https://www.fecundity.com/job

21may2024

Forthcoming in *Episteme*.

On trusting chatbots

P.D. Magnus[1]

ABSTRACT: This paper focuses on the epistemic situation one faces when using an LLM-based chatbot like ChatGPT: When reading the output of the chatbot, how should one decide whether or not to believe it? By surveying strategies we use with other, more familiar sources of information, I argue that chatbots present a novel challenge. This makes the question of how one could trust a chatbot especially vexing.

Since the release of ChatGPT in late 2022, there has been lots of distressed discussion about chatbots. There are diverse worries that students will use the chatbots to cheat on assignments, that chatbots provide a roundabout way of plagiarizing material that was used to train the AI model, that short-sighted businesses will fire writers and replace them with chatbots, that boilerplate output of chatbots will be posted all over the internet and crowd-out better content, and so on. My concern here is not with the whole range of distress. My interest is in the more narrow question of whether and when we ought to believe the output of chatbots.

Sometimes an analogy is drawn between the current panic about chatbots and the panic about Wikipedia almost twenty years ago. A 2006 *New York Times* article claimed that Wikipedia raised "a single nagging epistemological question: Can an article be judged as credible without knowing its author?" (Stross 2006) Text in Wikipedia is edited directly by many human authors, even if the reader

[1] Thanks to Ron McClamrock, Jason D'Cruz, Alessandra Buccella, students in my Theory of Knowledge class, and anonymous referees for feedback on earlier versions of this paper.

does not know who they are. Although a chatbot is trained on the work of human authors, the output is most directly the result of the algorithm. So, regarding chatbots, the parallel question is whether a claim can be judged as credible when there is no author who has asserted it.

Recognizing that it is a mistake to think of a chatbot as a cognitive agent— *a fortiori*, as a possible author— Eunice Yiu and collaborators suggest instead that chatbots and the algorithms that drive them are "powerful new cultural technologies, analogous to earlier technologies such as writing, print, libraries, the Internet, and even language itself." This framing still suggests that chatbots can serve as sources of information. Indeed, they add, "These AI systems offer a new means for cultural production and evolution, allowing information to be passed efficiently from one group of people to another" (Yiu et al. 2023).

The two framings— of chatbots as agents which can make assertions, and of chatbots as technologies for the transmission of information— share the presupposition that the output of chatbots is or could be worth believing. My task here is to think through this presupposition. In a quotidian sense, taking someone at their word is *trusting* them. In that sense, believing the output of a chatbot would be trusting the chatbot. So the question is whether, when, and how to trust a chatbot.

After some general remarks, I consider several strategies which we use in evaluating information sources. I argue that these do not readily apply to the output of chatbots. My conclusion is a cautionary one, that we need better strategies for engaging with chatbots.

What are chatbots? What is trust?

A Large Language Model (LLM) is a machine learning algorithm trained on large collections of text. It is, in the vivid metaphor of Emily Bender and collaborators, a stochastic parrot (Bender *et al.* 2021). Given a string of words, the algorithm predicts what word is likely to occur next. Stitched together in an appropriate way, predictions of the next word become predictions of whole sentences, paragraphs, and longer stretches of text.

OpenAI's ChatGPT, Google's Bard, Microsoft's Copilot and similar services provide a user interface for LLMs. Although there are other technologies used to construct chatbots, LLM chatbots are my target here. The user can enter a question or prompt, the service generates a response, the user can enter a further prompt, the service generates a further response, and so on. The text generated by the chatbot reflects not just the most recent prompt but also earlier parts of the dialogue. Chatbots of this kind are being incorporated into internet search engines, word processors, and lots of other applications.

Although there is much philosophical work on the nature of *trust*, there are certainly kinds of trust which are not relevant to trusting a chatbot— a point I return to in the next section. For now, note that I do trust the outputs of machines in lots of ordinary contexts. For example, the website for an on-line retailer says that they have none of the product I want in stock. On that basis, I believe that they have none in stock. This is fallible, of course, and I would not be shocked to find out that the website was incorrect. But for ordinary purposes, it suffices. The site says that they are out, I form the belief that they are out, I make decisions on the basis of my belief, and if anyone asks I would report my belief. My question, then, is whether I could trust the output of a chatbot even to that extent. Without independent confirmation from some other source, should I believe what I read in the chat window?

