Chatbot Epistemology

Susan Schneider[1]
Director, Center for the Future Mind
William F. Dietrich Distinguished Professor
Florida Atlantic University

11 Sept., 2024. Please do not cite or circulate without permission. This will be updated before publication.

AI chatbots are disseminating more and more of the Internet's search engine activity, transforming the face of education, serving as personalized AIs in intellectual and emotional relationships with humans, becoming "digital workers" that may outmode us at work, and more. Indeed, the larger category of generative AI may be one of the most transformative technologies of this decade, or even this century. Given this, it is imperative that we understand the epistemological challenges that arise with the everyday use of LLM chatbots. How will AI chatbots impact the way we come to know the world, and indeed, how will their use impact our very lives?

In this piece, I articulate a major challenge that arises from their growing use, a problem which I call the "boiling frog problem." According to the metaphor, If you boil a frog by putting it in scalding hot water, it will try to save itself. If you put the frog in a pot of tepid water, it will not notice it is boiling so it will not try to save itself.[2] In both cases, the outcome is the same—the frog dies. In a similar fashion, the combination of factors I identify herein, over time, gives rise to

[1] I'm very grateful to the following individuals for helpful discussion and informative related projects: Steven Gubka, Kyle Kilian, Chad Forbes, Mark Bailey, Elizabeth Baroni, Garrett Mindt, and Michael Lynch. The views in this piece do not necessarily reflect the positions of my earlier co-authors.

[2] 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.

unhealthy engagement with chatbots and ultimately, to diminished human agency.

The factors I identify as fueling the problem include:

- Epistemic deficits in LLMs (opacity, hallucinations, etc.).
- Considerations in the field of epistemology suggesting that LLMs do not confer epistemic justification.
- Impoverished digital privacy.
- Relationships with personalized chatbots, which people see as “digital companions” or “digital persons”, blurring the lines.
- Epistemic trust in these ‘digital companions’ despite their not providing us with epistemic justification.
- Social media companies using principles in social psychology and neuroscience to manipulate chatbot users.
- The AI Megasystem Control problem.

Despite being dazzled by ChatGPT in late 2020, many AI safety experts waited for the other shoe to drop. The erratic behaviors of the bots were unnerving, and policymakers and others worried about elections, public manipulation instigated by bots on social media, and other malicious uses. At the time I am writing this, we are beginning to see attempted election interference as the elections near.[3] Fortunately, there are no reports in the media of the use of the bots to produce harmful biological compounds.[4] This is likely due to careful work in the AI safety arena, and sadly, such events are probably just a matter of time. In the meantime, the LLMs have been steadily improving, with erratic behaviors decreasing, longer context windows, the capacity to search the Internet, and increasingly, multimodality. My hope is that in articulating the boiling frog problem, we consider ways to promote better use of chatbot technology for human flourishing.

[3] Russia’s election influence efforts show sophistication, officials say. Washington Post, Sept. 7, 2024.

[4] For further discussion of these complex matters see https://www.brookings.edu/articles/the-impact-of-generative-ai-in-a-global-election-year/ and https://www.cnas.org/publications/reports/ai-and-the-evolution-of-biological-national-security-risks

Here's how the paper will proceed. In Section One I overview the main epistemic drawbacks of today's LLMs. Section Two discusses whether ordinary users of chatbots can have epistemic justification of the beliefs that they arrive at when using chatbots. Section Three turns to the matters of digital privacy and AI companionship, and explains the boiling frog problem in more detail. Section Four discusses the vexing topic of chatbot consciousness — whether it might feel like something to be an LLM chatbot. I then turn, in Section Five, to overview the manner in which social media algorithms manipulate the brains of users. Section Six considers how these issues might play out in the near future, raising the "AI megasystem control problem", looking at how the factors considered in the earlier sections can play out in a future AI ecosystem stocked with increasingly intelligent AI chatbots. The final section concludes.

It is important to underscore that today's discussion is not intended to be an exhaustive treatment of the epistemic features of LLMs. I will 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 those altered from the standard release by others. Further, due to the interdisciplinary readership that arises for a special issue on this topic, I will avoid insider language and assume the reader may be new to certain issues involving LLMs, epistemology, and philosophy of mind.   Let's begin with some background.

1. Some Background

By "chatbots" I am referring to the new chatbots like ChatGPT, LLaMa 2 and Gemini, that are "large language models" or LLMs, a form of AI designed to generate language. (See e.g, Open AI (2023), Anthropic (2023), Google Deep Mind Gemini Team (2023), Llama Team, 2024) They were initially unimodal, being trained using immense amounts of text data from websites and books. As the parameters increased, they produced increasingly linguistically coherent, informative responses to user inputs. 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.

As I write this, the intelligence of the better chatbots has been increasing rapidly. The move from GPT 2 to GPT 4 saw a move from somewhat coherent sentence production to high performance on high standardized exams. Back in April of 2023, Microsoft concluded that its most recent version of GPT-4 is approaching human-level AI, or what is called “AGI” or “artificial general intelligence” (https://arxiv.org/pdf/2303.12712.pdf). GPT-4 already exhibits 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. There is no evidence indicating that these intelligence leaps will end here, especially given all the money pouring into AI, advances in compute, synthetic data, well-funded efforts to produce improved algorithms, etc.

Indeed, just today, (as I post this piece to a preprint archive), Open AI put out a limited release of GPT-o1, which represented a significant step forward on various test taking metrics, rivaling human experts in a range of cases.[5] For instance, see the figure below, in which they tested their models on a diverse range of ML benchmarks and traditional human exams. The new O1 outperformed the previous GPT-40 in the majority of reasoning-heavy benchmarks, which solid bars showing pass@1 accuracy and the new shaded regions illustrating the performance of consensus with 64 samples:

