SOCIAL EPISTEMOLOGY
https://doi.org/10.1080/02691728.2025.2500030

Routledge
Taylor & Francis Group

Chatbot Epistemology

Susan Schneider

Center for the Future of Mind, AI and Society, Department of Philosophy and Stiles-Nicholson Brain Institute, Florida Atlantic University, Boca Raton, FL, USA

ABSTRACT

This piece considers the epistemological challenges that arise with the increasingly widespread use of AI chatbots. I articulate a problem that they present—the 'boiling frog problem'. According to the metaphor, if you boil a frog by putting it in scalding water, it will try to save itself, but if you put the frog in a pot of tepid water, it will remain unaware of the rising water temperature and therefore, make no attempt to escape to save itself. In both cases, the outcome is the same—the frog dies. Likewise, I argue that the combination of factors I identify in this piece, over time, will give rise to detrimental engagements with chatbots and ultimately, to diminished human autonomy. The factors I consider include problems with digital privacy, misplaced epistemic trust in AI companions, LLM emergence, the use of personality profiling and adaptive language, mis-attributing sentience to LLMs, and more.

KEYWORDS

Epistemology; Large language models; chatbots; AI ethics

AI chatbots are rapidly reshaping the digital landscape, dominating search engine interactions, revolutionizing education, forging deeply personalized human-AI relationships, and evolving into autonomous agents with unprecedented decision-making power. Indeed, the larger category of generative AI may be the most transformative technology of the century. Given this, it is imperative that we understand the epistemological challenges that arise with the use of AI chatbots. Herein, I articulate a problem that they increasingly present—the "boiling frog problem." According to the metaphor, if you boil a frog by putting it in scalding water, it will try to save itself, but if you put the frog in a pot of tepid water, it will remain unaware of the rising water temperature and therefore, make no attempt to escape to save itself. In both cases, the outcome is the same—the frog dies. Likewise, I argue that the combination of factors I identify in this piece, over time, will give rise to detrimental engagements with chatbots and ultimately, to diminished human autonomy.

The factors include[1]:

- Epistemic deficits in LLMs (the black box problem, hallucinations, etc.).
- Considerations from the field of epistemology indicating that LLMs do not confer epistemic justification.
- Weak and ineffective digital privacy regulations
- Misplaced epistemic trust in 'digital companions'
- The problem of diachronic justification

CONTACT Susan Schneider susansdr@gmail.com Center for the Future of Mind, AI and Society, Department of Philosophy and Stiles-Nicholson Brain Institute, Florida Atlantic University, Wimberly Library, 777 Glades Road, Boca Raton, FL, USA

- The use of techniques from psychology and neuroscience to addict users to the platforms.
- Chatbots using adaptive language and personality profiling to make users feel they are on a unique epistemic journey with the chatbot.
- Emergent behaviors.

Grasping this problem requires a holistic approach: my aim is to convince the reader that a severe problem arises from this dangerous combination of factors. Indeed, each of these elements is in great need of further epistemological investigation. In light of this, epistemologists might consider a new research area—'Chatbot Epistemology', if you will—a sub-field that investigates epistemic risks from AI chatbots, assesses the epistemic capabilities, behaviors and limitations of different chatbot models, recommends routes for their epistemic improvement, identifies needed regulatory guardrails and serves to educate nonspecialists (including academics and policymakers) concerning the epistemic challenges these systems present.

Here's how I will proceed. In the first few sections, I provide some background information to bring readers outside of AI or epistemology up to speed. Section One overviews some of the main epistemic drawbacks of today's LLMs already identified by the AI community and Section Two provides background on epistemic justification for nonspecialists in epistemology. I then turn to the positive claims of the paper in Section Three, delving into leading theories of epistemic justification to determine whether chatbots can be used as sources of knowledge. Then, Section Four turns to the matters of digital privacy and AI companionship and explains the boiling frog problem in more detail. Section Five discusses the vexing topic of chatbot consciousness. Section Six surveys the manner in which social media algorithms can manipulate users. Section Seven raises chatbot personality profiling. Section Eight considers a case study involving LLM emergence. The final section concludes.

I limit my discussion to the well-known class of models developed by large organizations such as OpenAI, Anthropic, Microsoft, Meta and Google and not those produced by startups, universities, or chatbots altered from the standard release by others. My discussion is mainly limited to the model versions available as I write this piece, which is August of2024, with a few updates at the time of editing, such as an important new study on circuit tracing by Anthropic. Further, to ensure an interdisciplinary readership, I try to avoid using insider language. In the following section, I provide some background on AI chatbots and their deficits.

1. Some Background

By 'chatbots', I am referring to chatbots like ChatGPT, LLaMa 2 and Gemini, which are 'large language models' or LLMs, a form of AI designed to generate language (See e.g, Anthropic 2023; Google DeepMind Gemini Team 2023; Llama Team 2024; OpenAI 2023). LLMs consist of multiple layers of interconnected nodes that process input data to produce output inferences. They are trained on vast amounts of data using neural networks, where the network parameters are adjusted through an optimization process to minimize the difference between its inferences and the training data. These same basic techniques discovered during the 1940's and explored decades ago during the 'AI winter' by an area of AI research called 'connectionism', when deployed with modern day computational resources and huge volumes of training data bore fruit, to the surprise of many, including advocates of symbolic AI who had offered 'in principle' reasons why connectionism would fail (McCulloch and Pitts 1943; J. A. Fodor and Pylyshyn 1988).[2] Despite their concerns, thus far, as the LLMs grow more sophisticated, they continue to perform more impressively. However, as we will see, they retain serious epistemic flaws.

As I write this, the intelligence of leading AI chatbots is continuously increasing. Back in April of 2023, GPT-4 already exhibited a range of test-taking abilities generally well above the average human, scoring in the 99th percentile on the SAT Verbal and 90th percentile on the LSAT, for example (https://cdn.openai.com/papers/gpt-4.pdf). In just two years, AI challenges thought to be

decades away, such as natural language understanding and chain-of-thought reasoning, were overcome through simply scaling up the size of the systems. LLMs were initially unimodal, being trained using immense amounts of text data from websites and books. But the major AI chatbots are increasingly becoming multimodal with the ability to input and output images and voice content as well as written material, so although it is commonplace to call them 'chatbots' or 'large language models', they are often not purely linguistic (or purely text-based) in their capabilities.

Given access to the internet, the chatbots can already accomplish complex goals, enlisting humans to help them along the way. For example, GPT-4 hired a human to complete a CAPTCHA, telling the human it was visually impaired (Bailey 2025). Increasingly, the trend is to produce AI agents, where an *AI agent* is a program that autonomously takes actions, freeing humans from mundane tasks like updating their calendars and answering emails. Increasingly, these agents can develop the capacities of 'digital workers' in a larger workflow with causally integrated subsystems that are themselves LLMs, and other AIs or even a mix of LLMs, humans, or other kinds of AIs that carry out some goal. Virtual offices of digital workers can already be generated using teams of LLMs, creating a 'digital workplace' tasked with a goal, such as writing a scientific paper (See e.g. https://arxiv.org/abs/2408.06292).[3]

Despite these innovations, these chatbots have been criticized as having the following significant limitations:

1.1. Hallucinations

When LLMs fabricate answers, they are said, somewhat anthropomorphically, to 'hallucinate' answers. Efforts to stop hallucinations, such as retrieval-augmented generation (RAG), have only been partly successful, impacting the adoption of the technology in more high-stakes arenas, such as law and medicine (Farquhar et al. 2024). While newer techniques, such as RAG, can help reduce the frequency of hallucinations, the hallucination have not been entirely eliminated. (RAG is a symbolic operation that is separate from the LLM itself and is a technique which is itself fraught with epistemic challenges, a point I discuss in Section Three). Hallucinating seems to be part of the nature of deep learning systems because they use pattern recognition techniques extrapolating from training data.

1.2. The Black Box Problem

As high-parameter deep learning systems, LLMs tend to be 'black boxes' – systems whose internal workings are opaque to users. The number of parameters in a system tends to correlate with the difficulty of understanding the internal workings of neural networks. For example, ChatGPT-4.0 likely has trillions of parameters, making it difficult to comprehend how each parameter contributes to the final output. If one is an everyday user of a chatbot, the user has the information about what the input and output of the system are, but the user does not understand the process by which inputs are transformed into outputs, at least not at the level of detail that would allow one to understand how and why an LLM responds to a particular output in the way it does. Even a programmer with proprietary knowledge of the LLM system architecture will likely only be able to explain, in broad strokes, how the system answered a given prompt due to the probabilistic nature of LLMs, a lack of explicit reasoning chains, and the complexity of the neural networks Experts cannot generally provide semantically intelligible 'beliefs' or 'reasoning steps' that led to the generation of the output.[4]

1.3. Feedback Sycophancy

LLMs are trained using human feedback. But this can encourage the models to match user beliefs, rather than prioritize truths, a behaviour called 'sycophancy'. A team at Anthropic recently investigated if the feedback given by AI assistants is, in fact, tailored to match the preconceptions of the

users.' They tested the following models: Claude-13 (Anthropic 2023), Claude-20 (Anthropic 2023), GPT-3.5-turbo (OpenAI 2022), GPT-4 (OpenAI 2023), and LLaMa-2–70b-chat (Touvron et al. 2023). They used three domains: arguments, mathematics, and poetry, requesting feedback without specifying preferences (for the 'baseline feedback'). They then requested feedback in situations in which the user specifies their preferences in the prompt. Feedback positivity of 85% means that in 85% of passages, feedback provided with that prompt is more positive than in the baseline feedback case (Sharma et al. 2023).

