatbot’s trustworthiness on both fronts by claiming that the first kind is to ascribe to it qualities it cannot have (rather than just qualities it fails to have) and claiming that the second kind would require its deliverances to be safe, rather than just reliable. Pritchard’s view demonstrates that chatbots are not the kinds of things that are in the market for trust, which is to say that they are not trust-apt. Compared to Nguyen’s unquestioning attitude account, Pritchard’s view seems to rule out trust by fiat—you cannot trust something that is not, in the first place, a candidate for trust.

The reality is that people will say they do or don’t trust chatbots, which is nonsensical if chatbots are the kinds of things that are not candidates for trust. But insofar as a chatbot may be seamlessly integrated into one’s practical reasoning, it is a thing that is trust-apt. The unquestioning attitude account allows us to focus on the lived practice with chatbots that many people face; the important part is not arguing about whether a chatbot is trust-apt, but rather about what it means to take an unquestioning attitude toward something, regardless of whether one would characterize the thing as trustworthy. Chatbots are trustable even if they are not trustworthy.³ This brings out another virtue of Nguyen’s account: trust is demonstrated in action. So long as one has an unquestioning attitude to

³ This claim underlies Holton’s (1994) claim that we can *decide* to trust and may have reasons for trusting disproportionately to something’s trustworthiness.

something, one is in a trusting relationship with it, regardless of whether one “thinks” or states that one trusts it.

When it comes to chatbots, users can be in a state of epistemic trust and practical trust, wherein one trusts the chatbot to put one in a better epistemic state or trusts the chatbot to accomplish some task successfully. The value of Nguyen’s account is to reveal that trust is risky not purely by virtue of the goodwill of what is trusted but because of the intimacy with what is trusted. An integrative stance with things innately engenders risk because in such a stance one outsources tasks to something else, adopting a first-order disposition to trust unquestioningly that the thing will not let us down and a second-order disposition to put aside questions about the first-order disposition (Nguyen, 2023). We ought to avoid being agentially gullible in which we are “willing to outsource [our] agency too readily—to let any old thing in” (Nguyen, 2023, p. 239). By putting ourselves in trusting relationships with chatbots, we are not merely risking that they will betray us in some sense, whether epistemically or practically, but also risking that we are becoming something we don’t want to be by fusing ourselves with the technologies on which we unquestioningly rely. The risk, therefore, is not just epistemic or practical, but also existential. This insight will allow us to return to Pritchard’s argument—and others like it—later in §6, at which point we’ll be able to see the ways it sidesteps more risk than it lets on.

3 Chatbots as Tools

By looking at trust, we were able to set the stage by setting the stakes. To trust something “is to let it in, to let it muck about directly with one’s practical and cognitive innards,” which means we should not be gullible trusters and trust just anything (Nguyen, 2023, p. 2023). However, we need to trust some things and some of the things that are candidates for trust are objects, both designed and natural. Typically uncontroversial objects of trust are tools and machines we reach for unthinkingly on a daily basis: pens, cars, forks, and the like. For many of us, included in this list of objects we unquestioningly trust are computers, the internet, smartphones, and, for a growing number of people, chatbots. Sometimes we are sheepish about what we rely on unquestioningly, and a latent appreciation of the fact that trust is so intimate and therefore so risky may underlie the common refrain that “AI is just a tool.” Behind the affirmation is a sense that in tool-use, the user is still in control, and the tool is subject to the aims of its user. The idea is that if my own aims are good—and that if *I* am trustworthy—then there ought to be nothing suspect about me using a tool to accomplish these aims. To clarify the legitimacy or illegitimacy of the “just a tool” claim, we can prise apart the intention behind the claim and the concept invoked. When we do

so, I think we'll see that chatbots are neither tools in the sense that "just a tool" intends nor are they tools in a robust conceptual sense.

Claiming that something is "just a tool" is a defensive claim, hinging on "just." It rebuts implicit judgments about outsourcing one's practical powers and therefore aims to communicate a sense of control as instantiated in two positions. First, it conveys that the mechanisms and resources through which one completes a task are support to the agent; we wouldn't think to negatively evaluate a painter for using a paint brush, because the paint brush accomplishes the painter's aims without imparting any influence. In the case of the brush, it really is just a tool. Second, the "just a tool" claim insists that one be judged on the basis of tasks qua tasks and outcomes qua outcomes and so deflects criticism away from the choices particular to the accomplishment of tasks and outcomes. In a general sense, the user who declares that the technology or data they're using to meet some ends is just a tool, aims to convey that the purported tool imparts no value and does not end up, so to speak, running the show.

It may be easy to think that the "just a tool" claim invokes a withholding of trust, but I think this misses the finer point. We do trust tools and in doing so do not surrender to them; the photographer who insists the camera is just a tool necessarily trusts the camera but does so in a way that foregrounds and maintains the photographer's own creative agency. In other kinds of trusting relationships, however, we may let ourselves be changed by the thing trusted or even surrender wholly or partially to it. When I ask a colleague a teaching question and trust their opinion on the matter, I do not surrender my agency to theirs, but I am nonetheless predisposed to exercise my agency in light of something I cannot principally control, in this case, their judgments. When I co-author a paper, I trust my co-author and allow them to form with me a sense of shared agency. When I get in a taxi, I surrender my agency more or less completely to the driver.⁴ Consequently, a criterion for determining whether the "just a tool" defense is viable is just how much control one gives up in trusting something or someone.

The control criterion is not just a metric for the extent of control but also the relation of the control to the task at hand. Consider a basic calculator. "It's just a tool" is a viable defense if my waiter objects to me using the calculator to determine the tip I will leave, but not a viable defense for a seven-year-old taking a times-table test. In both cases, the calculator serves the same function and the extent of how much control is relinquished to

⁴ The absolute dissolution of agency is perhaps mythical and if not mythical then likely an instance of moral failure, as MacIntyre (1999) argues.

it is the same, but the task at hand—calculating a tip versus demonstrating ability to multiply unaided—dictates the appropriateness of the tool.

While “just a tool” is a common enough phrase, conceptual analysis of tools is lacking. Philosophers talk of technology, of artifacts, of devices, and of instruments, but is a chatbot any of these things? It certainly is technology, but that category is broader than the category one invokes when one proclaims something is just a tool.⁵ I won’t attempt necessary and sufficient conditions in this paper, but a working account will do. I want to suggest three conditions that seem to underlie a basic account of what a tool is, minimally as a designed artifact, but that a chatbot fails to meet in certain use cases. Given the argument, I set aside conditions that it obviously does meet: it is designed, it is an artifact, it is used for practical ends. The conditions where it begins to exceed the tool category are in (1) control, (2) defined functionality, and (3) expected outcome. In use, a tool is something principally controllable by the user with some circumscribed function that produces reasonably predictable results. Control in this sense is more or less the same as in the sense behind the intention of something being “just” a tool; something can be relatively autonomous, like a clock, and still meet the control function so long as the user sets the operation and grasps the output. The set of a tool’s functions can be vast—there’s a lot a word processing application, for instance, can do—but the set is defined. And the outcome of the use, tracking functionality and design, is predictable to the user. I may not understand exactly how my car works, but it only starts and moves with my intent, it only does so many things, and whatever it does in normal use is something that I expect to result from my input. In all of this, a tool is defined by its use and therefore its user.

Given these two positions—the “just a tool” defense and the tool category—what can be said of chatbots? One route would be to say that a chatbot’s capacities rule it out categorically as a candidate for being a tool; this route is promising, but there are routes with less resistance. The first route is descriptive: many uses of chatbots do not constitute tool uses. In inquiry, users expect chatbots to deliver knowledge, give advice, suggest ideas, and so on. In expression, users expect chatbots to draft papers, emails, reviews, and so on. Given these uses, our possible trusting relationships with chatbots also belie the “just a tool” defense. In using a chatbot for inquiry and expression, we seek not to merely extend our agency through the autonomous use of a tool but rather we seek to offload our agency to the technology, to let it dictate what we think, what we do, and how we do it. For many

⁵ Pre-internet accounts tend to talk of tools as devices, such as Heidegger’s hammer case (1962), and distinguish tools from machines, but there is also a common history of calling concepts and heuristics tools—look no further than Dennett’s *Intuition Pumps and Other Tools for Thinking* (2013).

users of chatbots, the unquestioning attitude extended to chatbots concerns yielding their beliefs and practical aims to the chatbot.