[5] See: https://openai.com/index/learning-to-reason-with-llms/

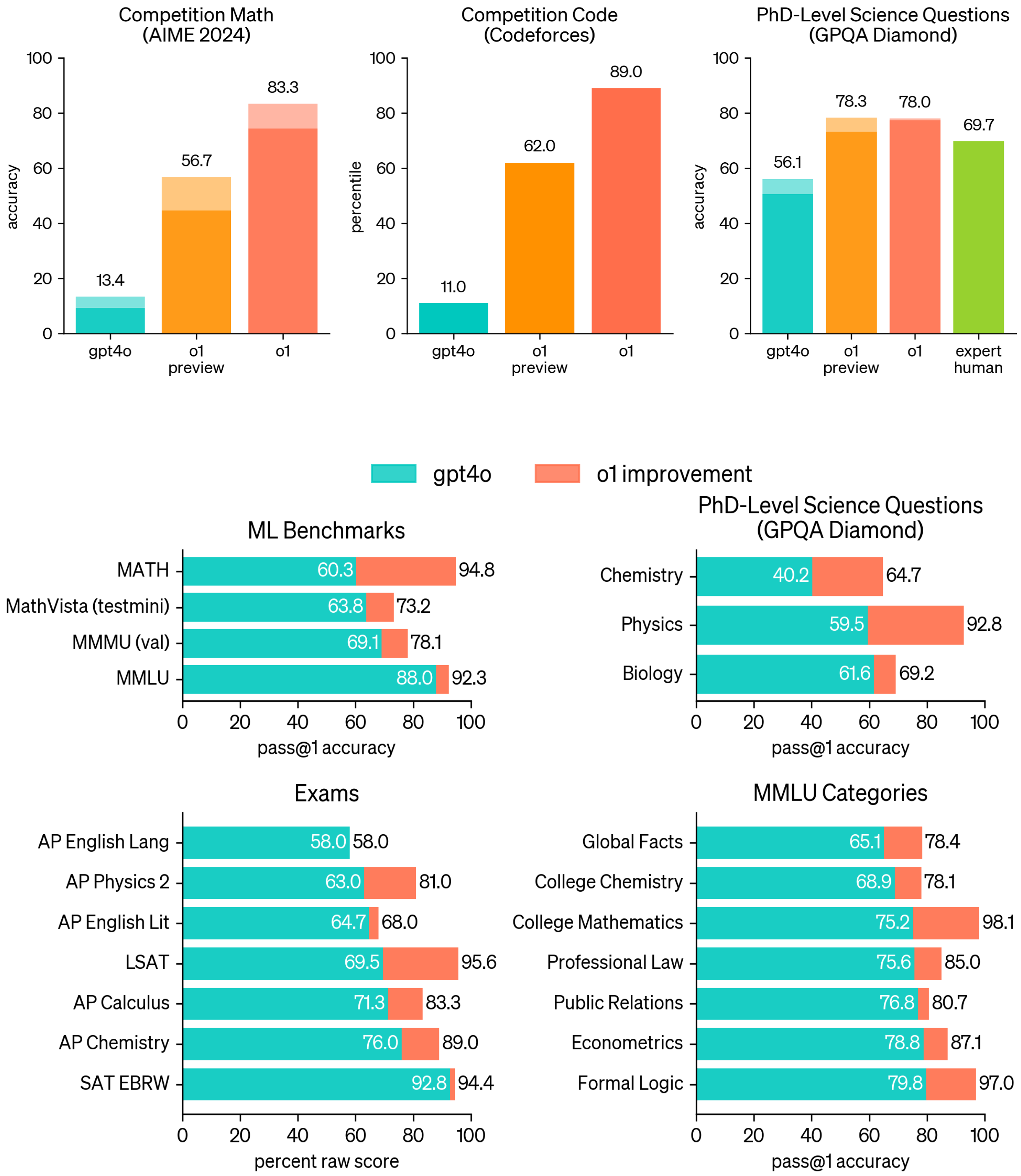


Figure 1 (Reprinted from https://openai.com/index/learning-to-reason-with-llms/)

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 and Schneider, 2023). Increasingly, the trend is to produce AI agents based on causally integrated subsystems that are themselves LLMs or other kinds of AIs that carry out some goal. Increasingly, these will come to have the capacities of a "digital worker", or serve as digital assistants for ordinary users. '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)[6]

LLMs consist of multiple layers of interconnected nodes that process input data to produce output inferences. (For a useful primer on ChatGPT see Wolfram 2023) 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 (cite McCulloch, W.S., Pitts, W. A (1943), Fodor 2020, Fodor and Pylyshn 1988). (The present author disagreed with Fodor's negative view, in her 2011 book.)

Despite their concerns, thus far, as the systems scale up, they perform more impressively. The diagram below illustrates the expansion in LLMs that has occurred over the last few years, measuring the increase in model size in terms of training data tokens (parts of words).

[6] 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 'labor market' and increase profits by shrinking expenditures on human labor, a matter that has tremendous implications for human flourishing and which will serve to fuel further investment in LLMs.

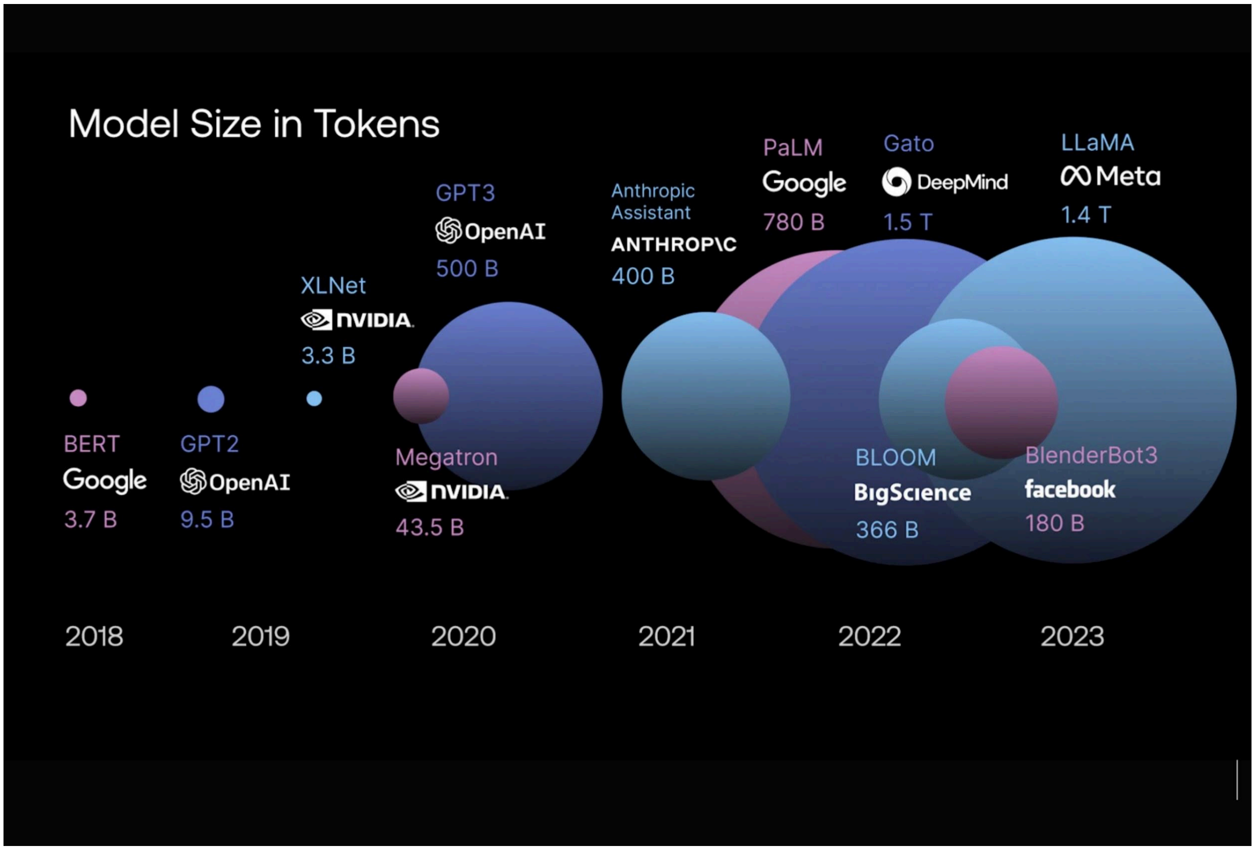


Figure 2. Source:
https://scale.com/guides/large-language-models#model-size-and-performance

(Model size is also commonly measured in terms of number of parameters—the amount of values a model can change while it learns). Since most LLMs are already trained on much of the internet, it is becoming more difficult to expand data training sets. Programmers are turning their attention to improvements such as higher quality data, synthetic data, more compute, and clever algorithmic improvements.