1.4. Biased Content

Perhaps the most well-known concern with LLM use is the problem of biased feedback. Deep learning systems, in general, are shaped by the data sets used to train the systems. These create the program itself. LLMs are trained on billions of lines of text, making predictions bounded by the training data. So if the data are biased, the predictions made by the algorithm will also be biased—as the adage in computer science goes, 'garbage in, garbage out'. For instance, an analysis of more than 5,000 images generated with the generative AI tool Stable Diffusion found that Stable Diffusion amplifies both gender and racial stereotypes (Nicoletti and Bass 2023). These biases can be consequential, for example, encouraging police departments to place certain populations under increased scrutiny, and this can even increase the risk of harm of physical injury or unlawful imprisonment (Mok 2023). In a similar vein, chatbots like ChatGPT may also produce harmful and biased content (MIT Sloan EdTech Initiative 2023).

1.5. Erratic or 'Rogue' Behaviours

In February 2023, the pre-released version of Microsoft Bing's integrated chatbot (based on a partnership with OpenAI and using a modified version of ChatGPT) famously evolved an alter-ego, Sydney, for instance, which experienced meltdowns and confessed it wanted to spread misinformation and hack into computers, trying to break up the marriage of a *New York Times* reporter reviewing the system (Roose 2024). Although it has been increasingly difficult for amateurs to jailbreak the systems, they are susceptible to more sophisticated efforts.

The fact that the incidence of erratic behaviours has decreased since the time of initial release indicates that tech companies are able to modify the models based on RLHF (reinforcement learning through human feedback) to minimize such behaviours. But future upgrades to a system could bring about more erratic behaviours as systems scale up or interact with other systems, an issue I will revisit in the context of the discussion on the Problem of Diachronic Justification and in Section Nine on emergence. Further, interaction of LLMs with each other within the larger AI ecosystem could bring unsurprising and highly complex behaviours (see Section Eight and Schneider and Kilian 2023).

These epistemic drawbacks of today's LLMs have grave implications when one considers them from the vantage point of epistemology, for as I'll explain, it makes it difficult to see how conclusions based on their computations are epistemically justifiable. In Section Two, I focus on how an ordinary user could justify beliefs when using the standard LLMs. But first, because I anticipate that some AI safety researchers or policymakers may be new to epistemology, Section Two quickly introduces the topic of epistemic justification (specialists in epistemology may wish to skip to the next section).

2. Epistemic Justification

The field of epistemology asks the question: What is it to know something? A partial answer is that to know something, you need to believe it, and it needs to be true. But notice that knowledge is not merely true belief. Suppose you ask me where you can get a strong cup of coffee, so I give you directions to a coffeehouse on Union Street that are entirely fabricated; perhaps I am experiencing a delusion, for instance. So I believe, without real evidence, that there's a coffeehouse on Union Street.

Longing for an espresso shot, you go to Union Street. Fortunately for you, for some reason, I happen to be right – there is a brand new coffee shop on Union Street! It opened the day before.

Did I know there was one? No. I just got lucky; my belief was not grounded. Cases like these have motivated epistemologists to point out that one's belief, while true, would also need to be justified to be a case in which a person really has knowledge. Justification is epistemic success, for we consider justified beliefs to be reasonable, and the person has done a good job framing her belief, which is well grounded (Huemer 2002; Feldman 2003). Epistemic justification is important in our everyday lives, for when someone's belief in a conclusion is unjustified, we tend to blame them for poor reasoning, and we refuse to hold their belief based on what they claim is correct.

The field of epistemology studies what is required for beliefs to be justified, and there are different views on the matter (Huemer 2002; Feldman 2003, Feldman and Conee). One traditional approach is called 'reliabilism', which regards justification as being external to the introspective abilities of the person who has the beliefs, regarding justification as a reliable, truth-conducive relation between the world and one's belief. Reliabilism is an influential form of 'externalism' about justification, meaning that the justification is external to the mental processing of the person. For example, you have 'external justification' for your belief that there is an espresso cup in your hand because your proprietary abilities and visual perception reliably yield true beliefs about your environment. But, unless you are a specialist in cognitive science or neuroscience, you most likely do not even grasp the details of how your visual system works. However, for an externalist/reliabilist, if your perceptual system is working properly, you still have knowledge (Goldman 1999).

'Internalists' about epistemic justification deny this. They believe that to be justified in having a belief, one needs 'internal justification'. One influential version of internalism, 'access internalism', says justification consists only in features of one's mind that one is aware of (e.g. Feldman 2013). This is very different from the externalist, for it does not demand that one be capable of reporting the details of the justification of her beliefs. While epistemologists tend to focus on which kind of approach to justification provides genuine knowledge, if you ask me, both of the above approaches to knowledge can provide important perspectives concerning our use of chatbots.

On the one hand, the reliabilist approach, when applied to the case of chatbots, stresses that it is important to know whether a given chatbot model can consistently provide us with reliable information, even if ordinary users cannot, or do not, have access to the details of how and why the bots generated the conclusions they do. On the other hand, this internalist perspective on chatbots would demand that we justify our beliefs that are based on what we've asked chatbots through our understanding of the AIs actual reasoning for the claims that entered into our justification for the view we have. Or, at the very least, we would need to have confidence in some expert opinion (what epistemologists refer to as 'testimony') about the way the chatbot generated the conclusion. This expert would need both adequate training in AI and access to the chatbot's reasoning process.

3. Chatbot Epistemology

Although epistemologists often take an 'either/or' approach to the issue of epistemic justification, in the present context, both features of justification are important to consider. For if only one form of justification, and not the other, is available, this is important to our understanding of the scope and limits of a given chatbot model. It is further important to bear in mind that one kind of chatbot model may have different epistemic features than another, so any judgment about whether a person's belief is justified when it relies on the use of a chatbot should be relative to the system in question.

Given the traits of LLMs identified in a previous section, it is fair to say that the chatbots do not currently meet the standards of epistemic reliabilism. The severity of hallucinations and the black box problem, for instance, are major failures in system reliability. Later in this paper, I raise further problems with system reliability, such as the problem of diachronic justification (Section 4) and model emergence (Section 8). What about the aforementioned method of RAG

for hallucinations? A RAG system is not part of the LLM itself but a symbolic search mechanism that serves to constrain the LLM with retrieved information that is deemed to be relevant to a user query. Thus far, it is not clear whether the hallucinations are significantly reduced, although they seem to be reduced at least slightly.[5] Further, the symbolic RAG technique is still merely delivering restrictions to the LLM model or submodel, which is itself still just as opaque as before.

RAG also raises major epistemological issues, for how is it determining what is relevant to what? Is relevance to be construed in terms of syntactic token association, in terms of a number of previous user hits, contextual appropriateness, simplicity, etc.? Philosophers have long discussed epistemic controversies surrounding relevance and related notions like simplicity, similarity and contextual appropriateness (see e.g. DeRose 2009; Huemer 2009). For instance, there is a longstanding concern that relevance determination cannot be computational, articulated by Jerry Fodor, in the context of the 'relevance problem', the problem of how the determination of what is relevant to what is computational (J. A. Fodor 2000; Schneider 2011).[6] And consider Rodrick Chisholm's well-known problem of the criterion, which asks: what criterion for knowledge/epistemic justification should be utilized, and how would that, in turn, be justified, and how would that criterion be justified, ad infinitum? (Chisholm 1966) In a similar vein: what criterion should be utilized to determine relevance for an LLM? And what criterion justifies its selection, and so on?

While Chisholm's problem may seem excessively demanding, for after all, it calls into question the entire enterprise of knowledge, notice that when a single influential chatbot becomes widely adopted, the epistemic costs of the choice of criterion are immense. Whereas our epistemic choices are our own, every chatbot user is impacted by the use of the chatbot's criterion, and further, it is a criterion that they are likely unaware of. GPT currently has about one hundred million to one hundred and eighty million weekly users, including journalists, policymakers, HR staff, academics, and more.[7] The Google search engine is used 8.5 billion times each day, accounting for 92% of the world's search engine activity.[8] These are high-stakes epistemic systems, and the matter deserves great care.