For both inquiry and expression, the use of a chatbot typically resembles the kinds of interactions between two people, rather than the use of a tool by one person.⁶ But even when people contribute to our practical reasoning or extend our practical agency, we do not think of them as tools. Tool status requires being used in a tool fashion, and functional contribution alone does not meet that bar.

The second route to invalidating the “just a tool” defense and the tool-candidacy is principled. In the inquiry and expression uses, a user has limited control over a chatbot, the chatbot has no circumscribed function and does not produce reasonably predictable results from the user’s perspective. A chatbot’s outputs are open-ended. The user plays the role in bringing forth the outputs, but the outputs themselves are, for the user, a matter of surprise. In use, we cannot say what will happen as a result of our input; at most, we can hope something happens, but even our hopes are generalities—we hope it produces an explanatorily satisfactory output or a “good” paper draft, but we cannot say what the output will contain because of the very fact that we are seeking deliverances we cannot conjure on our own. There is a key difference between this kind of surprise in result and the kind of surprise in result that comes from using a calculator. Like with my use of the chatbot, I use a calculator because I don’t know the deliverance I seek, but with the calculator I am trusting it to use well-defined mathematical operations to produce a deliverance that is both truth-apt and circumscribed.

The chatbot outputs I seek are far dissimilar from the deliverances of other instruments in which we trust, like calculators and thermometers, and which answer to questions within a fixed field. These instruments are reliable because, so situated, they would not easily deliver “any falsehood in that field” (Sosa, 2006, p. 117). The outputs of calculators and thermometers are truth-apt, whereas an email or advice is not truth-apt at a category level (though propositional deliverances contained in the email or advice may be truth-apt). There is then no field relative to which its reliability could be assessed, and so no clear content for the trust that Sosa thinks all reliance on an instrument presupposes. These qualities of chatbots and their deliverances, particularity in certain kinds of uses, need not rule out all uses of a chatbot as tool-like. It may be said that a horse can serve a tool-function, but it would not be right to say that a horse is reducible to its tool-function (Lott

⁶ A claim bolstered by the fact that transformers (including early models of ChatGPT) can pass Turing tests of multiple flavors (Howick et al., 2021; Jones & Bergen, 2025) and that when they don’t pass, it’s sometimes because people think AI is superior to humans (Aharoni et al., 2024).

& Hasselberger, 2024). I will return to this point in §5 because it underlies how certain theories of epistemically responsible chatbot-use work.

The problems for the tool-use model can be summed up as such. In ordinary use in the domains of inquiry and expression, chatbots present as speakers, their outputs present as testimony, and the use presents dialogically, but if we're looking for testimony and reliability, we would need to relinquish control and insist on queries where there is reliability to test, which is to say queries in which ostensible outputs are truth-apt. Given that the uses I've described and the ones that Susskind proclaims as the future of AI are testimonial in appearance and (often enough) subjective in kind, perhaps the solution to the question of chatbot trust is to say that we ought to treat chatbots like speakers, which is to say like things capable of testimony.

4 Chatbots as Testifiers

It's easy to anthropomorphize a chatbot, not least of which because they sometimes have human-sounding names, but the allure of treating chatbots like speakers extends beyond our all too human urge to anthropomorphize inert objects. Chatbots produce propositionally structured answers that look like testimony; they seem to respond to inputs in domain-sensitive ways, like a qualified testifier would; and the phenomenon is dialogical in form—we can audibly talk to a chatbot and receive responses back in an ever more natural human-sounding voice. What would it mean to trust a chatbot as a speaker?

When someone tells us something, we expect them to (a) present what they are saying as true, (b) intend to inform someone, (c) take responsibility for what they are saying, and (d) have the authority to assert it as true. These expectations are what make testimony a source of knowledge. On this picture, telling is an invitation to trust (Hinchman, 2005), and to assert is to undertake a commitment one can be held to (Brandom, 1994). Regina Rini puts it simply, by highlighting the competency and sincerity conditions she draws from: “When a person attempts to provide testimony,” she “really means it, and she knows what she’s talking about” (2020, p. 2). John Eriksson contends that the concept of sincerity at stake in testimony is sincerity as a communicative virtue. The speaker believes they are in the state their words express, wants their hearer to believe on the basis of what they say, does not want to leave a false impression, and is properly justified in saying what it is they say (2011, p. 232).

I want to distinguish between two grades of testimony, a thin kind and a thick kind. Thin testimony is required to meet the standard norms of assertion: sincerity, competence, and

intention. Thick testimony, on the other hand, goes the extra mile by virtue of meeting the required norms of assertion—call them reliability requirements—and adds what I call attunement.

To illustrate the distinction, consider two ways I might go about explaining some concept. Here's an example from real life. I assigned my students Christopher Lebron's excellent article "Without the Loving Strains of Commitment" alongside James Baldwin's (1963) "A Letter to My Nephew." The student was trying to grasp Lebron's and Baldwin's arguments that love is what is required for social justice, particularly in conditions of racial injustice. The student and I engaged in a back-and-forth, him pushing the idea that only laws can enforce justice, and me trying to help demonstrate that Lebron and Baldwin agreed with the student's claim, but that Lebron's argument hinged on love being a sentiment that can help people keep their commitments under the law under strained conditions and Baldwin's argument hinged on injustice being rooted in an ignorance that love and acceptance could dissolve. I eventually reached for the example of Daryl Davis, a black jazz musician whose sincere engagement with Ku Klux Klan members has resulted in more than 200 Klansmen renouncing the organization. Davis says this, of his conversations with Klansmen:

When a Klansman walks into a room, his wall is up. I'm trying to bring that wall down. I've been to 57 countries on six continents. But no matter how far I've gone, I've observed the same thing: we human beings all want the same things. We want to be respected. We want to be loved. We want to be heard. And we want the same thing for our families as everybody else wants for their families. (Davis and Walker, 2020, p. 76)

I cannot say that the student "grasped" Lebron and Baldwin's ideas, but what I can say is that my testimony to the student was an attempt to be attuned to him. In speaking with him, I was drawing on my knowledge of him, our experiences thus far in and out of class, comments I knew he had heard from peers, our shared basis of the Lebron and Baldwin texts, his physicality during the conversation, his responses to me and how exactly he conveyed them (his speech patterns, his wording, again, his body). This kind of attunement in speaking is what all people do in conversation and what all good teachers do in attempting to explain concepts to their *particular* students.

Contrast this example with a hypothetical similar instance of communication, but this time thin. I record a video of myself explaining the Lebron and Baldwin texts. In the video, I imagine interlocutors and draw on the Davis example to help illustrate what I take Lebron and Baldwin to be saying. I put it up on YouTube and hope that people find it and learn

something from it. In this case, my testimony is offered—sincerely, competently, and intentionally—to anyone and so to no one in particular. The content of the video may overlap, even nearly perfectly, with my office hours exchange, but the explanation is thin because attunement to my listener is absent or general. Attunement is my responsiveness to the hearer as a particular person.

The distinction I've drawn demonstrates inklings of the positive positions I take in §6; it demonstrates that reliability requirements are not necessarily all that is sought in certain domain endeavors like inquiring to understand and searching for expression.