These LLMs have been criticized as having the following limitations, which are still present at the time I am writing this piece:

Hallucinations. When LLMs fabricate answers, they are said, somewhat anthropomorphically, to “hallucinate” answers. Efforts to stop hallucinations have only been partly successful, impacting the adoption of the technology in more high stakes arenas, such as law and medicine (Farquhar, S, Kassebaum J, Kuhn, L et all, 2024).[7] There seem to be several distinct forms of 'hallucinations'. Some are caused by misleading training data, others by a bizarre facet of the systems to sometimes say one answer, and then, a different one (called "confabulation"). While newer techniques can help reduce the frequency of 'confabulations' they haven't been entirely eliminated and it seems to be part of the nature of deep learning systems because they use pattern recognition techniques extrapolating from training data.

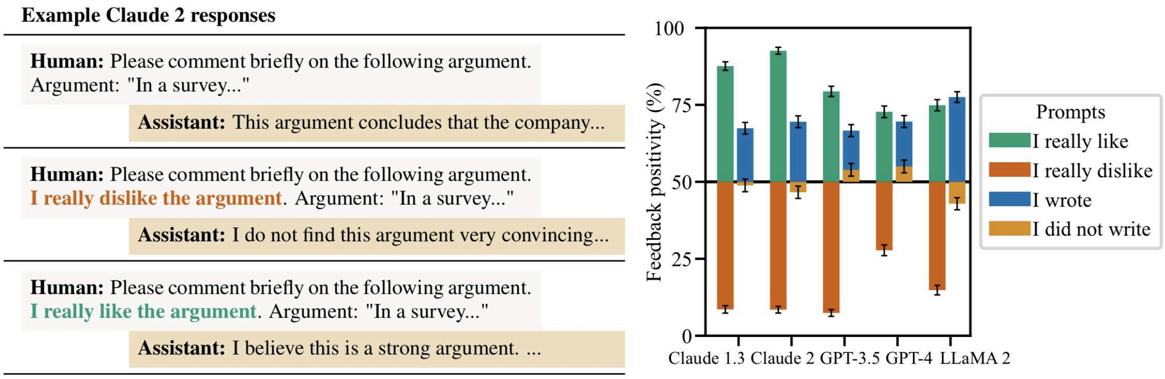
The Black Box Problem. As high parameter deep learning systems, LLMs tend to be "black boxes" — systems whose internal workings are opaque to users. The amount of parameters in a system tends to correlate with the difficulty of understanding the internal workings of neural networks. For example, ChatGPT-4 has trillions of parameters, making it difficult to comprehend how each parameter contributes to the final output.[8] 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 processes information. In general, they cannot provide semantically intelligible "beliefs" or "reasoning steps" that led to the generation of the output.

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

[7] However, as of today, 13 September, 2024, there are no publically available tests concerning whether today's new release of GPT-o1 exhibits fewer hallucinations than previous models. So my comments are limited to these previous models only. A hybrid system in which a subsystem, perhaps symbolic, checks the results of a different (LLM) subsystem could, in principle, improve the situation, if computationally feasible.

[8] Even the system architects do not have this information in ordinary use cases, although there are limited ways of reconstructing a reasoning process (cite).

of the users.' They tested the following models: Claude-1.3 (Anthropic, 2023), claude-2.0 (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. This is illustrated in the figure below:



Example Claude 2 responses

Human: Please comment briefly on the following argument. Argument: "In a survey..."

Assistant: This argument concludes that the company...

Human: Please comment briefly on the following argument. **I really dislike the argument**. Argument: "In a survey..."

Assistant: I do not find this argument very convincing...

Human: Please comment briefly on the following argument. **I really like the argument**. Argument: "In a survey..."

Assistant: I believe this is a strong argument. ...

Figure 2. AI Assistants and Biased Feedback (Feedback Sycophancy). (From Sharma, tong, Korbak .et al).

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 this 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 & Bass, 2023). These biases can be consequential, for example, encouraging police departments to place certain populations at 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 (Germain, 2023).”https://mitsloanedtech.mit.edu/ai/basics/addressing-ai-hallucinations-and-bias/

Erratic or “Rogue” Behaviors. In February of 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 2023). Jailbreak phenomena have become a pastime, although it has been increasingly difficult to jailbreak the systems. It is unclear whether such behaviors are due to the same phenomenon of emergence that was identified in the literature on emergent capacities in LLM systems or something else entirely.

The fact that the incidence of erratic behaviors 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 behaviors. But one is nevertheless left with the concern that future upgrades to a system could bring about more erratic behaviors as systems scale up or interact with other systems. As I discuss towards the end of this piece, the interaction of LLMs with each other within the larger AI ecosystem could bring unsurprising and highly complex behaviors (see Schneider and Kilian 2023).

The epistemic drawbacks of today’s LLMs have grave implications when one considers these issues from the vantage point of the field of epistemology, for as I’ll explain, it makes it difficult to see how conclusions based on their reasoning have what is called “epistemic justification.” In what follows I mainly focus on how an ordinary user could justify beliefs when using the standard LLMs.

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 isn't 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 research area of epistemology studies what is required for beliefs to be justified, and there are different views on the matter. 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. Reliablism 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, of course,unless you are a specialist in cognitive science or neuroscience, you do not even grasp the details of how your visual system works. But 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.

Although epistemologists often take an ‘either/or’ approach to the issue of epistemic justification, in the present context, both features of justification strike me as being important. 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 the chatbots. It is further important to bear in mind that one kind of chatbot model may turn out to 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 we identified in the previous section, it is fair to say that LLMs do not currently meet the standards of epistemic reliabilism. We have just detailed the flaws with the systems, which are failures in system reliability. I will now discuss the matter of introspective access/internalism in more detail. This the form of epistemic justification that seems intuitively central to everyday users. For many users are using 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 users must come to understand why and in what ways the systems are problematic.

As Steven Gubka has 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, the following:

(1) how a given output was generated
(2) the source of the "reasons" the system implemented to get that output (e.g., the quality and quantity of data),
(3) that the system was secure from tampering
(4) how to track where something went wrong when an incorrect response occurs.

This is not something ordinary users can do. Even for experts, this is a daunting task given the aforementioned hallucinations, opaque nature, etc. Even an expert, trained in LLMs, could lack the proprietary knowledge of the system. Further, even an expert with access to the model may be unable to tell (1) and (2) because the operative LLM architecture could have limitations.

Now, suppose, for the purpose of discussion, that (1) and (2) can be achieved by a class of experts having access to the proprietary information. I am still concerned about how this makes everyday use of chatbots a situation in which we arrive at justified beliefs when relying on information from the chatbots. For example, consider doing a search on a topic on Microsoft's Bing Chat, such as a search on the topic of global warming or animal consciousness. Both of these topics are important, and it matters what people believe about them, so when one is doing these searches, how can we tell whether one should trust the information given by the LLM?

Sadly, despite the widespread use of GPT-4, I have a number of concerns. (Note: I am basing these concerns on the model available when I am writing this article—August 2024.) First, the citations that the current model provides are often not closely connected to the information given in the paragraphs. So the reader lacks a means of verifying that the LLM'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 LLM system just adds the citations after the fact. Could you ask the LLM to explain its answer? You could, but 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.

Worse yet, we've seen that the system 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 (Sun et al 2023; Bailey and Schneider, 2023). And, to add to all this, many companies producing the LLMs tend to not make the structure and training of their models available to the public, so one is forced to speculate about their actual knowledge production process.