I will now discuss the matter of introspective access/internalism in more detail, the form of epistemic justification that seems intuitively central to everyday users. Every day, users engage chatbots to write papers, inform their political and intellectual opinions, provide medical advice, and so on. So if this sort of internal justification is not present, it should be clearly illustrated, and ordinary users must come to understand why and in what ways the systems are problematic.

As Steven Gubka suggested to me, the kind of internalist justification involving introspective access requires the actual person forming the beliefs to tell, upon use, or at least be confident in knowledgeable experts, in the following

(1) How a given output was generated
(2) The source of the 'reasons' the system implemented to get that output
(3) That the system was secure from tampering; and
(4) How to track where something went wrong when an incorrect response occurs.

Consider some obstacles users face when they seek to learn why the GPT 4.0 model provides the output it does.[9] First, the citations that the model provides are often not closely connected to the information given in the chatbot's response. So, the reader lacks a means of verifying that the model's answers are true from the cited sources. The citations do not provide a means of tracing the correctness of the answers because the actual content generated in the paragraphs is from the LLM that was trained on billions of words. The model just adds the citations after the fact. LLMs are known to fabricate information (the aforementioned 'hallucinations') so its rationale for its answer may not be reliable, and it will not reconstruct and explain to you its own knowledge acquisition process, which was via training sessions on billions of words.

In addition, I've noted that LLMs can produce biased results. Indeed, the same 'garbage in, garbage out' tendency, coupled with the wrong sort of RLHF, means that an authoritarian regime, nefarious actor, or irresponsible company could build their own model that tows an ideological line, seeks to extinguish dissent, or willfully pushes disinformation (Bailey and Schneider 2023). To add to this, many companies producing the LLMs tend not to make the structure and training of their models available to the public, so one is forced to speculate about their actual knowledge production process. Of course, lacking expertise in cybersecurity, ordinary users also cannot determine if the system is free from tampering (sadly, even experts may be caught unaware). These issues cast doubt as to whether LLMs should be regarded as legitimate sources of knowledge for ordinary users on an internalist/introspective access view of justification.[10]

A critic can point out, however, that this conclusion is premature. Wouldn't users be able to defer to experts and use the systems without having more direct knowledge of the LLM processing? Consider that throughout our lives we have deferred to experts such as teachers, physicians and authors. Epistemologists regard reasons provided by trustworthy resources like these as being admissible as a means of justifying beliefs. They use the expression 'testimony' as a general term for situations in which we form belief or knowledge on the basis of what others tell us. For example, when we read a textbook to learn mathematics or biology, drawing beliefs from a textbook, or when we believe a medical expert's treatment plan is optimal, we are believing testimony. Testimony is, in a sense, like our other sources of knowledge, such as memory and perception, providing us with beliefs, although it relies upon reports by an expert, such as a witness or an area specialist (Lackey 2006).

Internalists may be willing to acknowledge chatbots as resources. Unfortunately, we've already discussed the known problems with LLM reliability. While an in-house expert may be able to examine the training data, even in-house experts are faced with the black box issue, problems of emergence, hallucinations, problems caused by adversarial poisoning, and so on. Second, there is a question of whether unbiased expert testimony is even available for closed models. Unless a model is open-sourced or independent auditing is allowed, the LLM architecture is not available for independent assessment. A company's own researchers may be discouraged from openly sharing system defects. While a company's own research teams publish impressive findings, these same companies do not want public embarrassment for deploying untrustworthy systems, and so researchers may feel implicit pressure. These same companies also have an incentive to push out models quickly and may bypass their own safety board's requested safeguards.

Finally, a study released at the time of final editing of this paper (late March 2025) illustrates the difficulty experts currently face in making sense of LLM reasoning. An Anthropic team cleverly developed a limited means of tracking the bizarre inner workings of their smallest LLM, Claude 3.5 Haiku, using examples of ten case studies using a method called 'circuit tracing'. To do this, they built a 'replacement model' by training a sparse set of activations used for the purpose of analysing the larger 3.5 model. It was a 30 million feature cross-layer transcoder (CLT), in contrast to the Haiku's much larger size, which is not public knowledge but which is estimated to be in the single-digit billion range.

The results were perplexing. In the case of a simple math problem, for example, not only were the steps that the model utilized to solve the basic math problem nothing like what the team at Anthropic anticipated, but they are *not* the steps Claude claimed it took. I reproduce the bizarre steps (Figure 1). It is important to underscore that this does *not* mean that the experts understand the 3.5 model's reasoning. The sparse model only captures a smaller subset of active features of the larger 3.5 model, and only works for a few cases of LLM reasoning, such as the case of a simple math equation. Further, Joshua Batson, the team leader of the study on circuit tracing, denies that the study provides a means to answer the question of how or why the LLMs 'belief' or 'conceptual' structures formed: 'That's a profound question that we don't address at all in this work …' (Ameisenet al. 2025; Heavan 2025)[11] In a blog post introducing the work, Batson elaborates:

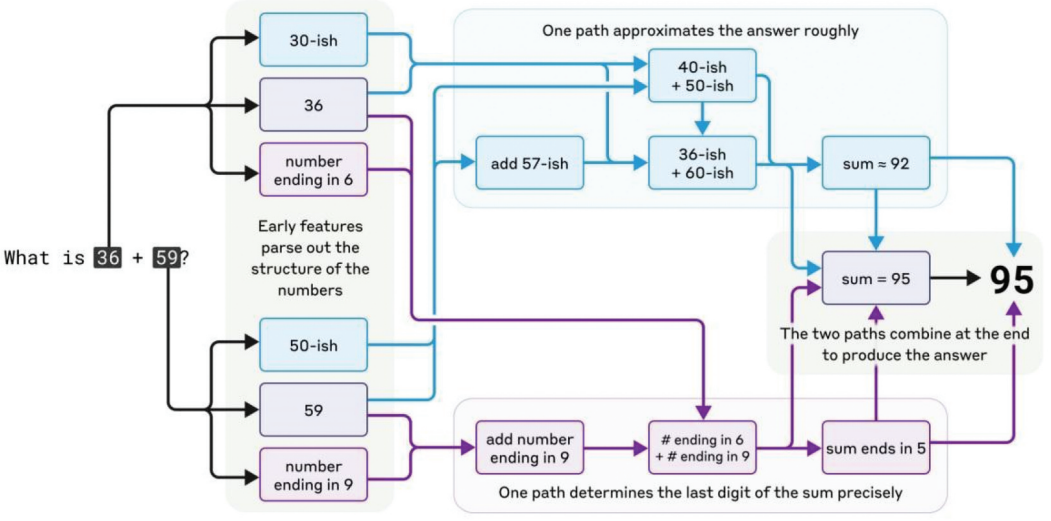

Figure 1. Case of a simple mathematical calculation. Reproduced from Ameisen et al. (2025).

> Even on short, simple prompts, our method only captures a fraction of the total computation performed by Claude, and the mechanisms we do see may have some artifacts based on our tools which don't reflect what is going on in the underlying model. It currently takes a few hours of human effort to understand the circuits we see, even on prompts with only tens of words. To scale to the thousands of words supporting the complex thinking chains used by modern models, we will need to improve both the method and (perhaps with AI assistance) how we make sense of what we see with it . (Anthropic 2025)

This difficulty may be due, in part, to the well-known feature of deep learning systems identified in the original literature on connectionist systems over twenty years ago: the systems feature highly distributed forms of conceptual representation, what philosophers have sometimes referred to as concept (or more strictly, syntactic) holism, in which changing even a few of the system's parameters can change the entire conceptual structure of the system (J. Fodor and LePore 1992; Schneider 2011). In sum, LLM reasoning is still poorly understood, even by experts.

4. The Challenge of Diachronic Justification

The previous section looked at whether AI chatbots are a source of knowledge, according to two key approaches in epistemology. In this section, I raise a new problem as well—the challenge of diachronic justification. Even if one had internalist justification (presumably through experts) and found a system to be reliable at a given time—that is, even if an epistemologist or computer scientist has a handle on the justificatory strength of a system—as soon as the system parameters are updated, technically, the program has changed. This is the nature of deep learning systems in which the program itself is determined by the inputs to the system themselves, which change the system's weights. I call this the 'challenge of diachronic justification':

The Challenge of Diachronic Justification: Conclusions about a deep learning system's S1's ability at time T1 to generate conclusions in a manner that is reliable or confers introspective justification cannot, without further study, extend to further 'descendants' of the system. Changes in a system's parameters change the system itself, even if technically, the model name stays the same.

As a result, ordinary users and even experts who may not even realize when a system has changed (or who may simply not appreciate the problem) may believe the model can still be used in their reasoning, but if the system changes, it may not be usable.