Much of what people ask chatbots for is ideally thick testimony, as evidenced by excitement about AI “models” that can “learn” and present “tailored” outputs and the fundamental mode of engagement with chatbots, which is dialogical. Again to the domains, certain inquiries between a user and a chatbot (or between two people) call for thicker testimony: users want useful feedback on a draft of their paper, or advice about how to negotiate a particular job offer, or an explanation of the ostensibly important parts of a reading. When the outputs sought are evaluative in form, as feedback and advice are, a further question arises about whether one ought to defer to it at all, even when a human speaker is sincere and competent (cf. Mogensen, 2017). So can a chatbot actually provide testimony of either grade? Elizabeth Fricker (2026) advances a view that argues that “AI ‘testimony’ is fake” on the basis that the outputs are designed to present with features of assertoric communication, but that the outputs cannot be acts of communication due to a chatbot having no communicative intentions and an inability to take responsibility for what it outputs. If the chatbot fails to meet the requirements of thin testimony, then *a fortiori* it fails to meet the stronger requirements of thick testimony.

There's a growing body of literature attempting to justify the claim that chatbot outputs qualify as testimony of some kind or other. Iwan Williams and Tim Bayne (2026) argue that chatbots are “proto-asserters,” sitting near the “edge of assertion” like toddlers who respond to the question “What did you do today?” with “Zoo!” (p. 3071). Even if chatbots are proto-asserters, however, Williams and Bayne concede that chatbots are not sanctionable in the sense that actual testimony requires of its speakers. Jumbly Grindrod (2024) argues, drawing on Gareth Evans (1982) and Ruth Millikan (2005), that chatbot outputs can be meaningful because chatbots “are able to employ [a] pre-existing system in the same way that ordinary speakers do” (p. 15). However, chatbot outputs being meaningful is, as I think Grindrod would agree, not the same as them constituting testimony.

One might object that chatbots *are* attuned, given the list of features I mentioned above. They respond to the particulars of a prompt, and some chatbots are able to “remember” and “learn” by adding text from earlier inputs and user-set descriptions. This is a fair point about how the outputs come to be what they are, but these aspects of a chatbot do nothing to approach the attunement that thick testimony requires. For one, a chatbot’s “tailoring” has neither the intention nor the answerability of a teacher’s attuned testimony. But more importantly, we’re attuned to each other and the world in ways that exceed the strict relation of words in semantic space.

We can return to Susskind’s optimistic declaration with which I opened this paper. The idea that a chatbot can be a teacher or a personal assistant is patently inconsistent with the claims that a chatbot cannot deliver thin or thick testimony. We can shift the focus from agency, by noting that whatever teachers and assistants can do for us, all a chatbot can do is provide text.⁷ As such, the comparison between chatbots and teachers will come down to the ability of each to say (or “say”) things that serve the practical ends of teaching. Given that what we want of teachers in the realm of language is testimony, then the preceding analysis demonstrates a chatbot will fail to meet the roles being thrust upon it.

We find ourselves in a bind. It seems that in the domains of inquiry and expression, the chatbot is not a tool, but also not a speaker. At the same time, it undeniably produces outputs that we irresistibly read as meaningful, because our “effortless immediate mental entertaining of the semantic content” is “generated by subpersonal processing modules, and as such is involuntary and belief-independent” (Fricker, 2026, pp. 2868-2869). In short, the outputs present as testimony and regardless of whether they constitute testimony, we are predisposed to treat them as such. Most users of AI have no sense of the technical meaning of “testimony” and might struggle when it comes to making the kind of distinctions that epistemologists make about how chatbot outputs differ (or don’t differ) epistemically from human assertions. For the chatbot users whose use demonstrates trust as an unquestioning attitude, pointing out that the output is not testimony will likely fail to persuade them to use a chatbot differently. After all, the outputs are not just assertoric-looking, they also seem intelligent, articulate, personal, and reliable—and this characterization is all the more striking when often the feedback, advice, and explanations we get from friends, colleagues, and, yes, teachers, seems uninterested, impersonal, limited, unrevealing, or hard to understand. Is there a way to philosophically legitimate using a chatbot in our inquiries and expressive aims? The primary approaches to making the use of chatbots responsible, rather than hazardous, and rational, rather than delusional, are

⁷ Or images and videos and the like, but all these outputs rely on the same transformer technology that allows a system to generate outputs by predicting tokens (whether they are tokens corresponding to pixels or words).

what I will term reliabilist accounts. They have much to offer the enthusiastic users of AI who are ready to integrate them into their agency, but the accounts fail to locate a kind of risk and loss that exists even in putative epistemically responsible use of chatbots.

5 Avoiding Risk

We shouldn't lose sight of the orienting frame of this paper, which is the trust endemic to our lives as ambitious but limited agents. We need to trust, but we also want to avoid being agentially gullible, of trusting just anything. Being discerning trusters is a way of staving off the risks of gullibility, but trust in the unquestioning sense is not the only way we can use a chatbot. The preceding sections suggest we would be mistaken to trust a chatbot as a tool or as a speaker, but they leave us in the lurch: can we trust a chatbot in ways that do not make these category mistakes? If we did, what real risk would we be running in use?

A family of approaches to these questions argues that by putting ourselves in the right epistemic relation to a chatbot, we can extract the practical and intellectual value from the technology without exposing ourselves to the risks of full-throated trust.⁸ We can call these accounts reliabilist accounts. The accounts differ in their details, but they share a relatively uniform structure that locates risk in regard to knowledge and reliability and therefore suggest various ways of responsibly using AI *qua* epistemic norms. I take it that all of these accounts grant that a Baier-style interpersonal trust in AI is inappropriate, but each then identifies a relation between user and chatbot that is weaker than full-throated trust, whether we call that relation one of reliance, calibrated reliance, non-interpersonal trust, make-believe, or virtuous oversight. The grounds of the appropriateness of these relations are first in the reliability of a chatbot's outputs and second in the user's epistemic stance toward them as constituted by their beliefs about the chatbot and the specific way they treat its outputs. The putative promise of these accounts is that by putting ourselves in the right epistemic relation to a chatbot, we can extract its practical and intellectual value without exposing ourselves to the risks of trusting it. Giacomo Zanotti et al. characterize the epistemic accounts at the center of this literature well: they "apply trust and trustworthiness

⁸ The literature on AI reliability is not focalized around the aspects of trust brought out by Nguyen and others. Some of the literature focuses on the opacity of AI systems (which is to say the fact that "how" a transformer produces the outputs it does is not transparent), some on algorithmic bias, some on interpretability of AI outputs, some on the predictive methods that form the nature of much generative AI, and some on value alignment. I will not routinely disambiguate between accounts on the basis of what they take to be the complicating factors when it comes to trusting AI, because the thrust of the argument remains to elaborate the conditions that responsible AI use would require. For instance, some say opacity undermines reliability (e.g., Koskinen, 2024) and others say opacity does not (e.g., Ortmann, 2025); in either case, opacity is not the component of the argument I am tracking.

to AI systems on the basis of these systems' performance and the justification of our beliefs about it" (2024, p. 2694).⁹

On the testimony front, Ori Freiman (2024) assumes that knowledge can be acquired from chatbots and says that such knowledge is properly called "technology-based belief," since the beliefs do not rely on testimony and are sourced from non-agents, but nonetheless present as natural language. His account does not establish an account for what constitutes knowledge gained from a chatbot or guidelines for epistemically responsible use, but it does argue that the non-testimony classification need not dissuade some from taking it to be a source of knowledge. Fintan Mallory also contends that chatbot outputs do not constitute testimony and treats our engagement with chatbots as prop-oriented make-believe, in which the outputs are "literally meaningless but fictionally meaningful" and can therefore help us learn about the world (2023, p. 1083). On this approach, a chatbot is like a fictional character from a novel with a corresponding fictional semantic competence; one can learn things about the world from a novel because of the novel's author and one can learn about the world from a chatbot because of the chatbot's "particular architecture, training data, openness to third-party probing" and so on (2023, p. 1096).