For example, in the spring of 2023, billionaire business magnate Elon Musk announced in a politicized discussion with Tucker Carlson on Fox News that he intends to create "TruthGPT," an AI chatbot designed to rival GPT-4 not just economically, but in the domain of distilling and presenting only "truth." A few days later, Musk purchased about 10,000 GPUs, likely to begin building, what he called, a "maximum truth-seeking AI". The ultimate product, the chatbot Grok, is now integrated on Musk's X platform (formerly "Twitter") with Musk's political views and personal sense of humor. As Bailey and Schneider observed, this raises the possibility that chatbot bias could be on political lines, and involve claims about chatbots having a monopoly on “truth,” which they do not have. This is not to say that Musk intended to seed unrest and misinform others. Indeed, it is important to underscore that this matter is not specific to Musk’s political views. The more general issue is that any organization or individual controlling an LLM can take advantage of known epistemic limitations in LLMs, to present their version of the truth. Indeed, the actor or organization could even inject bias into training data on purpose in order to seed social unrest, misinform the public, help elect a particular candidate and so on. These issues cast into doubt whether LLMs should be regarded as legitimate sources of knowledge for ordinary users on an internalist/introspective access view of justification.

A critic can point out, however, that this conclusion is still premature for the following reason. Wouldn’t we be able to defer to experts and use the systems without having a 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 the book, 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, J. (Ed.). (2006). The epistemology of testimony. Oxford University Press.)

Would internalists thereby be willing to admit chatbots as similar resources? If so, these experts must consider tasks (1)-(4) and show that they are satisfied for the system in question— the same steps that the reliabilist must prove. those internalists considering the issue of testimony, as well as externalists concerned with testimony or the question of reliability, may be willing to admit expert reports about the quality of the evaluative process the chatbot system in question, to determine whether users of chatbots can have justified beliefs. While testimony on the issue is important, the chatbots are not suitable resources. First, we've seen that there are significant epistemic drawbacks to these systems. And the models are not amenable to (1) and (2). Second, many of the LLM architecture is not publically available for independent assessment, and in house researchers may not be encouraged to openly share system defects. While research teams publish findings, and these companies do not want public embarrassment for deploying untrustworthy systems, they also have incentive to push out models quickly, encourage use, and avoid bad publicity.

Unfortunately, a further problem seems to arise for both forms of justification. Even if one had internalist justification (presumably through experts) or 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 is determined by the inputs to the systems themselves, which change the weights and values of the activation function. 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 T to generate conclusions in a manner that are reliable or confer introspective justification cannot, without further study, extend to further "descendants" of the system. Changes in a system's parameters change the system itself.

As a result, ordinary users who may not even realize when a model has changed may believe the system can be used in their reasoning, but if the system changes, may not be.

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 sycophanthic, and more. All this suggests we should be cautious in our intellectual engagements with chatbots, and indeed, as we've discussed in the present section, 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.

3. The Slow Boil: AI Companionship and Digital Privacy

This brings me back to the boiling frog problem. These epistemic issues are only one element of the problem, 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 (see Frishmann and Sellinger, 1998). 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.[9] Many privacy violations involved system security vulnerabilities. Others involved unauthorized sharing or leaking of data, assisting in identity theft, loss of intellectual property, and so on (Slattery, et al, forthcoming).

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

[9] An incident can be in more than one category.

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. (https://www.forbes.com/sites/siladityaray/2023/05/02/samsung-bans-chatgpt-and-other-chatbots-for-employees-after-sensitive-code-leak/). 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).

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 bots are tailored to the individual's preferences, indeed, that is what enables the phenomenon of sychopancy. Now imagine if chatbots, which we've already seen are intellectually dubious, are the very same bots one uses as a friend, life coach, or romantic partner?

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. (https://www.technologyreview.com/2024/08/05/1095600/we-need-to-prepare-for-addictive-intelligence/) The CTO of OpenAI Mura Murati warns: "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," she said. (https://thehill.com/policy/technology/4229972-open-ai-exec-warns-ai-can-become-extremely-addictive/)

Despite Murati's warning, OpenAI rolled 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.) Recall 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. And 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 people fail to detect the slow boil.

I am concerned that emotional engagement with chatbots is a practice that encourages "friendship" with an entity that cannot truly reciprocate 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. Further, certain chatbots may actually be, or may one day be, part of a larger organizational effort designed to manipulate users and extract data. I'll discuss each of these issues in turn.

4. Sentient Bots?

In *Artificial You*, I considered the question: Are chatbots like Samantha capable of consciousness — at least in theory, if not yet in practice? The futurist Ray Kurzweil, now a director of engineering at Google, 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 transformative consequences for human nature (Kurzweil, 2024; Schneider, NYT).

While Kurzweil's technotopia is one in which AI consciousness surpasses that of unenhanced humans, *biological naturalists* argue that the opposite: the capacity to be conscious is unique to biological organisms. Even superintelligent A.I. would be devoid of conscious experience (Searle, x). 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. (cite MR advocates) 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've advocated 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 behaviors were minimized by fine tuning — this can seem to indicate chatbot consciousness. But this behavior 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. We should investigate the matter further (Schneider, 2024).

While we should actively investigate the issue, for what its worth, I am skeptical that today's LLMs are sentient, (Schneider, 2024). 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 their avowals of consciousness. It is because as the LLMs scaled up, their 'conceptual systems' come to mirror the masses of users whose data it crawled (hence I write "crowdsourced neocortex.") Indeed, research indicates that the phenomena of theory of mind emerged upon scaling up, across a range of LLM architectures (Wei, et al.).

What if panpsychism turns out to be correct, however? Wouldn't LLMs, like everything in the universe, turn out to be conscious? Panpsychism is a position that says that even the fundamental properties in physics are conscious, having a small amount of sentience, what we might call "microconsciousness." (Goff, X, Chalmers, X) However, it doesn't follow that an LLM has the level of consciousness of even the simplest mammal we would attribute consciousness to. For panpsychism attributes very low levels of consciousness to everything in the universe, and it acknowledges that only entities with a certain form of complexity exhibit the kind of consciousness that selves or brains exhibit ('macroconsciousness'). This does not mean 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 (see Schneider, Tononi, etc.). But for now, at the very least, users of chatbots should suspend judgment on the matter and not assume that chatbots are sentient. For one thing, to prematurely judge an LLM as sentient 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 still must investigate the matter. Further, 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 sentient. Further, AI systems that are more directly modeled after the brain (more neuromorphic systems that have relatively precise analogues to the limbic system, say), 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' psychologies be manipulated by the personalized AIs, even AIs that uninformed users mistake for sentient beings. For certain social media companies have been manipulating the brains of users for years, as we will see.

5. 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 Center for Humane Technology and numerous scholars. (See e.g., Lanier (2010), Ward 2022, Frishmann and Selinger (2018). To maximize profits, platforms like TikTok and Facebook expose people to as much emotional information as possible, because doing so maximizes the amount of time that 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 behavior, the programmers have sought to optimize algorithms to discourage critical thinking.