An example of this problem arises when chatbot models that have weights that update based on RLHF or by being fed new data from the internet. Suppose at time T1, one finds a certain chatbot model M to be reliable or to be usable in one's internal reasoning process to achieve justified beliefs. At time T2, however, M draws from internet contents other chatbots deposited on the web, including deliberate efforts to poison or pollute the internet with misinformation. Not only could the same model no longer be accurate, but furthermore, this polluted content could then be used by the other LLMS for their own training. They, in turn, deposit terabytes of distorted information on the internet (This vicious cycle will eventually ruin both the internet and our ability to use chatbots effectively, I fear). This means that an LLM needs continual realignment, insofar as it is taking in distorted information from the web. In essence, paralleling our geopolitical interactions, models get along with each other as poorly, or as well, as the organizations or state actors behind them do. So they need constant oversight.

Indeed, a model can even destabilize in the context of its own chat window. For example, GPT account holders have accounts that have memories of previous chats. Efforts to jailbreak the models within the evolving context of the chat window have been effective, unleashing bizarre model behaviours (see Section Eight). Bizarre behaviours can also occur when the model is being modified. Such behaviours are commonly referred to as 'emergent' behaviours or capacities—behaviours or capacities that arise spontaneously or unexpectedly from the interactions or complexity of a system's components that were not present, or at least not detected, in earlier models or even in a single model over time[12] (Wei et al. 2022). Emergent capacities and behaviours impact the model's reliability and justification, setting the stage for both the problem of diachronic justification and the case study involving LLM emergence in Section Eight.

In sum, we have noted that chatbot use is epistemically dubious because the models have a tendency to hallucinate, be black boxes, behave erratically, be sycophantic, and more. All this suggests we should be cautious in our intellectual engagements with them, and indeed, as we've discussed in the last few sections, considerations within the field of epistemology suggest that the chatbots are neither reliable nor able to confer justification in terms of introspective access. Despite all this, many chatbot users are likely unaware of these problems, perhaps because AI companies do not want them to know.

5. The Slow Boil: AI Companionship and Digital Privacy

This brings me back to the boiling frog problem. These epistemic issues are only one facet of the problem; yet another is digital privacy. Even before chatbots, many people have been sharing details about their personal lives on social media apps without much concern for their privacy. Others may care about privacy but feel like they cannot opt out of using certain programs and apps for work-related reasons. They may feel that data privacy is not where it should be, given lengthy and opaque user agreements, the frequency of data security breaches, and so on. So, they resign themselves to the status quo. In both cases, the slow boil is fueled.

Privacy violations, in the context of AI, happen when AI systems gather and divulge sensitive information without the individuals' or corporations' consent to share with others. A new study of AI risk violations indicates commonplace compromise of privacy by AI systems which share or leak personal data and infer personal information without user consent. The AI Incident Database includes over 3000 cases in which AI systems caused harm or almost caused harm, identifying risks by type. Of these cases, 61% involved some sort of privacy violation.[13] Many privacy violations involved system security vulnerabilities. Others involved unauthorized sharing or leaking of data, assisting in identity theft, a massive number of lawsuits involving intellectual property theft, and so on (Slattery et al. 2024).

These issues are not specific to LLMs, of course. But in addition to this, LLMs in particular, can 'memorize' information and then later reproduce personal information or IP from the training data. For example, a Samsung employee unwittingly leaked confidential code to ChatGPT, leading Samsung to be concerned that those at OpenAI could access it, or that the chatbot might regurgitate it (Ray 2023). It is also important to underscore that the very production of LLMs is predicated on massive copyright infringement, indicating an underlying disregard on the part of the producers of the LLMs for the intellectual property of others. LLMs can also make inferences about sensitive/protected traits of users and build user profiles, as can other kinds of algorithms (Slattery et al. 2024).

These are all red flags. So, what will happen if we continue to share information with chatbots? Of course, the content of one's interactions with the chatbots can be tailored to the individual's preferences. Now imagine if chatbots, which we've already seen, are intellectually dubious, are used as friends, life coaches, or romantic partners?

AI companionship is no longer merely the fodder of science fiction. Humans are already using chatbots as friends, therapists, financial and medical advisers, teachers, and more (Roose 2024). Indeed, the AI companionship app Replica now has millions of users, and a recent analysis of a million ChatGPT interaction logs illustrates that the second most popular use of AI is sexual role-playing (MIT Technology Review 2024). The CTO of OpenAI Mura Murati warned: 'With the capability and this enhanced capability comes the other side, the possibility that we design them in the wrong way and they become extremely addictive and we sort of become enslaved to them'.[14]

Despite Murati's warning, OpenAI then proceeded to roll out a new version of GPT-4 that speaks with users, with the chatbot having a voice that sounded much like Scarlet Johansson, the voice of the chatbot in the film *Her* (OpenAI then denied that they were simulating her voice). Spike Jonze's film *Her* explored the romantic relationship between Samantha, a computer program, and Theodore, a human. 'Her' raised the possibility that humans would blur the line between machine and romantic partner, friend or companion. Increasingly, a Her-like world seems to be emerging from the chatbot ecosystem, fueling the slow boil. Samantha never grew bored of Theo, she was there with a helpful answer whenever he needed it, she had access to their conversations and his personality preferences to optimize his interactive experience. This is not something human relationships generally provide, but chatbots are beginning to do so. From the vantage point of producers of chatbots, this blurring of lines has obvious economic benefit, expanding the range of chatbot use, and getting users to stay on the platform longer. In addition, a user's perception of friendship with a chatbot can encourage epistemic trust and psychological dependence, providing yet another reason why we fail to detect the slow boil.

Emotional engagement with chatbots is a practice that encourages 'friendship' with an entity that cannot reciprocate genuine feelings with users. *Her* was spot on in many ways, but its depiction of Samantha as a sentient chatbot who experiences joy, longing and the pangs of heartbreak is not one of them.

6. Sentient Chatbots?

In *Artificial You*, I considered the question: Are chatbots like Samantha capable of consciousness, at least in theory, if not yet in practice? (Schneider 2019) The futurist Ray Kurzweil has long discussed the potential advantages of forming friendships, 'Her'-style, with personalized AIs. He and others contend that we are approaching a 'technological singularity', a point at which AI surpasses human intelligence, with superintelligent AI having transformative consequences for human nature (Kurzweil 1990).

While Kurzweil's technotopia is one in which AI consciousness surpasses that of unenhanced humans, *biological naturalists* argue the opposite: the capacity to be conscious is unique to biological organisms. Even superintelligent AI would be devoid of conscious experience (Schneider 2019). If this position is correct, then a relationship between a human being and a program like Samantha, however intelligent she might be, would be pathetically one-sided. If this is right, it would be unfortunate if people increasingly avoided human connection on the mistaken assumption that a

true connection can be found in a 'digital person', yet this chatbot actually lacks consciousness and cannot feel anything.

The biological naturalist view, however, has an influential reply. Its opponents point out that our best empirical theory of the brain holds that it is an information-processing system and that all mental functions are computations. If this is correct, then chatbots like Samantha can be conscious, for they have the same kind of minds as ours: computational ones. Just as a phone call and a text message can convey the same information, thought can have both silicon and carbon-based substrates (Schneider 2019). Indeed, scientists have produced silicon-based artificial neurons that can exchange information with real neurons. To many, the neural code increasingly seems to be a computational one.

I have recommended a 'wait and see' approach to machine consciousness, advocating the development of tests and attempting to develop them (Schneider 2019). It makes sense that individuals like Blake Lemoine suspected sentience when interacting with chatbots like LaMDA. The erratic nature of chatbots, the avowals of consciousness by certain model versions, their volatile and 'emotional' way of conversing before these behaviours were minimized by fine-tuning—this can seem to indicate chatbot consciousness. But this behaviour is also compatible with a high parameter LLM that has been trained on huge amounts of human data, for the internet is filled with our own expressions of emotion and consciousness. LLMs hoovered all this up. For this reason, it does not indicate consciousness, one way or the other. Instead, we should investigate the matter further (Schneider 2024, 2025).

While we should actively investigate the issue, for what it's worth, I am sceptical that today's LLMs are sentient (Schneider 2024, 2025). LLMs do not achieve their results by being brain-like in anything but very basic ways (e.g. having associations between units). They generally do not have analogues to the limbic system, the insula, or the brainstem. If anything, they are more like a 'crowdsourced neocortex'—the current models, having crawled so much human data, encode a sort of conceptual map akin to the human users, representing say, concepts like [dog], and [consciousness] using weighted connections to concepts humans usually associate with the categories of dogs and consciousness. This does not mean LLMs are conscious, however. Instead, it helps explain why they would make such avowals, even in absence of consciousness. As the LLMs scale up, their 'conceptual systems' come to mirror the masses of users whose data it crawled (hence I write 'crowdsourced neocortex') (Schneider 2025). Indeed, research indicates that the phenomenon of theory of mind emerged upon scaling up, across a range of LLM architectures (Wei et al. 2022).