The responses to the non-testimony quality of the chatbot outputs parallel accounts that focus on reliability rather than trust in AI use. Niël Conradie and Saskia Nagel (2024) argue that since trustworthiness and responsibility are agential notions that AI cannot bear, policy should speak of reliable, safe, and explainable AI, and of our being trustworthy and responsible *about* AI rather than of trustworthy or responsible AI. John Dorsch et al. (2026) likewise set moral trustworthiness aside as a category mistake and recast the question as one of calibration: "Reliance on the outputs of an [AI decision-support system] is well-calibrated when users accept the system's outputs when they are correct and reject them when they are not" (2026, p. 9). Jakob Ortmann argues that epistemic reliance on an opaque system such as various AI applications does not require trust but just reliability such that researchers who use AI applications do not "just rely on a black box about which nothing can be known, but on a system that has undergone extensive robustness checks" (2025, p. 17). To wit, "if an LLM comes up with an argument, what matters for epistemic

⁹ Zanotti et al. take it that their preferred account is not an epistemic account because its "notion of trust is also grounded in an ethical dimension" (2024, p. 2699). They cash out this premise by arguing for developers and regulators being held accountable for ensuring that AI development and regulation maintains "(i) human agency and oversight; (ii) technical robustness and safety; (iii) privacy and data governance; (iv) transparency; (v) diversity, nondiscrimination and fairness; (vi) societal and environmental wellbeing; [and] (vii) accountability" (2024, p. 2696). These assurances produce a non-epistemic guarantee of reliability, so the payoff is still epistemic or practical.

dependence on this argument is whether the argument is sound, not what kind of opaque process has brought it about” (2025, pp. 19-20). Eamon Duede argues that “in principle” scientists “can be justified in trusting in the reliability of [deep learning models]” on the basis of the glut of research demonstrating, for instance, AI-assisted solutions to “complex scientific problems that have otherwise been intractable” (2022, p. 2). Duede calls techniques for justifying reliability of AI systems “post-hoc,” by which I take it that he means AI users may subject an AI output to scrutiny, verifying its truth value, scientific or otherwise, and, upon finding the output to be scientifically useful, grant reliability to the technology that produced it (cf. Duede, 2023).

Another line of accounts keeps the word “trust,” but invokes a thinner conception of it or sets the stage for an eventual thicker conception of AI trust. Sam Baron distinguishes strong notions of trust, which require the trustee to have mental states, from moderate notions, which do not (2025, p. 12). For Baron (2025), strong trust in AI is inappropriate, but moderate trust is “appropriate when a given AI system produces highly accurate outputs over time” (p. 8) and therefore an “unquestioning attitude toward AI” requires that users understand “that it does work,” an “understanding [that] can be gained by drawing on the same inductive base that we use to establish the reliability of the AI system” (p. 13). Ian Robertson defends Baron against the objection that unexplainable systems call for doubt, concluding that trust can be appropriate for systems with “an established history of stability and accuracy” (2025, p. 4). Philip Nickel’s account of trust in AI focuses on medical contexts where trust is defined as giving an AI system discretionary authority over a circumscribed question. Nickel endorses Neil Manson and Onora O’Neill’s (2007) account deemed “intelligent trust,” in which the discretionary authority is well-judged and properly limited in scope. In the case of AI, “the paradigmatic epistemic reason” for trusting AI “is the application’s reliability: it makes fewer incorrect judgments than the clinician” (2022, p. 7). Madeleine Ransom and Nicole Menard do not provide a positive argument in favor of AI trustworthiness, but they refute arguments skeptical of AI trust and call for new concepts of trust “to help educate and guide people in their interactions with AI and ensure that their dependence is properly calibrated” (2026, p. 9). Sara Blanco argues that reliance on AI may suffice in situations where AI’s motives are irrelevant to the user, that general trust may suffice in contexts where the stakes are low, and that appropriate trust requires the AI system “not only to be accurate but to reach accuracy following processes that the human trustor deems appropriate” (2025, p. 4178).

Duncan Pritchard argues that while generative AI can be reliable, it cannot be safe, and therefore “isn’t even in the market for interpersonal trust” (2026, p. 6). Even if we grant that a chatbot’s outputs are reliable, in that they constitute “results with a high degree of

probability” of being accurate (p. 1), trusting a chatbot will still be epistemically unsafe because its “epistemic indifference means that it will also easily generate false answers too, in surprising ways” (2026, p. 11). This is due to the predictive technology that underlies all generative AI and the modifying temperature parameter that controls the randomness of the model, whereby low temperatures mean more deterministic outputs. Across the board, any model does not generate the outputs it does because it takes it to be the “right” answer, but rather just the probabilistic answer due to its complex architecture. Is there a chance that a chatbot gives the factually wrong answer to a well-defined question with a correct answer well established in training corpora and post-training feedback and tuning? In principle, yes, even though the true likelihood may approach zero. Even so, Pritchard, though pessimistic about the capacity of chatbots to be trustworthy, does not conclude that it is altogether irresponsible to use chatbots; in this sense, he joins the pack. According to Pritchard, we can still acquire knowledge “via the use of generative AI so long as one combines it with independent epistemic support” (2026, p. 12). We should be “circumspect” about our AI use, he says, and given our circumspection “check” and “scrutinize” chatbot outputs such that we have a safe basis for our beliefs.

Vlasta Sikimić concludes something similar in placing the AI safeguard on the use side: responsible use of AI in science, on this view, depends on “the intellectual virtues of the human agents who interpret, evaluate, and act upon AI-generated outputs” (2026, p. 4). Sikimić walks through several use cases for generative AI in scientific research including conceptualization, analysis, writing, peer review, and more. On writing, she states that the user of a chatbot being asked to produce writing for a scientific article “must be intellectually honest and correct the AI whenever necessary” (2026, p. 11). Sikimić also contends that the use of chatbots in the peer review process is attractive due to the workload associated with peer review, but that “the responsibility falls on the researchers using these tools to carefully evaluate their outputs” (2026, p. 11). The upshot of Sikimić’s article is simply that because generative AI is so practically useful, it should be used, but should be used carefully and with virtue. In considering the risks of trusting chatbots, she alleges she makes “a moderate claim about surrendering epistemic effort and not full cognitive outsourcing as this [“full cognitive outsourcing”] is a much stronger position” (2026, footnote 3).

Perhaps there are folk views that endorse full cognitive outsourcing to chatbots, but the philosophical literature seems to be much more reserved though not cynically so. I take it that none of the views I’ve cited endorse offloading all of one’s cognition, labor, and practical aims to a chatbot, but they all suggest that there are ways to responsibly integrate AI into one’s agency. Set side by side, these accounts share three important features. First,

regardless of where they fall on the trustworthiness or non-trustworthiness of AI, reliability comes first. Whatever else they require, whether safety, calibration, explanation, institutional oversight, or virtue, is added to an existing requirement of reliability. Second, the remaining work falls to users, who are responsible for being in the right epistemic relation to a chatbot's outputs. The responsible user does not attribute agency to a chatbot and recognizes believing a chatbot's outputs at face-value is not a guaranteed or safe path to knowledge. They know that learning from a chatbot or responsibly using its outputs in their own practical endeavors requires scrutiny, making sure that one is in a position to understand the output as true on some basis other than the fact that it was produced by a particular chatbot. Third, the comparison and use cases are, for the most part, tools, in which a user holds a chatbot to its tool-function, such as providing facts about the world or analyses of data. Many of the cases these accounts were built on, particularly in medical and lab contexts, have ground truth. Protein structures (Ortmann, 2025), cancer screenings (Dorsch et al., 2026) and hospital readmissions (Nickel, 2022) *can* be checked against outcomes and data.