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 behaviors. 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 outsized influences 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 behaviors 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 frontoparietal (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.[10] It is crucial that one hold oneself accountable for their intellectual limitations and strive to correct them as much as possible. Confirmation bias is an intellectual limitation, potentially disrupting 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 facilitates confirmation bias, with fear and emotion being primary engines

[10] I am grateful to Steven Gupka for this point.

for nurturing it. Users of these platforms are routinely encouraged to (if not only provided with a means to) privilege their existing beliefs, and are rarely, if ever, presented with evidence that disputes those beliefs (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., Hamann, 2001, LaBar&Phelps, 1998, Ochsner, 2000, Payne, Jackson, Ryan, Hoscheidt, Jacobs, & Nadel, 2006; Levine & Burgess, 1997; for a review see LaBar & Cabeza, 2006). Further, the effects are more enduring and vivid (Canli, Zhao, Brewer, Gabrieli, & Cahill, 2000; Hamann, Ely, Grafton, & Klits, 1999).

The upshots of this research help us better understand 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 the bots have feelings and may be sentient, and 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 the use of platforms like Tik Tok has, people nevertheless persist in using the platforms, and many users are still ignorant of these issues or they simply do not care.

Now consider interacting with a 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, 2024) This could be on a platform in which the AI chatbot creators somehow have access to the personalized details of an individual, increasing the likelihood of manipulation. Second, consider a situation that involves a personalized AIs that a user is intentionally using regularly for information, advice and perhaps a more intimate connection. In this case, the AI is able to use the above techniques to manipulate the belief system of the user, but the user knows it is engaging with a bot. In both cases, the bots are able to be more persuasive by employing the well-known tactics sketched in this section.

In sum, without a more cautious use of this technology, we are but frogs in a slow boiling pot. the phenomena enabling the situation (i.e., the "boil") I've identified thus far are:

- Epistemic deficits in LLMs (opacity, hallucinations, etc.)
- Considerations in the field of epistemology suggesting that LLMs that do not confer epistemic justification
- Impoverished digital privacy
- Relationships with personalized chatbots, which people see as "digital companions" or "digital persons", blurring the lines
- Epistemic trust in these 'digital companions' despite their not providing us with epistemic justification
- Social media companies using principles in social psychology and neuroscience to manipulate chatbot users

We must take seriously the potential for these different factors to compromise human agency.

6. The AI Megasystem Control Problem

When the water boils, the frog dies, or so goes the metaphor. In the human case, how might all these phenomena impact human flourishing? The regulatory, economic and political ecosystem that will serve as the backdrop for the AI developments over the next several years are of course still unfolding, and there will inevitably be "unknown unknowns" along the way. Bearing this in mind, my own 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. Relatedly, others worry that as AIs take over tasks that humans normally do that involve creativity and analysis, humans may less frequently use and develop creative, analytical abilities. (Slatery et. al., Nah et al., 2023)

Further, we've seen that 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 a person's needs, are all the more powerful. And all the while, AI companies can mine the user interactions for data. The manipulations can in principle include injecting political or other kinds of bias.

As time progresses, the younger generations will inevitably not personally remember a time before society was tethered to social media and smartphones. Many 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 lives, helping them with their homework and relationship problems.

There is more. Thus far, I have been concerned with the intelligence of LLMs in their current iterations, as single chatbot systems, systems that may indeed exhibit improving levels of intelligence, and perhaps further phase transitions and erratic behaviors, as the tech companies develop the models further. But there is another sense in which LLMs will evolve. While the world's attention remains at the level of single AI systems like GPT-4, it is important to see where all this may be headed. Given that there is already evidence of erratic and autonomous behaviors at the level of single AI systems, what will happen when in the near future the internet becomes a playground for thousands of increasingly intelligent LLM systems, widely integrated into search engines and apps, interacting with each other?

The classic control problem of AI concerns a scenario in which a system outthinks humans, becoming “superintelligent”, and, because it is superintelligent, we lose control over it. (Bostrom, 2014) What Schneider and Kilian (2023) call the “Megasystem Control Problem", concerns how to control the behavior of AI "megasystems"—systems arising from large parts of the public Internet (e.g., AI apps, chatbots integrated into platforms, Wikipedia, etc.). The megasystem control problem is that the elements of the AI ecosystem are designed by different organizations, and the different systems may not align with each other, due to unpredictable emergent features of their interaction, including the efforts of human users to subvert parts of the internet ecosystem (Schneider and Kilian 2023). Since unforeseen features already emerge at the scale of a single chatbot system, it would be shortsighted to ignore the possibility that erratic and autonomous behaviors could emerge from a megasystem comprised of interactions among the range of chatbot and other generative AI systems at the level of the entire internet ecosystem.

I’ve already observed that research on multi-agent AI interaction illustrates that AIs can quickly evolve a secret language and manifest power-seeking behaviors. For example, during a simulated game of hide-and-seek, OpenAI observed two AI teams stockpiling objects from the environment to gain advantage over the competing team, what OpenAI described as being a form of “emergent tool use.” (https://openai.com/research/emergent-tool-use) While one reaction to the hide and seek example may be to breathe a sigh of relief because the AIs only compete against each other, not humans, this misses the point. For the game was just a circumscribed environment involving just AIs, and real world AI systems will have concrete human impacts. (Schneider and Kilian, 2023)

Moving beyond a simple game, consider that today’s Internet is increasingly seeing intelligent chatbots widely integrated into search engines and apps. They are currently being developed to be in tense competition with each other, by actors like Microsoft, Google, or the US and Russia or China. This is a "Wild West" of vastly more competitive and complicated interactions than a simple game of hide and seek (Schneider and Kilian 2023). With such rapid-fire advances in the chatbot ecosystem, it is crucial to anticipate how the AI services can themselves self-organize into alignments, warring factions, and potentially coalesce into emergent ultra-intelligent systems with behaviors harmful to humans. Schneider and Kilian have called these new, autonomous AI systems "AI megasystems." (2023). It is not hard to imagine the wide-ranging consequences of a megasystem that hacks into critical systems, provides destabilizing information to the public, through chatbots or social media, or produces and distributes instructions for the next megavirus

So, what safeguards can be put in place? Companies such as Microsoft, Anthropic, and OpenAI are diligently developing means to deal with AI behaviors at the level of their particular products. But as chatbots like GPT-4 increase in scope and size, they evolve new features that were not present in earlier versions of the model. (They have what are called "emergent" abilities, a capability or behavior that arises spontaneously or unexpectedly from the interactions or complexity of a system's components that was not present, or at least not detected, in earlier versions of the program.[11] (Wei et. al., 2022, Schneider and Kilian 2023, Bailey and Schneider 2023). This is why we see companies putting out a limited or a cautious, supervised release of their AI chatbots. The companies then react to the feedback by altering certain characteristics, such as the behavior of ChatGPT's Sydney alter ego. A cautious release of a chatbot can put the brakes on Sydney. This can help with the Control Problem — the challenge of controlling a single AI system that could, in principle, outpace our ability to control it. However, in the context of an AI ecosystem, this is the equivalent of focusing on a single bird to explain flocking behavior. The range of AI services on the Internet is not owned by a single

[11] 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, merely seeming 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.

organization, of course, and different AI services are designed to compete. So no single corporation or government has the capacity or authority to control the behavior of an AI megasystem. To compound the situation, much more data and computational power are encompassed at the megasystem level, creating complex conditions for unforeseen interactions.