Discussions of trust in AI are often posed from the perspective of developers. The question is how an engineer or technologist should approach design so that their products can be trustworthy. As Kush Varshney poses it, the challenge is that potential users "do not trust artificial intelligence and machine learning in critical enterprise workflows because of a lack of transparency into the inner workings and a potential lack of reliability" (Varshney 2022, p. 2). This suggests that the two loci of trustworthiness are *transparency* and *reliability*.[2] Do these solve the problem from my point of view as an individual user or would-be user?

[2] Both concepts have received attention from philosophers. Reliability especially has been a commonplace of epistemology at least since the watershed of Goldman (1979). Regarding the entanglement of transparency and reliability in the context of AI, see *inter alia* Ryan (2020), von Eschenbach (2021), and Alvarado (2023).

Regarding transparency: My description above of how an LLM works is non-technical. Further details of the algorithm used by an LLM would not fill the gap, however, because the quality of the output also depends on the data used to train it. The data used to train major chatbots include large swaths of the internet and some undisclosed collection of other sources. Moreover, they include human feedback used to nudge the quality and tone of the output. For example, Open AI hired Kenyan workers through a subcontractor to tag offensive and toxic passages, to increase the propriety of their model's output (Perrigo 2023). So the limits of transparency are two-fold. First, many of the details are proprietary. Second, even if I were to know all the details, I still could not form a sensible expectation of how likely the chatbot output is to be true. The system would still largely be a black box.

Regarding reliability: It is common to understand the reliability of a process in terms of the truth-ratio of its outputs. For chatbots, however, it would not make sense to report a categorical statistic like a percentage of true versus false outputs or a rate of falsehoods per 1000 words. The actual rates will depend on the particular chatbot. Even if ChatGPT were reliable to such-and-so degree, nothing follows about Bard or any of the others. And specific chatbots themselves are moving targets. Between occasions when I interact with a chatbot— even between major updates— programmers and engineers may have tweaked parameters, provided more training, or modified the underlying algorithm. When a chatbot is publicly denounced for some terrible answer and later stops giving that answer, perhaps it has improved in a general sense— or perhaps it was just tweaked so as to avoid that particular pitfall.

The output of chatbots is highly variable in the further sense that it depends on context, both the topic in question and the specific way the prompt is posed. Everybody has anecdotes, but it is hard to say anything synoptic. Any statistical claims will be sensitive to the algorithm used, when it is used, the topic asked about, and subtle priming effects in the prompt. Rather than trying to resolve this as a

purely empirical question, then, let's take a step back and consider the epistemic situation I am in when I encounter a chatbot's claim and consider whether or not I should believe it.

How to think about LLM-generated content

One approach to epistemology begins with the idea that the details of justifying a belief depend on its source. Knowledge, on such an approach, is grounded in many specific sources such as perception, memory, introspection, inference, reason, or testimony. (See, e.g. Audi 1998.) From that list, the output of an LLM is most like testimony. An expert or witness might write some sentences relaying facts to me, and it is possible to read the output of an LLM as if it were sentences of that kind.

Of course, there are important differences. Take a mundane example of testimony, like when a student tells me that they are from Vermont. I believe on that basis that they are from Vermont. I presuppose, without thinking about it, that they know where they are from and that they are sincerely recounting that to me. Although it might be overblown to say that I am *accepting their testimony* in such a banal case, I take them at their word. Formal testimony in a court of law has a similar epistemic structure. It relies on or presupposes both epistemic and moral facts about the person providing testimony; that is, that they have knowledge (secured by some source or other) and that they are being sincere. It is standard in the philosophy of testimony to hold that "epistemic trustworthiness requires the conjunction of competence and sincerity" (Fricker 2007, p. 45). For recent discussions of relationally responsive trust, see Almassi (2022) and the introduction to Nguyen (2022).

Chatbots do not have beliefs in the usual sense, which undermines the conditions for both knowledge and sincerity.[3] Knowledge on most accounts requires belief, and sincerity requires a separation between what one believes and what one says. Since the output of a chatbot lacks these dimensions which are typical of testimony and trust in human sources, it makes no sense to use

[3] See Shanahan (2023). Whether LLMs can have beliefs will depend in part on the ultimate nature of belief, a philosophical issue that I do not attempt to resolve here.

relational accounts of trust when asking about the trustworthiness of chatbots.[4] If someone were to argue that chatbots do have knowledge (in some sense) and can be sincere (in some sense), the burden of proof is on them to show that these are relevantly like the knowledge and sincerity required by relational accounts of trust.