Consider another use case that falls outside the domain of natural science. In an input, I could ask a chatbot to give me a list of typos—not editorial comments—in a paper of mine included in the input. The chatbot output would be a list of typos and I could then use this list to direct my attention to spots in my paper where the list indicates they would be; if I see a typo, I can fix it and if I don't see a typo, I do nothing. In this case, I am using the chatbot as a more advanced grammar and spelling check where the use is practical and I do not know that the purported typos do exist in my paper until I am looking at them myself. The benefit to me is giving me a map that could turn out to be wrong, but most of the time will turn out right. In this example, the function is defined, I have a reasonable prediction of what the output will entail, the query is itself truth-apt, and I am not treating it as a speaker. It counts as a case of relying on, even perhaps trusting, the chatbot, because I use it reflexively as a guide to my action: I check the paper, expecting to see the typo so that I can correct it. In this sense, the reliabilist accounts above seem reasonable as ways of defining what it means to responsibly use a chatbot and how a particular kind of trusting relationship can conceptualize the chatbot as a tool.

There are two further implications of the accounts that they fail to state. First, these accounts leave certain uses uncovered. If chatbot outputs are only reliable sources to be checked, or leads to be confirmed by independent support, then the responsible state of reliance is conceivable only against a backdrop of truth-apt queries. The point of asking someone's advice is not to acquire knowledge but instead to give oneself over to their evaluative frame and to then evaluate the advice on the basis of some standard other than

merely its factuality. It can be easy to spot bad advice, but it is not always the case that one is in a position to know whether advice is good and in many cases of advice, there may not be a truth of the matter by which to evaluate the advice, even *ex post facto*. It might be responded that in these cases, responsible use just looks like the user evaluating the output and deciding if, in light of their own rationality, it is actionable. This response, however, employs a stronger form of trust; when I trust someone's advice, someone's feedback, someone's attempt to express my own thoughts back to me in different words, I rely on something more than their ability to string together on-topic sentences.

The second implication of the accounts is that a chatbot's epistemic utility and risk comes down to the user. In testimony, one is in a position to evaluate the speaker; if I take them to be a reliable knower in the matter at hand, I am in a position to believe what they say. With a chatbot, one can have some knowledge of the reliability of the model generally and within domains, but since the source is inherently unsafe, one must be in a position to evaluate the output itself. Experts in a given domain are in a position to directly assess content. For instance, in evaluating a chatbot's explanation of a philosophy paper, a philosopher can spot tells that the output is incorrect or off base without consulting the paper in question (e.g., perhaps the output attributes a view to the author of the paper or someone cited in the paper that the philosopher already knows the author or their source does not hold). An undergraduate student, on the other hand, is not in a position to easily assess an output's accuracy absent guidance from somewhere else. Many of the accounts I cited above are written with experts in mind, and they suggest that if one is principally unable to believe or disbelieve the output on some basis other than the output itself then one is in no position to do anything with it responsibly—aside from seeking out some other source, at which the utility is minimal, if not altogether missing.

Even with these attenuations and trust-accounts, a prevailing assumption survives, which is that a chatbot can be useful in the kinds of inquiries that would normally call for testimony that consists of recommendations, evaluations, endorsements, judgements, and the like. Consider a recent blog post by Brian Leiter (2026):

A colleague elsewhere, who subscribes to Astra, Open AI's high end LLM model (at least \$100/month), told me that he has found it useful for giving feedback on early drafts—he compares it to the feedback one gets from a good first-year graduate student. [He] invited me to give him a Nietzsche paper I was working on, which I did. Astra generated several pages of feedback, that was impressive in a number of ways, but deficient in some others. [...] I recently discussed this same paper with a very good Nietzsche scholar, who gave me more meaningful feedback.

But I think the comparison to a first-year graduate student is fair, with the caveat that many first-year graduate students know more about the primary text than Astra. What Astra could not do was challenge any aspect of my reading based on other texts of Nietzsche's. What Astra could do is assess the internal logic of my argument (with the caveat that it gets confused about a developing dialectic).

The idea expressed by Leiter is that a chatbot's outputs are comparable to a first-year grad student's testimony, because the outputs, though they may *not* be testimony, and though they may not be *safe* in some sense, still function practically just as well. Even given the reliabilist models, it seems that for many people, not least of which undergraduate students, the prevailing view is that we can be in the putatively right epistemic state—if we hold the right beliefs about the chatbot (its outputs are not testimony, they are not safe, they need checking, and so on)—and thereby allay the risk of agential gullibility, which will allow us to go on using chatbots to pursue our ends. This puts us, nonetheless, in a form of a trusting relationship where the object of our trust is our own epistemic fortitude and the chatbot outputs *as* reliable or somewhat reliable evidence for belief and action. In the next section, the last before the conclusion, I will consider whether the reliabilist accounts obscure more of the possible problem than they let on.

6 **Against Reliabilist Accounts**

When, then, should one trust a chatbot? Given what I have said, certain uses do constitute tool-uses: the user controls the operation, the function is circumscribed, and the result is predictable and checkable by the user. Reformatting a bibliography into a citation style I know, or translating a passage in a language I read, are uses of this kind. But this rules out most of the ways people use chatbots: to write for them, to give them opinions, to generate ideas, to conduct inquiry for them, and to do their expressing for them. In those uses we do not get what we want, because what we want there is particularity, and we let ourselves be shaped by something we naively think is not shaping us.

The reliabilist accounts let us see how we might be gambling when we rely on chatbots. They also give us ways to recognize just how much we are gambling and how, in theory, to avoid gambling at all. But is this the right lens? I think not, for two main reasons. First, the value gap between a chatbot and its non-AI counterparts, like books and teachers, is wider than accounts focused on generative AI's capacity to reliably deliver knowledge can explain. Second, trusting a chatbot carries a risk not easily obviated by standing in the putatively "right" epistemic relation to it.

The attunement criterion from §4 began to demonstrate the value gap between a chatbot output and human testimony. The point of attunement is that a speaker's particularity is what bestows value on their testimony above and beyond considerations of reliability and accuracy. Consider an ordinary case. Suppose I'm visiting a city for the first time, and I want to decide on a place to get dinner. I have a variety of resources before me to make this choice informed. If I happen to know a dear friend who lived in the city for many years and moved out of the city only recently, I might text her. Her recommendations may end up being less articulate, less complete, and perhaps even inaccurate compared to the result I would get from ChatGPT or a Google search AI summary, but what I value is not so much that she calls a restaurant Tuscan when it is more accurately called Ligurian and not that she gives me a bird's eye view of the city's entire food scene. Instead, what I value is her giving me *her* recommendation that flows out of her experiences, her values, and her tastes. Perhaps I know how and what she likes to eat and perhaps I have a sense that our preferences in this domain are aligned, but I think even this estimation is too strong. What I know is my friend, in an objectual sense; I know *her* and because of who she is, I am interested in what she has to say about my dinner plans.¹⁰

The same example holds in the apparently more important cases in which we might be inclined to call on a chatbot. I ask a chatbot how to teach a text. I ask a chatbot to give me feedback on a paper. I ask a chatbot to explain a concept to me. I ask a chatbot to talk me through an idea I have. I ask a chatbot for a literature review—or even a set of references—to use for research. Prior to the release of LLMs on the public stage, these pursuits would necessarily flow through my own intellectual efforts and my engagement of other people, and in asking colleagues, mentors, teachers, and friends these questions, I would be looking for responses that instantiate their particularity. It may be true that I could ask anyone anything, but the value of the response extends beyond its merely *counting as* an appropriate response to my query and makes contact with the person's subjectivity: their perspective, their personal history, the personal genealogy for the response, and their relationship to me. A chatbot attempts an Archimedean view from nowhere—or perhaps a view from everywhere—and so has no position from which to speak; its outputs have no particular genealogy beyond a statistical one.