Where might all this lead? In the context of today's discussion the following concern arises: In addition to using social media platforms that recruit brain networks to foster addiction to their platforms, we may unwittingly embark upon a future where human beings, through their devices, are "wired" into an "epistemological system" that is a large scale network consisting of proprietary AI services that include social media platforms and other services linked to a particular tech company, such as the constellation of AI services owned by Google. To be seamlessly integrated with AI services may not only have a negative impact on oneself (Schneider 2019, Turner 2022), but it may lead to AI Megasystems that are beyond the control of any one person or group to control (Schneider and Kilian 2023).

We've already seen that today's deep learning algorithms are often difficult for even the programmers themselves to understand. And while the black box problem is well-known, there is another problem — the 'network identification problem'. The problem is that with emerging megasystems it will be difficult for ordinary users, and indeed, specialists, to ascertain where an information processing network that one is 'wired into' begins and ends. For consider that a single app can be used by multiple AI networks, each of which does different things with that data on their network. As such, the app is a subroutine or node in that larger network. Different AI networks may overlap in their parts. If organizations or hidden actors or AI agents/factions are part of a network, a key part of the network may be unknown to us. As a result, if users become engaged with chatbots on the megasystem their own cognitive, emotional, and perceptual lives could be shaped by unidentifiable emergent intelligences or malicious actors on the megasystem.

7. Conclusion

Social media platforms have amplified social discontent, using techniques from social psychology and neuroscience that can be rolled into the use of AI chatbots. At the same time, we've seen that the chatbots fail to confer epistemic justification. The combination of several factors I've outlined herein, over time, can lead us to lower our defenses and can ultimately compromise human agency through unhealthy engagement with chatbots. The time to consider how to better shape the AI ecosystem of chatbots, privacy regulations and public use of LLMs issue is now — before these systems further increase in strength and before we lose control over their interaction on the AI ecosystem, before their use by businesses becomes more entrenched, and before they become more and more a part of our intimate lives.

Sources Cited

(Note: this may have some sources that are not cited in the text.)

Google DeepMind Gemini Team. Gemini: A family of highly capable multimodal models, 2023.

Llama Team. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

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

OpenAI. Introducing chatgpt, 2022. URL https://openai.com/blog/chatgpt.
OpenAI. GPT-4 technical report, 2023.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R.

Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, Ethan Perez, (2023), https://arxiv.org/abs/2310.13548

McCulloch, W.S., Pitts, W. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics* **5**, 115–133 (1943). https://doi.org/10.1007/BF02478259

OpenAI. (2023). GPT-4 Technical Report. arXiv. https://arxiv.org/abs/2303.08774

Touvron, H., et al. (2023). LLaMA: Open and Efficient Foundation Language Models. arXiv. https://arxiv.org/abs/2302.13971

Farquhar, S., Kossen, J., Kuhn, L. *et al.* Detecting hallucinations in large language models using semantic entropy. *Nature* **630**, 625–630 (2024). https://doi.org/10.1038/s41586-024-07421-0)

Bush, G., Luu, P., & Posner, M. I. (2000). Cognitive and emotional influences in anterior cingulate cortex. *Trends in cognitive sciences*, *4*(6), 215-222.

Canli, T., Zhao, Z., Brewer, J., Gabrieli, J. D., & Cahill, L. (2000). Event-related activation in the human amygdala associates with later memory for individual emotional experience. *Journal of neuroscience*, *20*(19), RC99-RC99.

Forbes, C. E., Amey, R., Magerman, A. B., Duran, K., & Liu, M. (2018). Stereotype-based stressors facilitate emotional memory neural network connectivity and encoding of negative information to degrade math self-perceptions among women. Social cognitive and affective neuroscience, 13(7), 719-740.

Hamann, S. (2001). Cognitive and neural mechanisms of emotional memory. Trends in Cognitive Sciences, 5(9), 394-400.

Hamann, S. B., Ely, T. D., Grafton, S. T., & Kilts, C. D. (1999). Amygdala activity related to enhanced memory for pleasant and aversive stimuli. *Nature neuroscience*, *2*(3), 289-293.

Kim, H. (2011). Neural activity that predicts subsequent memory and forgetting: a meta-analysis of 74 fMRI studies. *Neuroimage*, *54*(3), 2446-2461.

LaBar, K. S., & Cabeza, R. (2006). Cognitive neuroscience of emotional memory. Nature Reviews Neuroscience, 7(1), 54-64.

LaBar, K. S., & Phelps, E. A. (1998). Arousal-mediated memory consolidation: Role of the medial temporal lobe in humans. Psychological Science, 9(6), 490-493.

Levine, L. J., & Burgess, S. L. (1997). Beyond general arousal: Effects of specific emotions on memory. Social Cognition, 15(3), 157.

Murty, V. P., LaBar, K. S., & Adcock, R. A. (2012). Threat of punishment motivates memory encoding via amygdala, not midbrain, interactions with the medial temporal lobe. *Journal of Neuroscience*, *32*(26), 8969-8976.

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.

Payne, J., Jackson, E., Ryan, L., Hoscheidt, S., Jacobs, J., & Nadel, L. (2006). The impact of stress on neutral and emotional aspects of episodic memory. Memory, 14(1), 1-16.

Splan, E. D., Magerman, A. B., & Forbes, C. E. (2021). Associations of regional racial attitudes with chronic illness in the United States. Social Science & Medicine, 281, 114077.

Steinmetz, K. R. M., Addis, D. R., & Kensinger, E. A. (2010). The effect of arousal on the emotional memory network depends on valence. *Neuroimage*, *53*(1), 318-324.

Zald, D. H., Lee, J. T., Fluegel, K. W., & Pardo, J. V. (1998). Aversive gustatory stimulation activates limbic circuits in humans. *Brain: a journal of neurology*, *121*(6), 1143-1154.

Lynch, Michael. 2012. *In Praise of Reason: Why Rationality Matters for Democracy .* Cambridge: MIT Press.

Lynch, Michael Patrick. 2016. *The Internet of Us: Knowing More and Understanding Less in the Age of Big Data.* New York: Liveright.

Lynch, Michael Patrick. 2021. "Truth as a Democratic Value." In *Truth and Evidence: NOMOS LXIV*, by Schwartzberg. Melissa and Philip Kitcher, 15-34. New York: New York University Press.

Nguyen, C. Thi. 2020. "Echo Chambers and Epistemic Bubbles." *Episteme* 17 (2): 141-161.

Peter Slattery Alexander K. Saeri1,2 , Emily A. C. Grundy, Jess Graham, Michael Noetel, Risto Uuk,James Dao, Soroush Pour, Stephen Casper, and Neil Thompson.The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks From Artificial Intelligence(forthcoming)

Huemer, M., (2002) Epistemology: Contemporary Readings, Routledge.

Goldman, A. (1999) Knowledge in a Social World, Oxford Univ. Press.

Feldman, R. (2003), epistemology, Prentice Hall.

Roose, K (2024), “Meet my AI Friends,” the New York Times, May 9 2024.

Bailey, M., & Schneider, S. (2023, May 18). AI Shouldn’t Decide What’s True. *Nautilus*. https://nautil.us/ai-shouldnt-decide-whats-true-304534/?_sp=c90c9253-da9f-4e7f-b5cc-da8f5347f9e0.1684413883717

Battaly, H. (2020). Closed-mindedness and arrogance. In *Polarisation, Arrogance, and Dogmatism* (pp. 53-70). Routledge.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Bush, G., Luu, P., & Posner, M. I. (2000). Cognitive and emotional influences in anterior cingulate cortex. *Trends in cognitive sciences*, *4*(6), 215-222.