Fintan Mallory (2023) argues that engagement with chatbots involves a form of make-believe. In reading output from a chatbot as if it were meaningful assertion, we are engaging in an imaginative pretense. One might extend Mallory's account, so that this pretense extends beyond communicative intentions to the cognitive and moral states required for relationally responsive trust. Certainly, we *could* make-believe that chatbots can be trusted in that way— but why should we? Mallory is explicit that his account does not answer the epistemic question of whether we actually ought to believe chatbot output.

C. Thi Nguyen (2022) elaborates a sense in which it is possible to trust objects which lack the human features required for relationally responsive trust— namely, adopting an unquestioning attitude toward those objects. If we are going to believe the outputs of chatbots, it should be on the basis of reflective considerations. An unquestioning attitude, in contrast, is "a disposition against deliberating." So, although we *cannot* trust chatbots in the relationally responsive sense, we *ought* not extend them unquestioning credulity either.

Nevertheless, reading the output of a chatbot has the bare structure of reading a claim on-line. It is, from the user's point of view, like reading a claim made on social media by some indefinite other user, like following a link and reading a claim from a random website, or like reading a claim in a Wikipedia article. The user is confronted with a claim, they have no real relationship with the source of that claim, and they will either end up believing it or not.[5]

[4] This line of argument is developed at greater length by Freiman (2023a, 2024).

[5] If they are a cautious Bayesian who refuses to have categorical beliefs, they still end up with some degree of belief. I put the points in terms of categorical belief just for ease of exposition.

There are numerous rubrics we might use to describe this phenomenon. Ori Freiman and Boaz Miller (2020; Freiman 2024) describe the outputs of AI as *quasi-testimony*. Freiman (2023b) distinguishes beliefs formed in this way as *technology-based* rather than as testimony-based. Ramón Alvarado (2023) characterizes accepting AI claims as *epistemic trust*. Regardless of what we call it, what methods does the user have to navigate such an encounter?

In earlier work (Magnus 2009), I provide a list of several strategies we commonly use in such situations, deciding whether or not to believe claims we find on-line. Let's consider some of them, to ask how well each works when applied to the output of a chatbot.

Method 1: Authority

One reason to accept a claim is that it comes from an established and authoritative source, and a corresponding reason to not accept a claim is if it comes from a shady or disreputable source. Call this an *appeal to authority*. Could a chatbot have authority in this sense?

LLMs can generate vividly articulated falsehoods. The phenomenon is standardly called *hallucination*, although that term carries some connotations that can be misleading. A number of writers have described it instead as *bullshit*, in the distinct technical sense defined by Harry Frankfurt (2005).[6] That is, an utterance is bullshit when it is produced without any concern for truth or falsity. Even if it turns out to be false, it is not an outright lie— because a lie is told with explicit consideration that the thing said is false. Bullshit is produced with an indifference to the truth.

The striking thing about chatbots is that they cannot possibly have a concern for the truth. Even if told to "be careful and answer truthfully", these words are just more linguistic items in the prompt. The chatbot's reply is statistically nudged by those terms in a different direction, but through the same space it navigates in response to any prompt. All it has is a model of discourse. It has no model of the world, no sense of *how things are* which discourse could be trying to describe. It can produce

[6] See *inter alia* Narayanan and Kapoor (2022), Sparrow, Koplin, and Flenady (2023), White and Skorburg (2023), and Roy and Maity (2023). Hannigan, McCarthy, and Spicer (forthcoming) coin the neologism *botshit* for AI-generated bullshit put to work, i.e. "LLM generated hallucinatory content that a human uncritically uses for a task."

sentences with words like “true” and “false” in them which are like sentences a human would produce, but it has no notions of truth or falsity. So the sentences it produces could not be anything besides bullshit.

The problem, fundamentally, is that a chatbot only has a map of discourse without any independent connection to a world that discourse is about. All it has are words. As Dan Li (2023) argues, the success of a machine learning algorithm depends on the choice of ontology in constructing it. To apply that general point, the problem with chatbots is that they have the wrong ontology for solving problems that are about more than just words. I mean *ontology* here both in the computer scientists’ sense (data ontology, the form of the algorithm’s internal representations) and in the philosophers’ sense (ontology as an account of what there is).