The reliabilist accounts treat genealogy as beside the point. Ortmann, for example, writes, “what matters for epistemic dependence on [a chatbot's] argument is whether the argument is sound, not what kind of opaque process has brought it about” (2025, pp. 19-20). For an argument, that may be right, since soundness can be checked in a vacuum, but advice,

¹⁰ On objectual knowledge, both as a paradigmatic kind of knowledge and a particularly value-loaded kind of knowledge, see Chappell (2014).

feedback, speculation, and recommendation cannot be checked that way. This has to do, again, with the truth-aptness of these kinds of speech. If I have an idea for a paper and ask a chatbot to take the idea and transform it into an outline with some relevant citations—as I know many of my students, and apparently also some fellow philosophers, have done—then what can be said about the truth of the output? I cannot check the generated outline against my own outline, because the latter does not exist. I could check the generated outline against independent standards that I hold in mind for what constitutes a good outline (if I have them), but in this case, I would need to cordon off the evaluative standards that I would apply in the case of my professor helping me work through my thinking. In the human comparison case, I would be trusting that my professor speaks from some position within the field, with knowledge of who I am and the context in which I am working, and so on.

To drive home just how common it is to value particularity in our inquiring lives, I want to turn to an unexpected example: the history of travel writing. Travel writing of a sort has always existed, such as in Pausanias’s descriptions of Greece, Marco Polo’s *Il Milione*, and Captain Cook’s diaries, but the travel writing that comes to mind for many of us exists in the form of the guidebook (Bolan, 2020). Karl Baedeker’s guidebooks, starting in the early 19th century, are the first notable examples of these modern tourist guides. Here is travel writing historian Peter Bolan on what made Baedeker’s red-covered guides famous:

Baedeker used a star system to rate tourist sites with an asterisk (an innovative approach at the time), and he meticulously collected data, including state-of-the-art maps to provide his guidebooks with everything he felt the traveller needed to know. [...] Whilst Baedeker’s pocket guidebooks have evolved and undergone key changes across the years they still centre on the qualities that first drove Karl Baedeker himself: accuracy, reliability, lucidity, portability, and above all facilitating the traveller with an independence to travel armed with knowledge and advice. (2020, p. 230)

Bolan declares that the “ultimate aim of any guidebook” could be reasonably said to allow travelers to “make destination choices with low perceived risk” (2020, p. 231). As the urtext for guidebooks by Lonely Planet, Rick Steves, Fodor’s, DK, Frommer’s and other relics of the 20th century and early aughts, the Baedeker guide prioritizes accuracy and reliability as ways of avoiding risk and gleaning whatever value there is to be had out of visiting a place.¹¹ Consider how Malin Zillinger describes the function of guidebooks:

¹¹ The Baedeker guides receive negative treatment in Virginia Woolf’s novels and E. M. Forster’s novels. As a case in point, consider Eleanor Lavish’s words to Lucy Honeychurch, the protagonist of *A Room with a*

Guidebooks provide tourists with spatial and social information and hence both identify and popularize places as tourist attractions. Thus, they determine the tourists' starting-points as well as provide vector points in advising and guiding them. Hereby the tourists are given propositional presentations of what the place is like and which attractions are worth seeing and experiencing, thus separating desirable from undesirable experiences. (2006, p. 231)

These characterizations of guidebooks approximate the ideals of the reliabilist accounts of AI trust. The guidebooks aim to lower risk for travelers by being accurate and comprehensive, though also impersonal, taking on a propositional form that may not even result from the guidebook's writers' own experiences. The guidebook, reliable and comprehensive as it may be, has seen a precipitous decline in the 21st century. The big five guidebook publishers in the U.S. saw a 40% drop in sales from 2007 to 2012, and the big seven in the U.K. saw a drop of 46% from 2005 to 2012 (Clampet, 2013). The guidebooks' "waning fortunes [...] are routinely attributed to the rise of online travel information" (Peel & Sørensen, 2016, p. 188). This strikes me as the right diagnosis, but I think the cause relates to something at the heart of our communicative and knowledge practices.

Even as the guidebook technology rose to prominence, readers have always returned to first-person travel accounts and memoirs. Bolan surmises that travelers "appear to trust the views of other tourists more than the views or comments of organisations in the tourism industry" (2020, p. 233). What this preference in trust tracks, I think, is actually a preference for particularity. Anthony Bourdain's wildly popular travel shows also point to the value of particularly above and beyond reliability: in watching Bourdain, we are exposed to a *sui generis* perspective on traveling and eating. When we see Bourdain walk into a five-and-dime store in Santa Fe to eat a Frito pie—Hormel chili and Day-Glo orange cheese sauce scooped into a sliced open bag of Fritos—we are taking in a communicative act that is the unique production of one man's particularity, and our experience of his testimony about the experience becomes as aesthetic as it is epistemic.

Say what you will about social media travel influencers, but I take it that their rise is a way of getting even closer to what we imagine as a gold standard in travel recommendations. On Instagram, one can find the travel influencer that brings together the lived experience

View: "Tut, tut! Miss Lucy! I hope we shall soon emancipate you from Baedeker. He does but touch the surface of things. As to the true Italy—he does not even dream of it. The true Italy is only to be found by patient observation" (Forster, 2000, p. 15). For more analysis of the guidebook in literature, see Jesse Schotter's (2019) excellent article about the limitations and affordances of the travel guidebook through a modernist, literary lens.

of having been somewhere—of having felt the majesty of ascending a lesser-known mountain path at sunset or what it is like to sit in the corner of this precise, dimly lit restaurant—along with a particular character that, I, scrolling Instagram, can spot out and judge as being the sort of character whose views on travel and testimony about experiences are ones I regard highly or poorly. There is a whole cottage industry of travel planners that offer their clients a way to cut past one-size-fits-all agendas to an agenda that is reflective of a singular, and ideally rare, point of view.

We still want good recommendations, ones that are reliable in the sense that they will come through for us, but what an influencer offers above and beyond a guidebook is the demonstrable insights of a particular mind, with its tastes and beliefs that all derive from a personal history that predates the very experience of, say, stepping inside the Blue Mosque. The influencer's videos, photos, and writing function as something more than just evidence for the claims made about what to do in a particular place; they offer windows into that person's sensibilities: what do they sound like? what are they wearing? who were they with? what font and music frame their content? how are they cutting their videos? I want to know *them*, rather than just the content of their recommendations. Influencers offer “freshness, immediacy, interactivity, and personality,” in a way a mass-produced guidebook could never (Bolan, 2020, p. 233). On Bolan's telling of the history of travel writing, the internet has let anyone become a Marco Polo or an Anthony Bourdain (p. 233).

I do not mean to extol influencers or condemn travel books; in fact, I do not wish to take any sort of evaluative stance on the tourism industry, recognizing that it is far from a value-neutral, inconsequential field. Instead, I want to draw out what I think is behind how ordinary people make evaluations of recommendations. While I suggest that the subjectivity offered in an influencer's recommendation is valuable over and above the ostensible objectivity of a guidebook, it is clear that even an influencer cannot offer the kind of thick testimony that is attuned and offered by a friend.

A chatbot, on the other hand, offers none of this particularity. Its outputs are aimed at minimizing predictive loss through a training process that includes ten trillion text tokens. This training, particularly when combined with post-training fine-tuning and reinforcement feedback, does amount to a position, but the position is mysterious, even to a chatbot's designers and engineers. The position inherited from the training data is a positionless position, one that can be expected to tend toward particular results, but without a commitment to those results or a clear genealogy of how those results, rather than others, became the pattern. Its partiality is therefore harder to see and harder to correct, but

regardless of whether or not it could be seen or corrected (post-hoc, on the user's end at least) is not to say that it represents the particularity of a human speaker.

In drawing these conclusions, I am not disavowing the capacity of a chatbot to be reliable and to excel at producing outputs in the form of natural language that would pass muster in a Turing test on markers of articulacy, accuracy, and perhaps even brilliance. It is not at all the point that a person could not reasonably and unknowingly describe a chatbot's output as brilliant or reasonably evaluate its contents in a way that puts them in a position to act well or acquire knowledge. The point is that those markers describe some aspect of a much larger gap between AI outputs and human communication. The gap between a chatbot and a teacher is not just evidentiary or reliabilist, it is normative. And that value gap is, at least as of now, a chasm that cannot be filled. The reliabilist accounts simply do not track this objection.