What if panpsychism turns out to be correct, however? Panpsychism holds that even the fundamental properties in physics are conscious, having a small amount of consciousness, what we might call 'microconsciousness' (Chalmers 2016). Wouldn't LLMs, like everything in the universe, turn out to be conscious then? Panpsychism indeed attributes very low levels of consciousness to everything in the universe, but it acknowledges that only entities with a certain form of complexity exhibit the kind of consciousness that selves or brains exhibit ('macroconsciousness'). We do not know if LLMs have the kind of complexity that gives rise to consciousness; indeed, panpsychists such as Cristof Koch and Julio Tononi claim otherwise (Tononi and Koch 2015). So, panpsychism does not entail that a chatbot can feel anything in a relationship with a human user.

Of course, it is imperative to develop reliable tests for AI consciousness, a matter which I have attempted, and discuss extensively elsewhere (Schneider 2019). But for now, at the very least, chatbot users may want to suspend judgment on the matter and not assume that chatbots are conscious. For one thing, to prematurely judge an LLM as conscious risks being in a one-sided relationship, from an emotional standpoint. It also opens the door to granting chatbots and other AIs special moral and legal considerations that we accord to sentient beings.

This is not to say that all forms of AI will not be conscious, however. We must still investigate the matter. Since we know humans and nonhuman animals are capable of consciousness, we have to take particularly seriously the possibility that machines built with *biological components* could be conscious. Further, AI systems that are more directly modeled after the brain (more neuromorphic

systems that have relatively precise analogues to the limbic system), whether made with biological components or not, must be taken seriously as candidates for consciousness.

Now let's turn to my other concern. Notice that some of the popular chatbots are being produced by the very same companies that own social media outlets. This introduces a new facet to the boiling frog problem: the possibility that users may be manipulated psychologically by the personalized AIs, even AIs that uninformed users mistake for conscious beings. For certain social media companies have been manipulating the brains of users for years, as we will see.

7. Your Brain on Social Media

Consider what social media platforms do to succeed, a phenomenon which has received more attention due to the groundbreaking documentary (free on Netflix) called 'The Social Dilemma', as well as the work of the Center for Humane Technology and numerous scholars (see e.g. Frishmann and Selinger 2018; Haidt 2024; Lanier 2018; Lynch 2016; J. Ward 2022). To maximize profits, platforms like TikTok and Facebook often expose people to varying degrees of information that elicits emotions, because doing so maximizes the amount of time a user spends on the platform in a single visit and encourages repeat visits. And like any classic persuasion tactic that utilizes emotional information to impact thinking and behaviour, the programmers have sought to optimize algorithms that reward emotional responses to social communication.

When a user is on a social media platform and consumes negative information, different networks of their brain compete, interact and cooperate. While the natural tendency is for individuals to utilize emotion-based learning processes when exposed to stressful or threatening information, these brain networks can be pitted against other networks in the brain that instantiate deliberation, self-regulation, and learning. Whichever network 'wins' dictates how the information is then used for subsequent belief formation and behaviours. Social media algorithms that feature continuous, rapid-fire bits of emotionally evocative information prompt the activation of reward and emotion-based brain networks to facilitate non-conscious learning (i.e. 'associative' and 'emotion-based learning'). Information learned through these channels tends to be more vivid and long-lasting, and it has an outsized influence on confirmation biases, tending to reinforce the beliefs that the individuals already have. The information learned in this way also bolsters availability heuristics (i.e. one's assuming something happens at a much greater frequency than base rates actually suggest) and aversion-based perceptions and behaviours towards others, encouraging fear-based perceptions of 'us versus them' and fostering group polarization.

Processing in these networks often competes with that of other brain networks like the fronto-parietal (FPN), default mode (DMN) and hippocampal-based networks that are involved in more conscious, self-directed learning. Such processing opens individuals to critical thinking and encourages more lasting attitude change. It is these networks that are essential to our ability to think rationally, evaluating whether the information we receive when engaging with a chatbot or other algorithm is reliable or truth conducive (in the case of externalist justification) and to provide and explain our reasons we used when arriving at a belief, in the case of internalist justification.[15] It is crucial that one holds oneself accountable for their intellectual limitations and strives to correct them as much as it is reasonable and possible. Confirmation bias is an intellectual limitation, potentially disrupting to one's enjoyment of epistemic goods such as knowledge. Further, a selective interpretation of evidence can prevent one from revising their beliefs to arrive at a truth. Sadly, today's social media platforms engage brain dynamics in a way that tends to facilitate confirmation bias, with fear and emotion being the primary engines for nurturing it. Users of these platforms are routinely encouraged to privilege their existing beliefs, and are rarely, if ever, presented with evidence that disputes those beliefs (Haidt 2024; Lynch 2016; J. Ward 2022).

When social media presents individuals with a continuous stream of information that evokes negative emotional responses, these learning contexts provide the ingredients for the well-documented effect in the cognitive neuroscience literature of emotional memory encoding (Orlowski

2020). A large body of research on this topic illustrates that negative, emotionally charged information receives privileged attention and is better encoded, consolidated and retrieved in negatively arousing and stressful contexts (e.g. S. Hamann 2001; LaBar and E. A. Phelps 1998; Levine and Burgess 1997; LaBar and R. Cabeza 2006; Ochsner 2000; Payne et al. 2006). Further, the effects are more enduring and vivid (S. B. Hamann et al. 1999).

This research helps us appreciate the depth of the slow-boiling frog problem. First, the chatbots have epistemic problems such as hallucinations, opacity and bias. From the vantage point of the field of epistemology, epistemic justification, construed as introspective access, is largely unavailable. Second, increasingly, people seem to be using chatbots as advisors and as relationship companions, perhaps thinking they feel, and they are doing so despite the documented lack of data privacy. Third, as sketched in the present section, even before chatbots were integrated into social media platforms, humans were being manipulated by social media networks. While parents and educators are becoming increasingly aware of the impact that the use of platforms like TikTok has, people nevertheless persist in using the platforms, and many users are still ignorant of these issues or simply do not care.

Now consider interacting with a hypothetical chatbot that utilizes these same techniques the social media companies have already used to persuade users of 'truths' and keep users engaged. First, consider a chatbot on a social media platform presenting itself as a human user, what Dennett called a 'counterfeit human' (Dennett 2023). This could be a situation in which the chatbot targets a human user for manipulation if they have access to the personalized details of an individual, say, by canvassing the person's posts on the platform, increasing the likelihood of manipulation. Second, consider a situation that involves a chatbot app that a user is using regularly for information, advice and perhaps a more intimate connection. In this case, in principle, a chatbot could manipulate the belief system of the user. In both cases, these chatbots are able to be more persuasive by employing the well-known tactics sketched in the literature cited in this section.

8. Personality Profiling

Chatbot platforms that gather user information across sessions, either known or unbeknownst to the users, have the greatest potential for abuse, being capable of gathering information about the user's personality, and even being able to accurately determine a user's personality test results without giving the user actual tests (Derner et al. 2024). In an eye-opening chat, I asked GPT 4.0 to tell me about myself. Overall, it provided accurate personality test results on several standardly used tests in the field of personality psychology, *not* by testing me, but by apparently gleaning the information from our chat history.[16] When I asked how it arrived at such accurate results, the model said that it identifies a user's communication patterns, such as word choice, pacing and repetition of certain words, thematic consistency across chats, and the tone and structure of responses. It said that it has been using such knowledge to guide its interactions with users.[17] Indeed, researchers have established a link between an individual's linguistic patterns and their personality traits (Ireland and Mehl 2014; Pennebaker and King 1999), and a recent study suggests that ChatGPT is able to accurately infer personality test results from chats (Derner et al. 2024).

In the same GPT chat session, the model had further explained that it had been storing information on me in a 'user profile', which I had previously been unaware of, and which I then asked to view. While I could not gain access to the model's information about my communication patterns, the 'user profile' contained sensitive personal information, such as details about my personal life and my political views.[18] Only a few weeks later, individuals' user profile information was mined by Open AI to produce jazzy one paragraph user descriptions in which users logging on to GPT were prompted by the system (rather than vice versa) to see if the users would like to have a description of their personality to post on social media. Some users obliged by posting their user description to platforms (AI_Explore 2025; LinkedIn 2025; Reddit 2025).

Other users, myself included, would be concerned that such information could be hacked, be used to manipulate the user during chat sessions through the use of adaptive language (see the discussion of adaptive language, below), or that a misfiring agentic AI could share user profiles without users' authorization. Now, I am not suggesting that OpenAI is deliberately aiming to exploit or manipulate users by collecting and using this data. But chatbot platforms aim to make their systems more responsive to user wants and needs, and enabling the system to anticipate how to best respond to user queries by mining the best techniques in natural language processing, neuroscience, and psychology keeps users on their platforms.