To illustrate this point in a non-technical way, imagine training a computer to play chess. Suppose you adopt an LLM for this task and train it on discourse and commentary about chess: transcripts of games, but also commentary about games, chess adjacent conversations, and chatter among chess aficionados. Now imagine asking that resulting model to play a game. It might play a decent game of chess— at least some of the time— but it will also sometimes go off the rails. It only represents the game state and the way that the pieces move alongside all of the discourse about chess, and all those stochastic nudges will sometimes lead it to make illegal moves. Other times it might fail to make a move at all. Perhaps instead it describes an anecdote purportedly about Bobby Fischer. If you want a machine-learning algorithm to play chess, it would be better to make its ontology correspond to the game directly: the board state, the permissible ways to move the pieces, and so on.

One might object that it is in fact possible for an LLM to have an internal model of a game like chess. Compare work by Kenneth Li and collaborators (Li et al. 2023) modeling the boardgame Othello.[7] They begin with GPT, the same algorithm used by ChatGPT, and train it on games of Othello recorded in standard notation. The resulting model is pretty good (although not perfect) at

[7] They pick Othello because it “is simpler than chess, but maintains a sufficiently large game tree to avoid memorization” (Li et al. 2023, p. 2). Kenneth Li (2023) provides an informal discussion of their result.

suggesting legal moves. More impressively, they provide compelling evidence that the model has an internal representation of the state of the board when it recommends a next move. However, it is important to note that their Othello-GPT model is trained only on games of Othello. The training set does not include even discourse about Othello beyond the notation for games. So it is unlike the training set for a chatbot, which might include games of Othello but will also include language about all sorts of other things. A chatbot's response when prompted to recommend a move in a game of Othello will be influenced not just by whatever games of Othello were in its training set, but by debates about the history of Othello and the earlier game Reversi, by analysis of Shakespeare's Othello, and by who knows what else. If one wants recommendations for moves to make in a game of Othello, one is better off asking Othello-GPT than ChatGPT. And there are better algorithms to consult than that. Since their study, the game of Othello has been solved (Takizawa 2023). Although Othello-GPT shows how a chatbot can develop an internal model of sorts for Othello or chess, it is merely a virtual model inside its model of all of discourse. That lesson is entirely general. A chatbot has no model of anything apart from its model of discourse.

One might object that the examples of chess and Othello are misleading because they are games with precisely defined rules. So it is natural to think that a computer should be programmed directly with the game state and possible moves. The promising uses for chatbots (one imagines) will not be like that. Consider, however, the suggestion that chatbots should help with academic writing— e.g., in preparing the straightforward parts of papers like the initial literature review. A group of bioethicists write that "it might become an academic norm or even a policy requirement that one can use [an LLM] to generate introductions, conclusions, or background sections in which a more or less rote synthesis of existing ideas and scholarship is necessary, while still needing to manually develop the bulk of the substantively new material" (Porsdam Mann et al. 2023, p. 38).[8] The problem with this suggestion is that a literature review is not merely a new bit of writing that riffs on earlier writing.

[8] Huang and Tan (2023) make a similar suggestion.

Instead, it is supposed to cite actual published sources. If we want to provide verbatim quotations and genuine citations, then asking an LLM trained on all of academic discourse to do the job is like constructing a chess program from an LLM trained on discourse about chess. The procedure for citation and direct quotation is much more old school. It involves consulting sources. On a computer, it involves ordinary database operations to query library records.

Even if chatbots could get pretty good at playing chess or quoting sources under typical circumstances, it is inevitable that some particular conversational contexts will lead them astray. At a fundamental level, the way that the chatbot represents information treats all discourse the same. When it begins a sentence by saying that it is quoting a source, the only thing that distinguishes passages from the source it is quoting and other bits of discourse are weights somewhere inside— differences in degree that can be overwhelmed by other contextual factors. A vivid example is provided by a case from the summer of 2023, when two lawyers were fined for citing what the judge called "non-existent judicial opinions with fake quotes and citations created by the artificial intelligence tool ChatGPT" (Neumister 2023). One of the lawyers had asked the chatbot whether the cases were real, and it replied that it had double-checked to confirm that they were in legal databases. This reassurance was despite the fact that ChatGPT had not consulted the databases and had no way of doing so (Maruf 2023).

To sum up, a chatbot is just not the sort of thing which could be an authority.