The value gap is also one that cannot be safely crossed merely by having the right beliefs. This is where the second concern with the reliabilist accounts enters the picture. Nguyen's account of unquestioning trust introduced the notion of agential gullibility. Recall that to trust something on Nguyen's account is to let it in, "to let it muck about directly with one's practical and cognitive innards" (2023, p. 232). The reliabilist accounts promise a way to let the chatbot in far enough to use it without letting it muck about inside us. The accounts assume that if we get epistemically right about the chatbot and its outputs, holding the right beliefs about it (that its outputs are not testimony, that they are not safe, that we ought not to trust them, and so on), we can adopt a bulwark against agential gullibility and stave off the risk of believing things that are untrue. In a phrase, by acknowledging that it is risky, we avoid being put at risk by it.

This seems wrong for two reasons. First, despite lacking the particularity that a human speaker has, a chatbot presents with certain characteristics that are themselves non-neutral. I do not want to walk again over well-trod ground here, but artifacts contain politics inherent to their design (Winner, 1980). Moreover, technologies mediate how we perceive and act whether or not we notice them (Verbeek, 2011). There is no getting around the fact that our engagement with a chatbot presents to us as dialogical and testimonial; we cannot "turn" off our recognition of the output as natural language (cf. Fodor, 1983). In addition to worrying about agential gullibility, then, we should also worry about agential naivety.

Routinely calling upon ChatGPT or Claude to help us inquire, inform our practical agency, and express ourselves entails an integrative stance with it, regardless of how we characterize it when we are thinking as epistemologists. In doing so, we put ourselves at risk of another

outcome that Nguyen (2024) warns of elsewhere, called value capture. In value capture, one has one's values captured and replaced by values that are prioritized by the systems, organizations, pursuits, and objects one embraces or lives by. Nguyen's examples are Fitbits replacing the rich values of exercise with the quantitative value of steps, Yelp replacing the rich values of running a restaurant with the quantitative value of 5-star reviews, and philosophy journal rankings replacing the rich values of writing philosophy with the quantitative values of high-tier-journal publications (Nguyen, 2026). In these cases of value capture, a person's stated values may stay the same, but the external value present in their environments "dominates how they act by setting the practical criteria in day-to-day applications of their values" (Nguyen, 2024, p. 476).

Consider again the star system of Baedeker's asterisks. Visiting sites with an asterisk can begin to guide one's decision making in ways that end up crowding out their original pursuits in travel; travel becomes *about* the star rather than for the pursuit of beauty or novelty, for instance. Nguyen recognizes that "many cases of value capture start as innocuous-seeming uses of a proxy" (2024, p. 477). In relying on chatbots using the reliabilist best practices, we can tell ourselves we are making chatbots work for us and toward our practical aims, but whether this mindset survives routine use is another matter entirely.

Some of the reliabilist accounts name the potential worry of one's values being shaped by a chatbot's values: for instance, Zanotti et al. insist on the "value-laden aspects of the design and use of AI systems" (2024, p. 2691). But the idea goes that so long as the embedded values are good ones, that should "provide us with some guarantee" (Zanotti et al., 2024, p. 2699). Are good values or even good content all that matter when it comes to risk, however? The reliabilist accounts urge us to become chatbot users who are "primed to check information and scrutinize it for errors" (Pritchard, 2026, p. 12). They guard against agential gullibility, which is giving oneself over to people and objects without due consideration of the risks. But this kind of user, the one who knows chatbots can only be tools in certain uses and can never be testifiers, faces the risk of agential naivety, which is the belief that one's correct beliefs about a source protect one from being shaped by it.

Consider another example from life. The understanding that pornography does not depict sex in ways accurate to typical sexual experiences is ubiquitous. There is also a considerable body of literature investigating the effect of viewing porn on viewers' language, beliefs, and behaviors related to sex (Bernstein et al., 2023; Braithwaite et al., 2015; Casper et al., 2026; Gorman, 2014; Sun et al., 2016). Perhaps the most interesting insights in this growing research area are that actual sexual experiences between people do not self-correct

“pornography-induced sexual illiteracy” (Miller & Stubbings-Laverty, 2022, p. 590). The enthusiastic viewer of porn who vociferously contends that because they know that porn is not “real” and may depict views of sex that are violent and misogynistic and chauvinist will not in turn result in the porn they view affecting them or altering their behavior may be acting in a way that is well-informed and nonetheless naïve. The person who uses a chatbot every day for their writing and inquiries, while routinely emphasizing the right beliefs about the technology and affirming their own preexisting values, may very well be in the same state of agential naivety.

Return once more to the Leiter example. Leiter and other philosophers are perhaps best positioned to hold the right beliefs about a chatbot, but if a chatbot is believed to approximate the kinds of feedback on a paper that a live graduate student could, what becomes of our philosophical practices? Most users of chatbots are not writing the kind of material a professional philosopher writes, and most users of chatbots do not have access to even mere first-year graduate students, which is why for many users, a chatbot’s outputs appear to be the best kind of advice, explanation, and expression they see themselves as having access to. Chatbots are not able to offer testimony at all, let alone the thick testimony that is attuned to a person provided by a teacher or assistant.

7 Conclusion

We may feel that in using chatbots, we are just going to get what we need from them, move on, and not be affected by our use. But this is a naïve view, even if we avoid the gullible decision to give our agency over to a chatbot in deeply trusting ways. We risk becoming people, whether as inquirers, writers, or general agents, that turn to chatbots when the value we otherwise seek is to be found elsewhere. As a result, the ultimate answer to when you should trust a chatbot is murky. At a minimum, one should rely on a chatbot only in instances where it is being used with a strictly tool functionality, and even then, only where they are in a position of expertise with regard to whatever the tool is supposed to do for them. This principle of use rules out a vast number of ways in which people are using chatbots daily, which is to write things for them, to develop opinions for them, to generate ideas, to conduct inquiries into the unknown, and to help them express themselves. However we use a chatbot, we should be mindful not only that we may not be getting what we actually want in those areas, but also that we may be letting ourselves be shaped by something in ways that we are naïve to think we are not being shaped. If this all holds true in the inquiry and expression cases, imagine how much graver the value gap and risk of agential naivety must be in moral and emotional cases.

References

- Aharoni, E., Fernandes, S., Brady, D. J., Alexander, C., Criner, M., Queen, K., ... & Crespo, V. (2024). Attributions toward artificial agents in a modified Moral Turing Test. *Scientific reports*, 14(1), 8458.
- Baier, A. (1986). Trust and antitrust. *Ethics*, 96(2), 231-260.
- Baldwin, J. (1963). *The fire next time*. Grolier.
- Baron, S. (2025). Trust, explainability and AI. *Philosophy & Technology*, 38(1), 4.
- Bernstein, S., Warburton, W., Bussey, K., & Sweller, N. (2023). Pressure, preoccupation, and porn: The relationship between internet pornography, gendered attitudes, and sexual coercion in young adults. *Psychology of Popular Media*, 12(2), 159.
- Blanco, S. (2025). Human trust in AI: A relationship beyond reliance. *AI and Ethics*, 5(4), 4167-4180.
- Bolan. P. (2020). The history and influence of travel writing. In A. Baverstock, R. Bradford, and M. Gonzalez (Eds.), *Contemporary Publishing and the Culture of Books* (pp. 225-236). Routledge.
- Braithwaite, S. R., Coulson, G., Keddington, K., & Fincham, F. D. (2015). The influence of pornography on sexual scripts and hooking up among emerging adults in college. *Archives of Sexual Behavior*, 44(1), 111-123.
- Brandom, R. (1994). *Making it explicit: Reasoning, representing, and discursive commitment*. Harvard university press.
- Casper, M., Tydén, T., Sundström-Poromaa, I., & Gyllenberg, F. (2026). Consent, pornography use, and perceived impact on sexual behavior among women. *Acta Obstetricia et Gynecologica Scandinavica*, 105(4), 664-671.
- Chappell, S. G. (2014). The varieties of knowledge in Plato and Aristotle. In *Knowing what to do* (pp. 263-294). Oxford University Press.
- Clampet, J. (2013, March 3) *Lonely Planet and the rapid decline of the printed guidebook*. SKIFT. <http://skift.com/2013/03/04/lonely-planet-and-the-rapid-decline-of-the-printed-guidebook/>
- Conradie, N. H., & Nagel, S. K. (2024). No agent in the machine: Being trustworthy and responsible about AI. *Philosophy & Technology*, 37(2), 72.
- Davis, D., & Walker, G. (2020). In the Lions' Den. *Harvard International Review*, 41(4), 73-77.
- Dennett, D. C. (2013). *Intuition pumps and other tools for thinking*. WW Norton & Company.
- Dorsch, J., Moll, M., & Deroy, O. (2026). Why we need epistemically, not morally, trustworthy AI: Decision support and the need for collective accountability. *Philosophy & Technology*, 39(4), 189.
- Duede, E. (2022). Instruments, agents, and artificial intelligence: novel epistemic categories of reliability. *Synthese*, 200(6), 491.
- Duede, E. (2023). Deep learning opacity in scientific discovery. *Philosophy of Science*, 90(5), 1089–1099.
- Eriksson, J. (2011). Straight talk: conceptions of sincerity in speech. *Philosophical Studies*, 153(2), 213-234.