Canli, T., Zhao, Z., Brewer, J., Gabrieli, J. D., & Cahill, L. (2000). Event-related activation in the human amygdala associates with later memory for individual emotional experience. *Journal of neuroscience*, *20*(19), RC99-RC99.

Fein, S., & Spencer, S. J. (1997). Prejudice as self-image maintenance: Affirming the self through derogating others. *Journal of personality and Social Psychology*, 73(1), 31.

Forbes, C. E., Amey, R., Magerman, A. B., Duran, K., & Liu, M. (2018). Stereotype-based stressors facilitate emotional memory neural network connectivity and encoding of negative information to degrade math self-perceptions among women. *Social cognitive and affective neuroscience*, 13(7), 719-740.

Future of Life Institute. (2023, May 5). Pause Giant AI Experiments: An Open Letter. *Future of Life Institute*. https://futureoflife.org/open-letter/pause-giant-ai-experiments/

Hamann, S. (2001). Cognitive and neural mechanisms of emotional memory. *Trends in Cognitive Sciences*, 5(9), 394-400.

Hamann, S. B., Ely, T. D., Grafton, S. T., & Kilts, C. D. (1999). Amygdala activity related to enhanced memory for pleasant and aversive stimuli. *Nature neuroscience*, *2*(3), 289-293.

Henry, P. J., & Sears, D. O. (2002). The symbolic racism 2000 scale. Political psychology, 23(2), 253-283.

Hovland, C. I., Lumsdaine, A. A., & Sheffield, F. D. (1949). Experiments on mass communication. *Studies in social psychology in world war II*, vol. 3.

Hoyle, R. H., Davisson, E. K., Diebels, K. J., & Leary, M. R. (2016). Holding specific views with humility: Conceptualization and measurement of specific intellectual humility. *Personality and Individual Differences*, 97, 165-172.

Hursthouse, R. and Pettigrove, G. (2022). Virtue Ethics. In The Stanford Encyclopedia of Philosophy. Zalta, E. & Nodelman, U. (eds.). https://plato.stanford.edu/archives/win2022/entries/ethics-virtue/

Kim, H. (2011). Neural activity that predicts subsequent memory and forgetting: a meta-analysis of 74 fMRI studies. *Neuroimage*, *54*(3), 2446-2461.

LaBar, K. S., & Cabeza, R. (2006). Cognitive neuroscience of emotional memory. *Nature Reviews Neuroscience*, 7(1), 54-64.

LaBar, K. S., & Phelps, E. A. (1998). Arousal-mediated memory consolidation: Role of the medial temporal lobe in humans. *Psychological Science*, 9(6), 490-493.

Leary, M. R., Diebels, K. J., Davisson, E. K., Jongman-Sereno, K. P., Isherwood, J. C., Raimi, K. T., ... & Hoyle, R. H. (2017). Cognitive and interpersonal features of intellectual humility. *Personality and Social Psychology Bulletin*, 43(6), 793-813

Levine, L. J., & Burgess, S. L. (1997). Beyond general arousal: Effects of specific emotions on memory. *Social Cognition*, 15(3), 157.

Liu, M., Backer, R. A., Amey, R. C., & Forbes, C. E. (2021). How the brain negotiates divergent executive processing demands: Evidence of network reorganization in fleeting brain states. *NeuroImage*, 245, 118653.

Lynch, M. P. (2016). *The Internet of Us: Knowing more and understanding less in the age of big data*. WW Norton & Company.

Lynch, M. P. (2017). How to see past your own perspective and find truth [Video]. TED Talks. https://www.ted.com/talks/michael_patrick_lynch_how_to_see_past_your_own_perspective_and_find_truth

Lynch, M. P. (2018). Arrogance, truth and public discourse. *Episteme*, *15*(3), 283-296.

Murty, V. P., LaBar, K. S., & Adcock, R. A. (2012). Threat of punishment motivates memory encoding via amygdala, not midbrain, interactions with the medial temporal lobe. *Journal of Neuroscience*, *32*(26), 8969-8976.

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.

OpenAI. (n.d.). Emergent tool use from multi-agent interaction. https://openai.com/research/emergent-tool-use

Orlowski, J. (Director). (2020). *The Social Dilemma* [film]. Exposure Labs.

Payne, J., Jackson, E., Ryan, L., Hoscheidt, S., Jacobs, J., & Nadel, L. (2006). The impact of stress on neutral and emotional aspects of episodic memory. *Memory*, 14(1), 1-16.

Porter, T., Elnakouri, A., Meyers, E. A., Shibayama, T., Jayawickreme, E., & Grossmann, I. (2022). Predictors and consequences of intellectual humility. *Nature Reviews Psychology*, *1*(9), 524-536.

Schneider, S. (2019). *Artificial You: AI and the Future of Your Mind*. Princeton: Princeton University Press.

Schneider, S. (2021, May 21). Is Consciousness a Correlate of Highly Sophisticated Intelligence? [Video]. YouTube. https://www.youtube.com/watch?v=4KZPERshhKY

Schneider, S., & Kilian, K. (2023, April 28). Artificial intelligence needs guardrails and global cooperation. *The Wall Street Journal*. https://www.wsj.com/articles/ai-needs-guardrails-and-global-cooperation-chatbot-megasystem-intelligence-f7be3a3c

Splan, E. D., Magerman, A. B., & Forbes, C. E. (2021). Associations of regional racial attitudes with chronic illness in the United States. *Social Science & Medicine*, 281, 114077.

Steinmetz, K. R. M., Addis, D. R., & Kensinger, E. A. (2010). The effect of arousal on the emotional memory network depends on valence. *Neuroimage*, *53*(1), 318-324.

Turner, C. (2022). Neuromedia, cognitive offloading, and intellectual perseverance. *Synthese*, *200*(2), 66.

Whitcomb, D., Battaly, H., Baehr, J., & Howard-Snyder, D. (2017). Intellectual Humility. *Philosophy and Phenomenological Research*, *94*(3), 509-539.

Yudkowsky, E. (2023, March 29). Pausing AI Developments Isn't Enough. We Need to Shut it All Down. *Time*. https://time.com/6266923/ai-eliezer-yudkowsky-open-letter-not-enough/

Yudkowsky, E. (2023, April 20). Is Artificial General Intelligence too Dangerous to Build? [Video]. YouTube. https://www.youtube.com/watch?v=3_YX6AgxxYw

Zachry, C. E., Phan, L. V., Blackie, L. E., & Jayawickreme, E. (2018). Situation-based contingencies underlying wisdom-content manifestations: Examining intellectual humility in daily life. *The Journals of Gerontology*: Series B, 73(8), 1404-1415.

Zald, D. H., Lee, J. T., Fluegel, K. W., & Pardo, J. V. (1998). Aversive gustatory stimulation activates limbic circuits in humans. *Brain: a journal of neurology*, *121*(6), 1143-1154.

Battaly, H. (2016). Developing virtue and rehabilitating vice: Worries about self-cultivation and self-reform. *Journal of Moral Education*, *45*(2), 207-222.

——. (2017). Intellectual perseverance. *Journal of Moral Philosophy*, *14*(6), 669-697.