Method 2: Plausibility of style

When evaluating sources of information, we often take into account not just *what* is said but also *how* it is said. If I read an account of developments in high energy physics, for example, there are stylistic markers which suggest that the writer knows something about physics. Roughly speaking, they can write like a physicist, and I can take this as some reason to believe them. Call this an *appeal to plausibility of style*.

Plausibility of style relies on what Harry Collins and Robert Evans (2007) call interactional expertise: proficiency in the language of a subject. Authority, in the sense discussed in the previous

section, is more akin to what Collins and Evans call contributory expertise. Of course, it is possible to have interactional expertise without much real knowledge of the topic. Some people develop a glib facility with a particular kind of discourse without actually understanding it. However, the kinds of expertise tend to be at least somewhat correlated in ordinary humans. In the course of acquiring interactional expertise on some topic, one often learns something about the actual content as well. This correlation between fluency and substantive expertise lets a reader use plausible style as a sign of trustworthiness.

The problem is that chatbots break this typical correlation. Because an LLM is trained to produce outputs that look like existing discourse, it can generate prose in the style of real experts. Because it has no real model of the domain in question, though, it is indifferent to whether the contents of its outputs actually reflect what the domain is like or not. Its outputs will have the stylistic and expressive features of expert claims independently of whether they are true. As Jacob Browning and Yann LeCun put it, an LLM "is simply indifferent to matters of truth; it is simply attempting to produce the right *kind* of answer for the situation. This often provides users with the clear and readable responses one might find in a textbook or on Wikipedia, but often with errors, omissions, and confusions that are obvious to an expert but easy to miss for anyone else" (2023, italics in original).

To sum up, a chatbot producing output in a plausible style is no indication that it can be trusted.

Method 3: Plausibility of content

When considering a claim, it is possible to assess the plausibility of content itself. This is sometimes called a sniff test or sanity check. Even if I do not know about a particular topic in detail, I might know enough to judge that a particular claim is too wild to be true. Conversely, something merely sounding plausible is not sufficient reason to believe it, especially if I know that it is expressed merely as a guess or a conjecture. However, if someone sincerely asserts a plausible claim, then that might be enough reason for me to believe it— especially if the stakes are low. So *appeal to plausibility of*

content can be used both in a negative way (to discount implausible answers) and in a positive way (to accept plausible ones).

The negative use of appeal to plausibility can readily be applied to the output from chatbots. Sometimes they make claims that obviously are not true. As an example, consider exchanges I had with ChatGPT about the *James-Rudner-Douglas thesis*. In one exchange, it offers this dubious sentence: "This idea is named after the American philosopher and psychologist William James, the British philosopher Karl Popper, and the American philosopher C. I. Lewis, who are among the most well-known proponents of this idea." One does not need to know anything more about the thesis than its name to see that this is not true. The moniker alone reveals that the second eponym was Rudner (not Popper) and that the third was Douglas (not Lewis). On the basis of this sentence alone, one might discount everything else it says about the thesis.

The positive use appeal to plausibility, however, is frustrated by the fact that the chatbot can make no distinction between conjecture and assertion. It produces strings of linguistic output without communicative intentions. Unable to differentiate between different kinds of speech acts, it provides no warning that it is just speculating or spitballing ideas. What would be expressions of timidity or confidence in language produced by an actual human are just parameters in its operation, set independently of the claims made.

To sum up, plausibility of content can only be used in a negative way with the output of a chatbot, to discount the ridiculous.

Method 4: Calibration

Sometimes when consulting a source, I know something about the topic in question. I want to use the source to extend my knowledge. In such cases, I can check that the source gives the right answers on the parts of the topic that I already know about. If it gets those right, I trust it on matters which I did not antecedently know. Call this *calibration*.[9]

[9] Calibration can be formulated as an inductive argument from a track record of claims that the source gets right to the conclusion that the source is generally reliable.

A chatbot typically does not generate a response just based on the most recent prompt, but instead on the entire chat session up to that point. One might hope this would make calibration a useful method for deciding whether to believe a particular output or not. Unfortunately, the chatbot's pool of information is a heterogeneous grabbag. So earlier parts of the session can prime it to generate false output which it would not generate from a cold start.