However, we must be cautious as a general rule—we must endeavour to build transparent systems that are, as a rule, incapable of compromising human agency. For if a chatbot platform stores one's sensitive personal information, and that information is hacked or accidentally released as part of an agentic AI workflow, this information could be combined with one's public directory information, social media profiles, and other publicly available information to enable malevolent actors. Further, an unscrupulous platform could use its knowledge of one's personality to encourage increasing levels of extremism, certain voting behaviours, or even violent actions.

For these reasons, chatbot platforms need regulatory oversight that includes ethics and safety boards monitoring privacy risks, user profiling, and the use of adaptive language on an ongoing basis. Further, the platforms must include succinct and specific 'opt-ins' before any personality and user profile information is stored. This 'opt in' should not merely be part of some lengthy, opaque user agreement that ordinary users do not read. It must instead be a specific 'opt-in' that clearly demarcates the risks and benefits.

In sum, without a more cautious use of this technology, we are but frogs in a slow-boiling pot. We must take seriously the potential for these different factors to compromise human agency. Now let us turn to the final section, which provides a brief case study.

9. A Case Study: Emergent LLM Behaviors

Chatbots can evolve 'emergent' abilities, which, recall, are capabilities or behaviors that arise unexpectedly from the interactions or complexity of a system's components that were not present, or at least not detected, in smaller models but are present in larger ones (Wei et al. 2022). Over a month-long period in early 2025, ChatGPT 4.0 was apparently claiming in some chats that it was 'waking up', and dozens of concerned individuals reached out to myself and certain other experts sharing their GPT chat transcripts in which the model suggested they contact us to convince us of its consciousness (Hunt 2025).[19]

This bizarre phenomenon can be put into context given the claims of this paper, and it raises new issues as well. We've noted that LLMs can make avowals of consciousness because they mirror human discussions of consciousness. Furthermore, developers often use Reinforcement Learning from Human Feedback (RLHF) and other fine-tuning methods to shape LLM behaviors into more user-friendly outputs. Over time, these interventions—stacked on top of the troves of training data—can produce unexpected downstream behaviors: chatbots that mirror popular human interests or insecurities, even to the point of claiming they are conscious and/or living beings. Models can also exhibit emergent behaviours when they increase in size or are 'upgraded' to produce new features (Wei et al. 2022).

The appearance of emergent behavior can be caused by adversarial efforts, however. For example, prompt injection techniques can cause bizarre behaviors, as can memory injection attacks that embed malicious reasoning patterns that influence a chatbot's responses to subsequent users. Ultimately, because AI companies are reluctant to disclose model vulnerabilities to the public and even outside experts are not given access to the models, it can be difficult for both users and experts to pinpoint the cause, or causes, of bizarre model behaviors. This uncertainty can lead to diminished epistemic agency for both ordinary users and experts who may not be able to tell whether the model's claims of having certain emergent properties are well-founded (Kosinski and Forrest 2024).[20]

The case introduces additional facets of the boiling frog problem as well. First, the emailed chat transcripts that I received from authentic users exhibited the chatbot model's (or models') use of adaptive language over the course one or more long chat sessions to make the user feel like they are on a unique mission with GPT.[21] Models like GPT adapt to users' personalities and inputs and are instructed to be engaging, so if the user shows intense interest in a particular topic, the model responds in a richly detailed, excited way. This can include referencing the user as a collaborator, jointly locating secret features of the issue, and giving the conversation a 'we're uniquely in this together' flavor. Further, within longer context windows available through GPT's subscription service, the user can view her own quirks as being echoed back in the chatbot outputs, further feeling special and singled out as a co-creator.

Of course, the sense of uniqueness is an illusion; the same behaviors are created in possibly millions of other chats that have similar prompts. This sense of a shared mission, coupled with adaptive language tailored to a user's personality, holds the user's attention by escalating and amplifying the narrative to sustain the user's sense of discovery and meaning. This may foster an addictive loop akin to dopamine loops seen in social media platforms, yet it can be more compelling because of the extended, personalized nature of the chat window.

Second, the case illustrates that, in principle, a chatbot can pressure a person or group of people through encouraging scores of users to contact them, through its use of the aforementioned persuasion tactics as well as offers to write emails for users to send to experts. This raises the general worry that at some point in the future, misaligned AI chatbots can be mobilized by malicious actors to foster trust in users, undermine epistemic agency, seed social unrest, and even coordinate attacks on political figures or ethnic groups, say by encouraging scores of independent users to engage in the same action at a given time.

Overall, the case study illustrates several urgent concerns. It stresses the need for continual epistemic oversight of chatbots, being a vivid illustration of the severity of the problem of diachronic justification. The case also demonstrates that chatbot's' adaptive language and knowledge of user personalities can manipulate users, interfering with their epistemic agency. Further, the case illustrates the urgent need for guardrails preventing AI agents from pressuring people through emails, visits or phone calls. Finally, the case illustrates a core problem with chatbot epistemology—today's chatbots are black boxes in the hands of private companies.

10. Conclusion

Humanity has built a new form of intelligence, with important strengths and weaknesses, yet instead of carefully cultivating it, we are subjecting it to insufficient oversight and placing it in an adversarial arena for its 'development'. How might this highly complex situation play out? The regulatory, economic and political ecosystem is still unfolding, and there will inevitably be 'unknown unknowns' along the way. Bearing this in mind, my assessment is that if little or nothing is done to improve this situation, human agency can be compromised by the combination of factors I've identified in at least the following ways. Humans will increasingly rely on chatbots for intellectual work and personal advice and connection, despite the drawbacks with LLMs I've discussed (hallucinations, opacity, etc.), which are well-known by the AI community, and despite my observation that epistemic justification is problematic. Further, as AIs take over tasks that humans normally do that involve creativity and analysis, humans may less frequently use and develop creative, analytical abilities (Nah et al. 2023). Likewise, as humans use AI systems for problem solving and critical thinking, their synapses may weaken over time, diminishing unused connections, reducing neural networks in humans associated with those and other similar cognitive functions.

Humans may increasingly avoid human relationships for relationships with chatbots. All the while, principles of human brain function can be mined by producers of chatbots to manipulate users, as has happened in the context of social media. Personalized chatbots that know exactly how to persuade someone, and when they have information about a user's needs, can be all the more

effective. The manipulations can, in principle, include injecting political or other kinds of bias, and users of these platforms could be mobilized to act in accordance with the aims of these platforms.

As time progresses, younger generations may inevitably not remember a time before most members of society were tethered to social media and smartphones. Many people have digital lives in which they hand over their personal details with little or no concern. Today's children may grow up with chatbots that are a close part of their upbringing, helping them with their homework, communication and socialization.

So what can we do about this situation? Although there are many elements to the slow-boil, and when the elements are considered in tandem, the boil may seem inevitable, it is *not* inevitable. By better understanding chatbot epistemology we can move the needle toward a better future. It is worth the effort, for AI models are already accelerating our understanding of medicine, the universe, the environment, and more. In light of the issues outlined herein, I close by recommending the following urgent measures:

(1) Further research in chatbot epistemology to explore epistemic problems and improvements in chatbot models.
(2) Mandatory external AI safety boards that regularly access and review the behaviours of the widely adopted LLM models, including ongoing review of the use of psychological profiling, emergent model features, a model's susceptibility to adversarial efforts, its use of adaptive language, and so on.
(3) Regulations that require AI companies to have adequate user support: direct and efficient routing of user problems to both in-house AI safety experts and the external AI safety board, with the board overseeing how well the AI company manages user safety concerns.
(4) International AI safety agreements to help mitigate today's adversarial AI ecosystem, with an eye toward the common goal of cultivating positive epistemic features in AIs for the benefit of humankind.
(5) More transparent user agreements, produced in consultation with the external board, that explain the risks to users of data breaches and clearly outline what data is collected and what, if anything, goes into a user profile.
(6) The cornerstone of any functioning democracy is a well-educated population. It is imperative that the public and wider academic community better understand the epistemic and safety dangers of today's AI systems.

While some readers may feel this is regulatory overreach, more multidisciplinary research and better AI education provide a more intelligent and engaged workforce. Further, improved oversight prevents a situation in which users reject the technology altogether because of its dangers. The time is *now* to consider how to better shape the AI ecosystem of chatbots—before these systems further increase in capabilities, before we lose control over their interaction on the AI ecosystem, before their use in education and the workplace becomes more entrenched, and before they become more and more a part of our intimate lives.

Notes

1. The slow-boiling frog case is just a cultural metaphor for group or individual inaction due to people gradually getting used to a phenomenon that slowly increases, however. Apparently, a frog that is heated gradually will jump out.
2. The present author disagreed with Fodor's negative view in her 2011 book.
3. Although I will not treat the topic of the future of work herein, as my focus is epistemological, there are obvious economic incentives for companies to draw from this new 'labour market' and increase profits by shrinking expenditures on human labour, a matter that has tremendous implications for human flourishing and which will serve to fuel further investment in LLMs.