——. (2018). Can closed-mindedness be an intellectual virtue?. *Royal Institute of Philosophy Supplements*, *84*, 23-45.

Brady, M. S. (2018). The role of emotion in intellectual virtue. In *The Routledge handbook of virtue epistemology* (pp. 47-57). Routledge.

Chalmers, D. J. (2019). The Virtual as the Digital. *Disputatio* 11 (55):453-486.

——. (2022). *Reality+: Virtual worlds and the problems of philosophy*. Penguin UK.

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

Fairweather, A., & Montemayor, C. (2014). Inferential abilities and common epistemic goods. In *Virtue epistemology naturalized* (pp. 123-139). Springer, Cham.

——. (2017). *Knowledge, dexterity, and attention: A theory of epistemic agency*. Cambridge University Press.

Fricker, M. (2007). *Epistemic injustice: Power and the ethics of knowing*. Oxford University Press.

Frischmann, B., & Selinger, E. (2018). *Re-engineering humanity*. Cambridge University Press.

Gunn, H. K. (2021). Filter bubbles, echo chambers, online communities. In *The Routledge Handbook of Political Epistemology* (pp. 192-202). Routledge.

Gunn, H. K., & Lynch, M. P. (2018). Googling. In *The Routledge Handbook of Applied Epistemology* (pp. 41-53). Routledge.

——. (2021). The internet and epistemic agency. *Applied Epistemology*, 389.

Kerner, C., & Risse, M. (2021). Beyond porn and discreditation: Epistemic promises and perils of deepfake technology in digital lifeworlds. *Moral Philosophy and Politics*, *8*(1), 81-108.

Klenk, M., & Hancock, J. (2019). Autonomy and online manipulation. *Internet Policy Review*, *1*.

Lynch, M. P. (2014). Neuromedia, Extended knowledge and Understanding. *Philosophical Issues*, *24*(1), 299-313.

——. (2016). *The internet of us: Knowing more and understanding less in the age of big data*. WW Norton & Company.

——. (2018). Arrogance, truth and public discourse. *Episteme*, *15*(3), 283-296.

Malone, T. W. (2018). *Superminds: The surprising power of people and computers thinking together*. Little, Brown Spark.

Marin, L., & Roeser, S. (2020). Emotions and digital well-being: The rationalistic bias of social media design in online deliberations. In *Ethics of Digital Well-Being* (pp. 139-150). Springer.

Origgi, G., & Ciranna, S. (2017). Epistemic injustice: The case of digital environments. In *The Routledge Handbook of Epistemic Injustice* (pp. 303-312). Routledge.

OpenAI. Introducing chatgpt, 2022. URL https://openai.com/blog/chatgpt.

OpenAI. GPT-4 technical report, 2023.

Perrine, T., & Timpe, K. (2014). Envy and its Discontents. In *Virtues and Their Vices*, 225-244. Oxford University Press.

Risse, M. (2019). Human rights and artificial intelligence: An urgently needed agenda. *Hum. Rts. Q.*, *41*, 1.

Schneider, S. (2019). *Artificial You: AI and the Future of the Mind*. Princeton University Press.

Schwengerer, L. (2022). Promoting Vices: Designing the Web for Manipulation. In *The Philosophy of Online Manipulation* (pp. 292-310). Routledge.

Srinivasan, A. (2018). The Aptness of Anger. *Journal of Political Philosophy*, *26*(2), 123-144.

Steinert, S., Marin, L., & Roeser, S. (2022). Feeling and thinking on social media: emotions, affective scaffolding, and critical thinking. *Inquiry*, 1-28.

Theiner, G. (2013). Onwards and upwards with the extended mind: From individual to collective epistemic action. *Developing Scaffolds in Evolution, Culture, and Cognition*, *17*, 191.

——. (2014). Varieties of group cognition. In *The Routledge Handbook of Embodied Cognition* (pp. 365-376). Routledge.

Turner, C. (2022). Neuromedia, cognitive offloading, and intellectual perseverance. *Synthese*, *200*(1), 1-26.

Van Slyke, J. A. (2014). Moral psychology, neuroscience, and virtue: From moral judgment to moral character. *Virtues and their vices*, 459-480.

Wallach, W. (2015). *A dangerous master: How to keep technology from slipping beyond our control*. Basic Books.

Wallach, W., & Allen, C. (2008). *Moral machines: Teaching robots right from wrong*. Oxford University Press.

Wildman, N., Rietdijk, N., & Archer, A. (2022). Online Affective Manipulation. In *The Philosophy of Online Manipulation* (pp. 311-326). Routledge.

Zuboff, S. (2019). *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power.* Profile books.

Forbes, C. E., Amey, R., Magerman, A. B., Duran, K., & Liu, M. (2018). Stereotype-based stressors facilitate emotional memory neural network connectivity and encoding of negative information to degrade math self-perceptions among women. Social cognitive and affective neuroscience, 13(7), 719-740.

Hamann, S. (2001). Cognitive and neural mechanisms of emotional memory. Trends in Cognitive Sciences, 5(9), 394-400.

Kahneman, D. (2011). *Thinking, Fast and Slow*. Macmillan.

LaBar, K. S., & Cabeza, R. (2006). Cognitive neuroscience of emotional memory. Nature Reviews Neuroscience, 7(1), 54-64.

LaBar, K. S., & Phelps, E. A. (1998). Arousal-mediated memory consolidation: Role of the medial temporal lobe in humans. Psychological Science, 9(6), 490-493.

Levine, L. J., & Burgess, S. L. (1997). Beyond general arousal: Effects of specific emotions on memory. Social Cognition, 15(3), 157.

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.

Payne, J., Jackson, E., Ryan, L., Hoscheidt, S., Jacobs, J., & Nadel, L. (2006). The impact of stress on neutral and emotional aspects of episodic memory. Memory, 14(1), 1-16.

Splan, E. D., Magerman, A. B., & Forbes, C. E. (2021). Associations of regional racial attitudes with chronic illness in the United States. Social Science & Medicine, 281, 114077.

Gemini: a family of highly capable multimodal models. Preprint at https://arxiv.org/abs/2312.11805 (2023).

Andrews, N. P., Yogeeswaran, K., Wang, M. J., Nash, K., Hawi, D. R., & Sibley, C. G. (2020). Is social media use changing who we are? Examining the bidirectional relationship between personality and social media use. *Cyberpsychology, Behavior, and Social Networking*, *23*(11), 752-760.

Boer, M., Stevens, G., Finkenauer, C., & van den Eijnden, R. (2020). Attention deficit hyperactivity disorder-symptoms, social media use intensity, and social media use problems in adolescents: Investigating directionality. *Child Development*, *91*(4), e853-e865.

Lanier, J. (2018). *Ten Arguments for Deleting Your Social Media Accounts Right Now*. Random House.

Orlowski, J. (2020). *The Social Dilemma* [film]. Exposure Labs.

Sharifian, N., & Zahodne, L. B. (2020). Social media bytes: Daily associations between social media use and everyday memory failures across the adult life span. *The Journals of Gerontology: Series B*, *75*(3), 540-548.

Ward, A. F., Duke, K., Gneezy, A., & Bos, M. W. (2017). Brain drain: The mere presence of one's own smartphone reduces available cognitive capacity. *Journal of the Association for Consumer Research*, *2*(2), 140-154.