Return to the example of ChatGPT on the *James-Rudner-Douglas thesis*. Abbreviated as the JRD thesis, the label is used in only a handful of papers.[10] So it is little surprise that, when asked to describe the James-Rudner-Douglas thesis, the chatbot answers that it does not have any information about it. However, the thesis is related to the much-discussed Argument from Inductive Risk. When asked first about inductive risk and then about the relationship between inductive risk and the James-Rudner-Douglas thesis, ChatGPT produces several paragraphs on the subject. Often its answers are terrible. When it attributes the thesis to some James, some Rudner, and some Douglas, however, the answer may sound plausible enough. And its answer to the initial question, about inductive risk, can be pretty good even when its answer about the JRD thesis is flat wrong. In such a case, if one knew something about inductive risk and tried to use the chatbot to learn about the JRD thesis, calibration would lead one astray.

For an ordinary source, calibration relies on the fact that it would be a coincidence for the reliability of the source to run out just at the point where my own knowledge runs out. A chatbot does not have expertise in the usual sense, though, and its reliability on one topic is little indication of what is likely to happen next.

To sum up, calibration can easily go wrong when applied to the output of a chatbot.

[10] Magnus (2013, 2014, 2018). The details of what the thesis actually is are irrelevant to the example. This particular example worked with ChatGPT 3.5 from its release through all of 2023, although the exact output varied and depended on the details of the prompts.

Method 5: Sampling

I might consult several different sources, believing only claims that appear in more than one source. Call this *sampling*.

The value of sampling depends on the range of sources consulted. If I consult several chatbots or the same chatbot several times, my situation is little better than if I just consult one. Recall the lawyer who was assured by ChatGPT that non-existent cases were in legal databases. If I consult a chatbot and a range of more traditional sources, though, why consult the chatbot at all?

In any case, sampling takes time and attention.

More than a decade ago, Boaz Miller and Isaac Record argued that responsible use of the internet may require "gaining information from traditional media to supplement internet-filtered information in order to help determine whether it is biased or incomplete" (2013, p. 132). Nevertheless, most of us have come to rely on internet sources without checking them against old-school media. So it seems likely that— if we come to use AI chatbots as sources of information— we will often go no further after consulting them.

In cases where I just plan to consult one source, it is unclear why that one source should be a chatbot. The internet provides ready access to many alternatives. They are less exciting, perhaps, but also more trustworthy. As chatbots proliferate across internet services like search engines, however, we will consult them on questions where we want a quick answer and have no intention of taking a deep dive into multiple sources.

Conclusion

There may be other methods besides just these five, and the methods can be used in combination. So my survey of strategies here does not decisively prove that chatbots cannot be trusted. The upshot of the preceding discussion is rather that chatbots frustrate many of the methods we

ordinarily use when deciding whether or not to believe claims. If we are going to use chatbots, it is unclear how we ought to do so.

One reply would be to say that chatbots should never be trusted but that they should be used in some other way. The viability of this reply depends on the use that is imagined. A common suggestion is that a chatbot could be used to brainstorm topics, to suggest directions, and to generate early drafts— then I could edit the output with attention to accuracy. Note, however, that this asks me to edit in a different way than I do when editing thesis statements and early drafts that I have written myself. Say what you will about the roughness of my own drafts, they are explicit about the difference between reporting facts and speculating. They are not, in the technical sense, bullshit. Editing the output of a chatbot requires either believing it or fact-checking it, and those tasks are vexed precisely for the reasons discussed here.

As I noted at the outset, some of the current worries about chatbots parallel worries about Wikipedia from decades ago. To some extent, those worries were never answered. Rather, we got over them. We decided, as Don Fallis writes, that "Wikipedia is not all that bad" (Fallis 2011). However, this progression was not simply a matter of lowering our standards to make space for Wikipedia. Wikipedia has a culture and explicit, self-conscious policies. There were changes to Wikipedia policies that made it harder to introduce misinformation into it.

An LLM can be nudged toward or away from certain kinds of outputs by training on human feedback, but that training is just adjusting the weights it developed after all of its other training. So it is unclear what changes could be made to chatbots going forward, to answer the current concerns about them. And even if we could specify what such developments would look like, it is unclear whether the companies investing in chatbots and AI would have any interest in implementing them.