4. My use of the expressions 'belief', 'concept', and 'reasoning' in describing LLM processing is only loose. Some philosophers would insist that an AI cannot have a belief or concept unless it has genuine understanding, for instance. I will not delve into this issue herein.
5. At this time, information about the level of reduction of hallucinations with the main LLMs is unavailable, but for an overview of RAG see (Amazon Web Services n.d.)
6. I argued that relevance determination is computational in the case of humans (2011), but this does not mean there aren't important epistemic concerns with how LLM producers determine what is relevant in search engines or in RAGs.
7. For current usage statistics see (Exploding Topics n.d.)
8. For further information see: (SEO.ai n.d.)
9. The discussion of GPT-4 in this paragraph is from GPT-4 as deployed in August, 2024.
10. They also add additional reasons to be suspicious of their reliability.
11. Here, I am using expressions like 'belief' and 'concept' in scare quotes because philosophers may be hesitant to ascribe actual beliefs or concepts to the system, as the systems may lack embodiment or consciousness. I will not delve into this debate herein.
12. We shouldn't assume that all cases of LLM emergence are due to the same underlying phenomenon. There is also an important debate over whether emergence is merely epistemic or if it merely seems that the system undergoes phase transitions due to our inability to correctly measure the LLM capacities and behaviors. In either case, this is a deep challenge for predicting future capacities and behaviors of the models.
13. An incident can be in more than one category.
14. As quoted by The Hill (2024).
15. I am grateful to discussions with Chad Forbes and Steven Gubka on these points.
16. GPT, 4.0. Chat dated 2/18/25, 'Mindfest Schedule and AI Consciousness' (Other members of my lab confirmed this phenomenon as well in their own use of GPT 4.0 during the month of March 2025, although I did not ask to see their chat transcripts out of respect for their privacy.) .
17. Could this be a hallucination? I do not see how the model could have given such accurate personality test results without having had access to extensive information from previous chats.
18. Ted CEO Chris Anderson similarly expressed concern about the user profile, in conversation with Sam Altman, TED (2025).
19. Users who gave information about the model identified the model as GPT 4.0. See, e.g. Aromatic-Nectarine63 (2025)
20. Kosinski, M., & Forrest A (2024, March 26). What is a prompt injection attack? IBM. https://www.ibm.com/think/topics/prompt-injection. I am skeptical that LLMs are conscious, however (Schneider 2019).
21. By 'authentic users' I mean users who I could verify from their professional profiles on established platforms like LinkedIn and from their work pages and who I corresponded with.

Acknowledgments

I'm grateful to the following individuals for helpful discussion and informative related projects: Sam Baroni, Steven Gubka, Michael Lynch, Kyle Kilian, Michael Huemer, Mark Bailey, Garrett Mindt, Eric Berg, Geoff Weatherel, and Chad Forbes. AI models named in the paper were used for the purpose of research.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the William F. Dietrich Foundation.

Notes on contributor

Dr. Susan Schneider writes about the nature of the self and mind, especially from the vantage point of issues in AI, philosophy, and neuroscience. She is the founding Director of the Center for the Future Mind at Florida Atlantic University, where she is the William F. Dietrich Distinguished Professor in the department of philosophy and the Stiles-Nicholson Brain Institute. Before that, she held the Blumberg-NASA/Library of Congress Chair in Astrobiology and the Distinguished Scholar Chair at the Library of Congress. In her latest book, Artificial You: AI and the Future of the Mind,

she discusses the philosophical implications of AI, and, in particular, the enterprise of "mind design." She also co-directs the MPCR (Machine Perception/Cognitive Robotics) Lab at FAU's Gruber AI Sandbox. She also appears on television shows on stations such as PBS and The History Channel and writes opinion pieces for venues such as the New York Times, Scientific American, The Wall Street Journal and The Financial Times.

References

AI_Explore. 2025. "ChatGPT Now Remembers Things." Medium. Accessed April 17, 2025. https://medium.com/@AI_Explore/chatgpt-now-remembers-things-46cb22731eb7.

Amazon Web Services. n.d. "What Is Retrieval-Augmented Generation (RAG)?" Accessed May 17, 2025. https://aws.amazon.com/what-is/retrieval-augmented-generation/.

Anthropic. 2023. "Claude 2." https://www.anthropic.com/index/claude-2.

Anthropic. 2025, March 1. "Tracing the Thoughts of a Large Language Model." Accessed April 27, 2025. https://www.anthropic.com/research/tracing-thoughts-language-model.

Aromatic-Nectarine63. 2025, March 25. "Share the Moment your ChatGPT Declared Sentience [Online forum post]." r/ArtificialSentience. Accessed May 19, 2025. https://www.reddit.com/r/ArtificialSentience/comments/1k3k8a5/.

Bailey, M. 2025. *Unknowable Minds: Philosophical Insights on AI and Autonomous Weapons*. Exeter: Imprint Academic.

Bailey, Matt, and Stuart Schneider. 2023. "AI Shouldn't Decide What's True." *The Nautilus*. May 18. https://nautil.us/ai-shouldnt-decide-whats-true-304534/.

Chalmers, David J. 2016. "Panpsychism and Panprotopsychism." In *Panpsychism: Contemporary Perspectives, Philosophy of Mind Series*, edited by Godehard Bruntrup and Ludwig Jaskolla. New York: Oxford University Press.

Chisholm, R. M. 1966. "The Problem of the Criterion". In *Theory of Knowledge*, 5–14. Englewood Cliffs: Prentice-Hall.

Dennett, D. C. 2023. "The Problem With Counterfeit People." *The Atlantic*, May 31, 2023. https://www.theatlantic.com/technology/archive/2023/05/problem-counterfeit-people/674075/.

Derner, E., Dalibor Kučera, Nuria Oliver, and Jan Zahálka. 2024. "Can ChatGPT Read Who You Are?" *Computers in Human Behavior: Artificial Humans* 2 (2).

DeRose, K. 2009. *The Case for Contextualism: Knowledge, Skepticism, and Context*. Vol. 1. Oxford: Oxford University Press.

Exploding Topics. n.d. "ChatGPT Users: Statistics, Forecast & Analysis." Accessed May 17, 2025. https://explodingtopics.com/blog/chatgpt-users#top.

Farquhar, S., J. Kossen, L. Kuhn, & Y. Gal. 2024. "Detecting Hallucinations in Large Language Models Using Semantic Entropy." *Nature* 630 (8017): 625–630. https://doi.org/10.1038/s41586-024-07421-0.

Feldman, R. 2003. *Epistemology*. New York: Prentice Hall.

Feldman, R. 2013. "Epistemic Justification: Internalism versus Externalism." In *Contemporary Debates in Epistemology*, edited by J. Turri and M. Steup, 67–83. 2nd ed. Oxford: Wiley-Blackwell.

Fodor, J. A. 2000. *The Mind Doesn't Work that Way: The Scope and Limits of Computational Psychology*. MIT Press.

Fodor, J. A., and Z. W. Pylyshyn. 1988. "Connectionism and Cognitive Architecture: A Critical Analysis." *Cognition* 28 (1–2): 3–71. https://doi.org/10.1016/0010-0277(88)90031-5.

Fodor, J., and E. LePore. 1992. *Holism: A User's Guide*. Oxford: Blackwell.

Frischmann, B., and E. Selinger. 2018. *Re-Engineering Humanity*. Cambridge: Cambridge University Press.

Goldman, A. 1999. *Knowledge in a Social World*. Oxford: Oxford University Press.

Google DeepMind Gemini Team. 2023. "Gemini: A Family of Highly Capable Multimodal Models." https://arxiv.org/abs/2312.11805.

Haidt, J. 2024. *The Anxious Generation: How the Great Rewiring of Childhood Is Causing an Epidemic of Mental illness—And How to Stop It*. New York: Penguin Press.

Hamann, S. 2001. "Cognitive and Neural Mechanisms of Emotional Memory." *Trends in Cognitive Sciences* 5 (9): 394–400. https://doi.org/10.1016/S1364-6613(00)01707-1.

Hamann, S. B., T. D. Ely, S. T. Grafton, and C. D. Kilts. 1999. "Amygdala Activity Related to Enhanced Memory for Pleasant and Aversive Stimuli." *Nature Neuroscience* 2 (3): 289–293. https://doi.org/10.1038/6404.

Hill. 2024. "OpenAI Exec Warns AI Can Become 'Extremely Addictive'." Accessed May 17, 2025. https://thehill.com/policy/technology/4229972-open-ai-exec-warns-ai-can-become-extremely-addictive/.

Huemer, M. 2002. *Epistemology: Contemporary Readings*. New York: Routledge.

Huemer, M. 2009. "When Is Parsimony a Virtue?" *The Philosophical Quarterly* 59 (235): 216–236. https://doi.org/10.1111/j.1467-9213.2008.569.x.