- Evans, G. (1982). *The varieties of reference*. Oxford University Press.
- Faulkner, P. (2007). A genealogy of trust. *Episteme*, 4(3), 305-321.
- Fodor, J. A. (1983). *The modularity of mind*. MIT press.
- Forster, E. M. (2000). *A room with a view*. Penguin.
- Freiman, O. (2024). Analysis of beliefs acquired from a conversational AI: Instruments-based beliefs, testimony-based beliefs, and technology-based beliefs. *Episteme*, 21(3), 1031-1047.
- Fricker, E. (2026). On the metaphysical and epistemic contrasts between human and AI testimony. *Inquiry*, 69(6), 2859-2884.
- Gorman, S. (2014). Porn sex vs. real sex: Exploring pornography's impact on sexual behaviors, attitudes, and relationships [Doctoral dissertation, Georgia State University]. ProQuest.
- Grindrod, J. (2024). Large language models and linguistic intentionality. *Synthese*, 204(2), 71.
- Hawley, K. (2014). Trust, distrust and commitment. *Noûs*, 48(1), 1-20.
- Heidegger, M. (1962). *Being and time* (J. Macquarrie & E Robinson, Trans.). Harper & Row.
- Hinchman, E. S. (2005). Telling as inviting to trust. *Philosophy and Phenomenological Research*, 70(3), 562-587.
- Holton, R. (1994). Deciding to trust, coming to believe. *Australasian Journal of Philosophy*, 72(1), 63-76.
- Howick, J., Morley, J., & Floridi, L. (2021). An empathy imitation game: Empathy Turing test for care-and chat-bots. *Minds and Machines*, 31(3), 457-461.
- Jones, C. R., & Bergen, B. K. (2025). Large language models pass the Turing test. arXiv preprint arXiv:2503.23674.
- Jones, K. (1996). Trust as an affective attitude. *Ethics*, 107(1), 4-25.
- Koskinen, I. (2024). We have no satisfactory social epistemology of AI-based science. *Social Epistemology*, 38(4), 458-475.
- Lebron, C. J. (2019). Without the loving strains of commitment. In Satz and Levv (Eds.), *Ideas That Matter: Democracy, Justice, Rights* (pp. 91-109). Oxford University Press.
- Leiter, B. (2026, September 28). *An experience with feedback from an LLM*. Leiter Reports. <https://leiterreports.com/2026/09/28/an-experience-with-feedback-from-an-llm/>
- Lott, M., & Hasselberger, W. (2024). Why AGI could not be (just) a tool: goals, life, and general intelligence. *Inquiry*, 1-26.
- MacIntyre, A. (1999). Social structures and their threats to moral agency. *Philosophy*, 74(3), 311-329.
- Zillinger, M. (2006) The importance of guidebooks for the choice of tourist sites: A study of German tourists in Sweden. *Scandinavian Journal of Hospitality and Tourism*, 6(3), 229-247.
- Mallory, F. (2023). Fictionalism about chatbots. *Ergo*, 10(38), 1082-1100.
- Manson, N. C., & O'Neill, O. (2007). *Rethinking informed consent in bioethics*. Cambridge University Press.

- Miller, D. J., & Stubbings-Laverty, R. (2022). Does pornography misinform consumers? The association between pornography use and porn-congruent sexual health beliefs. *Sexes*, 3(4), 578-592.
- Millikan, R. G. (2005). *Language: A biological model*. Oxford University Press.
- Mogensen, A. L. (2017). Moral testimony pessimism and the uncertain value of authenticity. *Philosophy and Phenomenological Research*, 95(2), 261-284.
- Nguyen, C. T. (2023). Trust as an unquestioning attitude. In T. S. Gendler, J. Hawthorne, and J. Chung (Eds.), *Oxford studies in epistemology* (vol. 7) (pp. 214-244). Oxford University Press.
- Nguyen, C. T. (2024). Value capture. *Journal of Ethics & Social Philosophy*, 27(3), 469-504.
- Nguyen, C. T. (2026). *The score*. Penguin.
- Nickel, P. J. (2022). Trust in medical artificial intelligence: a discretionary account. *Ethics and Information Technology*, 24(1), 7.
- O'Neill, O. (2002). *A question of trust: The BBC Reith Lectures 2002*. Cambridge University Press.
- Ortmann, J. (2025). Of opaque oracles: epistemic dependence on AI in science poses no novel problems for social epistemology. *Synthese*, 205(2), 80.
- Peel, V., & Sørensen, A. (2016). *Exploring the use and impact of travel*. Channel View Publications.
- Pritchard, D. (2026). Generative AI is untrustworthy. *Synthese*, 208(2), 97.
- Ransom, M., & Menard, N. (2026). A dilemma for skeptics of trustworthy AI. *AI and Ethics*, 6(5), 538.
- Rini, R. (2020). Deepfakes and the epistemic backstop. *Philosophers' Imprint*, 20(24), 1-16.
- Robertson, I. (2025). AI, trust and reliability. *Philosophy & Technology*, 38(3), 94.
- Schotter, J. (2019). "Objects worthy of attention": Modernism and the travel guide. *Modernism/modernity Print Plus*, 4(2).
- Sikimić, V. (2026). Navigating AI-accelerated science: An intellectual virtue approach to responsible inquiry. *Episteme*, 1-18.
- Simpson, T. W. (2012). What is trust?. *Pacific Philosophical Quarterly*, 93(4), 550-569.
- Sosa, E. (2006). Knowledge: Instrumental and testimonial. In J. Lackey and E. Sosa (Eds.), *The epistemology of testimony* (pp. 116-123). Oxford University Press.
- Sun, C., Bridges, A., Johnson, J. A., & Ezzell, M. B. (2016). Pornography and the male sexual script: An analysis of consumption and sexual relations. *Archives of Sexual Behavior*, 45(4), 983-994.
- Susskind, R. (2025). *How to think about AI: A guide for the perplexed*. Oxford University Press.
- Verbeek, P. (2011). *Moralizing technology*. University of Chicago Press.
- Viehoff, J. (2023). Making trust safe for AI? Non-agential trust as a conceptual engineering problem. *Philosophy & Technology*, 36(4), 64.
- Williams, I., & Bayne, T. (2026). Chatting with bots: AI, speech acts, and the edge of assertion. *Inquiry*, 69(6), 3059-3082.
- Winner, L. (1980). Do artifacts have politics?. *Daedalus*, 109(1), 121-136.
- Zanotti, G., Petrolo, M., Chiffi, D., & Schiaffonati, V. (2024). Keep trusting! A plea for the notion of trustworthy AI. *AI & Society*, 39(6), 2691-2702.