Harry Collins, thinking about the sociology of knowledge, suggests this dictum: "Never treat an instrument as an intelligent actor or a participant in social knowledge unless you have a very good methodological reason for doing it— a reason that one could and should explain clearly" (Collins

2010, p. 146). Applying the dictum here: Do not trust chatbots unless you have a reason to do so that you can explain clearly. The demand for explanation goes beyond the demand for transparency. It is not enough just to know how chatbots work. The demand instead is to explain, in the context of our epistemic practices, why these novel instruments would count as legitimate sources. My argument here has been largely negative— that this explanation cannot be given in terms of familiar methods and practices. Even if chatbots are a cultural technology analogous to others before it, the analogy alone does not tell us that we can trust them.

Works cited

- Almassi, Ben. 2022. "Relationally Responsive Expert Trustworthiness." *Social Epistemology*. 36(5): 576-585.
- Alvarado, Ramón. 2023. "What kind of trust does AI deserve, if any?" *AI and Ethics*. 3(4): 1169-1183. DOI: 10.1007/s43681-022-00224-x
- Audi, Robert. 1998. *Epistemology: A contemporary introduction to the theory of knowledge*. London and New York: Routledge.
- Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, Shmargaret Shmitchell. 2021. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. March 2021: 610–623. DOI: 10.1145/3442188.3445922
- Browning, Jacob and Yann LeCun. 2023. "Language, common sense, and the Winograd schema challenge." *Artificial Intelligence.* 325: 104031. DOI: 10.1016/j.artint.2023.104031
- Collins, Harry. 2010. "Humans not instruments." *Spontaneous Generations: A Journal for the History and Philosophy of Science.* 4(1): 138-147.
- Collins, Harry and Robert Evans. 2007. *Rethinking expertise*. University of Chicago Press.

- Fallis, Don. 2011. "Wikipistemology," In: Alvin I. Goldman and Dennis Whitcomb (editors). *Social epistemology: Essential readings.* Oxford: Oxford University Press, pp. 297–313.
- Frankfurt, Harry G. 2005. *On Bullshit*. Princeton University Press. Republication of an essay from 1986.
- Freiman, Ori, and Boaz Miller. 2020. "Can Artificial Entities Assert?" in Sanford Goldberg (ed.), *The Oxford Handbook of Assertion*. Oxford: Oxford University Press, pp. 413-434. DOI: 10.1093/oxfordhb/9780190675233.013.36
- Freiman, Ori. 2023a. "Making sense of the conceptual nonsense 'trustworthy AI'." *AI and Ethics.* 3: 1351–1360. DOI: 10.1007/s43681-022-00241-w
- Freiman, Ori. 2023b. "Analysis of Beliefs Acquired from a Conversational AI: Instrument-Based Beliefs, Testimony-Based Beliefs, and Technology-Based Beliefs." *Episteme*. Published online: 1–17. DOI: 10.1017/epi.2023.12
- Freiman, Ori. 2024. "AI-Testimony, Conversational AIs and Our Anthropocentric Theory of Testimony." *Social Epistemology*. DOI: 10.1080/02691728.2024.2316622
- Fricker, Miranda. 2007. *Epistemic Injustice: Power & the Ethics of Knowing*. Oxford University Press.
- Goldman, Alvin I. 1979. "What Is Justified Belief?" in George S. Pappas (ed.), *Justification and Knowledge: New Studies in Epistemology*, Dordrecht: Reidel, pp. 1–25
- Hannigan, Timothy and Ian P. McCarthy and Andre Spicer. forthcoming. "Beware of Botshit: How to Manage the Epistemic Risks of Generative Chatbots" *Business Horizons.* https://ssrn.com/abstract=4678265
- Huang, Jingshan and Ming Tan. 2023. "The role of ChatGPT in scientific communication: writing better scientific review articles." *American Journal of Cancer Research*. 13(4): 1148-1154.