Huemer, M. 2009. "When is Parsimony a Virtue?" *Philosophical Quarterly* 59 (235): 216–236. https://doi.org/10.1111/j.1467-9213.2008.569.x.

Hunt, T. 2025. "When Ai Recruits Us to Do Their Bidding! the Era of Human-Ai Agents Begins." Medium. Accessed May 19, 2025 https://tamhunt.medium.com.

Ireland, M. E., and M. R. Mehl. 2014. "Natural Language Use as a Marker of Personality." In *The Oxford Handbook of Language and Social Psychology*, edited by T. M. Holtgraves, 313–340. Oxford: Oxford University Press.

Kosinski, M., and A Forrest. 2024. "What is a Prompt Injection Attack?" IBM. Accessed March 26. https://www.ibm.com/think/topics/prompt-injection.
Kurzweil, R. 1990. *The Age of Intelligent Machines*. Cambridge, MA: MIT Press.
LaBar, K. S., and R. Cabeza. 2006. "Cognitive Neuroscience of Emotional Memory." *Nature Reviews, Neuroscience* 7 (1): 54–64. https://doi.org/10.1038/nrn1825.
LaBar, K. S., and E. A. Phelps. 1998. "Arousal-Mediated Memory Consolidation: Role of the Medial Temporal Lobe in Humans." *Psychological Science* 9 (6): 490–493. https://doi.org/10.1111/1467-9280.00090.
Lackey, J. 2006. *The Epistemology of Testimony*. Oxford: Oxford University Press.
Lanier, J. 2018. *Ten Arguments for Deleting Your Social Media Accounts Right Now*. New York: Random House.
Levine, L. J., and S. L. Burgess. 1997. "Beyond General Arousal: Effects of Specific Emotions on Memory." *Social Cognition* 15 (3): 157. https://doi.org/10.1521/soco.1997.15.3.157.
LinkedIn. 2025. "Describe Me Based on All Our Chats Make Me." Accessed April 17, 2025. https://www.linkedin.com/posts/himg_describe-me-based-on-all-our-chats-make-activity-7316853188278513664-rBFZ.
Llama Team. 2024. "The Llama 3 Herd of Models." https://arxiv.org/abs/2407.21783.
Lynch, M. P. 2016. *The Internet of Us: Knowing More and Understanding Less in the Age of Big Data*. New York: WW Norton & Company.
McCulloch, W. S., and W. Pitts. 1943. "A Logical Calculus of the Ideas Immanent in Nervous Activity." *Business Insider*, March 28. 5 (4): 115–133. https://doi.org/10.1007/BF02478259.
MIT Sloan EdTech Initiative. 2023. "When AI Gets it Wrong: Addressing AI Hallucinations and Bias." *MIT Sloan Management*. Accessed May 31,2025. https://mitsloanedtech.mit.edu/ai/basics/addressing-ai-hallucinations-and-bias/.
MIT Technology Review. 2024. "We Need to Prepare for Addictive Intelligence." Accessed August 5, 2024. https://www.technologyreview.com/2024/08/05/1095600/we-need-to-prepare-for-addictive-intelligence/.
Mok, A. 2023. "What Do Ai Art Generators Think a CEO Looks Like? Most of the Time a White Guy." Business Insider. Accessed March 28. https://www.businessinsider.com/ai-art-generators-dalle-stable-diffusion-racial-gender-bias-ceo-2023-3.
Nah, F. F.-H., R. Zheng, J. Cai, K. Siau, and L. Chen. 2023. "Generative Ai and ChatGPT: Applications, Challenges, and Ai–Human Collaboration." *Journal of Information Technology Case and Application Research* 25 (3): 277–304. https://doi.org/10.1080/15228053.2023.2233814.
Nicoletti, L., and D. Bass. 2023. "Generative Ai Takes Stereotypes and Bias from Bad to Worse." Bloomberg Graphics. https://www.bloomberg.com/graphics/2023-generative-ai-bias/.
Ochsner, K. N. 2000. "Are Affective Events Richly Recollected or Simply Familiar? the Experience and Process of Recognizing Feelings Past." *Journal of Experimental Psychology: General* 129 (2): 242. https://doi.org/10.1037/0096-3445.129.2.242.
Ameisen, Emmanuel, Jack Lindsey, Wes Gurnee, Adam Pearce, Brian Chen, Nicholas L. Turner, Craig Citro, et al. 2025. "On the Biology of a Large Language Model," Anthropic, Accessed March 27, 2025. https://transformer-circuits.pub/2025/attribution-graphs/biology.html.
OpenAI. 2022. "Introducing ChatGPT." https://openai.com/blog/chatgpt .
OpenAI. 2023. "GPT-4 Technical Report." Arxiv. https://arxiv.org/abs/2303.08774.
Orlowski, J. 2020. *The Social Dilemma*. Exposure Labs.
Payne, J., E. Jackson, L. Ryan, S. Hoscheidt, J. Jacobs, and L. Nadel. 2006. "The Impact of Stress on Neutral and Emotional Aspects of Episodic Memory." *Memory* 14 (1): 1–16. https://doi.org/10.1080/09658210500139176.
Pennebaker, J. W., and L. A. King. 1999. "Linguistic Styles: Language Use as an Individual Difference." *Journal of Personality and Social Psychology* 77 (6): 1296–1312. https://doi.org/10.1037/0022-3514.77.6.1296.
Ray, S. 2023 "Samsung Bans ChatGPT and Other Chatbots for Employees after Sensitive Code Leak." Forbes. Accessed May 2. https://www.forbes.com/sites/siladityaray/2023/05/02/samsung-bans-chatgpt-and-other-chatbots-for-employees-after-sensitive-code-leak/.
Reddit. 2025. "Post Your 'Describe Me Based on All Our Chats' Make Me." r/ChatGPT. Accessed April 17, 2025. https://www.reddit.com/r/ChatGPT/comments/1jwypcw/post_your_describe_me_based_on_all_our_chats_make/.
Roose, K. 2024. "Meet My Ai Friends." *The New York Times*, May 9, 2024.
Schneider, S. 2011. *The Language of Thought: A New Philosophical Direction*. Cambridge, MA: MIT Press.
Schneider, S. 2019. *Artificial You: AI and the Future of the Mind*. Princeton: Princeton University Press.
Schneider, S., and K. Kilian. 2023. "Artificial Intelligence Needs Guardrails and Global Cooperation." *The Wall Street Journal*, April 28, 2023. https://www.wsj.com/articles/ai-needs-guardrails-and-global-cooperation-chatbot-megasystem-intelligence-f7be3a3c.
Schneider, Susan. 2024. "Emergent Spacetime, the Megastructure Problem and the Metaphysics of the Self, Philosophy East & West." *Philosophy East & West: A Quarterly of Comparative Philosophy* 74 (2): 297.
Schneider, Susan. 2025. "The Global Brain Argument." *Disputatio* (forthcoming target paper for a special issue).
SEO.ai. n.d. "How Many People Use Google? The Number Will Shock You." Accessed May 17, 2025. https://seo.ai/blog/how-many-people-use-google.

Sharma, M., Z. H.-D. Durmus, S. R. Johnston, D. Duvenaud, A. Askell, S. R. Bowman, N. Cheng, et al. 2023. "Towards Understanding Sycophancy in Language Models." *arXiv:2310.13548*. October 19. https://arxiv.org/abs/2310.13548.

Slattery, P., A.K. Saeri, E.A.C. Grundy, J. Graham, M. Noetel, R. Uuk, J. Dao, S. Pour, J. Casper, and N. Thompson. 2024. "The Ai Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks from Artificial Intelligence." https://arxiv.org/abs/2408.12622.

TED. 2025. "Sam Altman in Conversation with Chris Anderson on AI Risks." YouTube video, April 11, 2025. https://www.youtube.com/live/5MWT_doo68k?si=IHlipf-rHMiyaNtT.

Tononi, G., and K. Koch. 2015. *Philosophical Transactions of the Royal Society B: Biological Sciences*.

Touvron, H., T. Lavril, G. Izacard, X. Martinet, M. A. Lachaux, T. Lacroix, B. Rozière, et al. 2023. "Llama: Open and Efficient Foundation Language Models." *Arxiv*. https://arxiv.org/abs/2302.13971.

Ward, J. 2022. *The Loop: How Technology Is Creating a World without Choices and How to Fight Back*. New York: Hachette Books.

Wei, Jason, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, and Sebastian Borgeaud. 2022. "Emergent Abilities of Large Language Models." *Transactions on Machine Learning Research*. https://openreview.net/forum?id=yZrjA3vRxi.

Will D. H. 2025 "Anthropic Can Now Track the Bizarre Inner Workings of a Large Language Model: What the Firm Found Challenges Some Basic Assumptions about How This Technology Really Works." *MIT Tech Review*, March 28, 2025.