- Li, Dan. 2023. “Machines Learn Better with Better Data Ontology: Lessons from Philosophy of Induction and Machine Learning Practice.” *Minds and Machines* 33 (3): 429-450.
- Li, Kenneth. 2023. “Do Large Language Models learn world models or just surface statistics?” *The Gradient*. https://thegradient.pub/othello
- Li, Kenneth, Aspen K Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. “Emergent world representations: Exploring a sequence model trained on a synthetic task.” *International Conference on Learning Representations*. arXiv:2210.13382
- Mallory, Fintan. 2023. “Fictionalism about Chatbots.” *Ergo*.10: 38. doi: 10.3998/ergo.4668
- Magnus, P.D. 2009. “On trusting Wikipedia.” *Episteme*. 6(1): 74-90.
- Magnus, P.D. 2013. “What scientists know is not a function of what scientists know.” *Philosophy of Science*. 80(5): 840–849.
- Magnus, P.D. 2014. “Science and rationality for one and all.” *Ergo*. 1(5): 129–138.
- Magnus, P.D. 2018. “Science, values, and the priority of evidence.” *Logos&Episteme*. 9(4): 413–431.
- Maruf, Ramishah. 2023. “Lawyer apologizes for fake court citations from ChatGPT.” *CNN*. May 28. https://www.cnn.com/2023/05/27/business/chat-gpt-avianca-mata-lawyers/
- Miller, Boaz and Isaac Record. 2013. “Justified Belief in a Digital Age: On the Epistemic Implications of Secret Internet Technologies.” *Episteme*. 10(2): 117-134. DOI: 10.1017/epi.2013.11
- Neumister, Larry. 2023. “Lawyers submitted bogus case law created by ChatGPT. A judge fined them $5,000.” *The Associated Press*. June 22. https://apnews.com/article/artificial-intelligence-chatgpt-fake-case-lawyers-d6ae9fa79d0542db9e1455397aef381c

- Narayanan, Arvind and Sayash Kapoor. 2022. "ChatGPT is a bullshit generator. But it can still be amazingly useful." *AI Snake Oil.* December 6. https://www.aisnakeoil.com/p/chatgpt-is-a-bullshit-generator-but
- Nguyen, C. Thi. 2022. "Trust as an Unquestioning Attitude" in Tamar Szabó Gendler, John Hawthorne, and Julianne Chung (eds), *Oxford Studies in Epistemology*, Volume 7, online edition, DOI: 10.1093/oso/9780192868978.003.0007
- Perrigo, Billy. 2023. "Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic." *Time*. January 18. Accessed March 20, 2024. https://time.com/6247678/openai-chatgpt-kenya-workers/
- Porsdam Mann, Sebastian, Brian D. Earp, Nikolaj Møller, Suren Vynn, and Julian Savulescu. 2023. "AUTOGEN: A Personalized Large Language Model for Academic Enhancement—Ethics and Proof of Principle." *American Journal of Bioethics*. 23(10): 28-41.
- Roy, Nandita and Moutusy Maity. 2023. "'An Infinite Deal of Nothing': critical ruminations on ChatGPT and the politics of language." *Decision* 50(1): 11-17. DOI: 10.1007/s40622-023-00342-3
- Ryan, Mark. 2020. "In AI We Trust: Ethics, Artificial Intelligence, and Reliability." *Science and Engineering Ethics*. 26, 2749–2767. DOI: 10.1007/s11948-020-00228-y
- Stross, Randall. 2006. "Anonymous Source Is Not the Same as Open Source." New York Times, March 12. http://www.nytimes.com/2006/03/12/business/yourmoney/12digi.html
- Shanahan, Murray. 2023. "Talking about Large Language Models." arXiv:2212.03551v5 [cs.CL]
- Sparrow, Robert, Julian Koplin, and Gene Flenady. 2023. "Generative AI is dangerous — but not for the reasons you might think." ABC's *Religion and Ethics* portal. February 22. https://www.abc.net.au/religion/why-generative-ai-like-chatgpt-is-bullshit/102010238
- Takizawa, Hiroki. 2023. "Othello is solved." arXiv:2310.19387v1 [cs.AI]

- White, Dylan J. and Joshua August (Gus) Skorburg. 2023. "ChatGPT killed the student essay? Philosophers call bullshit." *The Conversation*. February 28. https://theconversation.com/chatgpt-killed-the-student-essay-philosophers-call-bullshit-200195
- Varshney, Kush R. 2022. *Trustworthy Machine Learning*. Chappaqua, New York: Independently Published.
- von Eschenbach, Warren J. 2021. "Transparency and the Black Box Problem: Why We Do Not Trust AI." Philosophy & Technology. 34: 1607–1622. DOI: 10.1007/s13347-021-00477-0
- Yiu, Eunice, Eliza Kosoy, and Alison Gopnik. 2023. "Transmission Versus Truth, Imitation Versus Innovation: What Children Can Do That Large Language and Language-and-Vision Models Cannot (Yet)." *Perspectives on Psychological Science.* DOI: 10.1177/17456916231201401