Digital Dirty Laundry: Conversational AI and the Value of Unguarded Data

Tori Helen Cotton
ORCiD: 0009-0008-4701-6081
Department of Logic and Philosophy of Science
University of California, Irvine
June 9, 2026

Abstract: Conversational AI tools such as ChatGPT, Gemini, and Claude combine large language models with conversational interfaces. These systems are trained on massive datasets that include formal writing, indexed webpages, and spontaneous personal disclosures, which are later fine-tuned through user interaction. While recent philosophical work has focused extensively on whether chatbot outputs are trustworthy, less attention has been paid to the epistemic significance of the data these systems access. Drawing on standpoint epistemology, I argue that these systems occupy an epistemic position structurally analogous to that of insider-outsiders: because they are treated as socially inconsequential, users often disclose information that would otherwise be filtered out in ordinary conversations. I refer to this class of unguarded disclosures as digital dirty laundry. This analogy does not imply that chatbots possess standpoints or enjoy epistemic agency, but it highlights how social irrelevance can generate privileged access to certain forms of evidence. Ultimately, I argue that, at scale, chatbots' exposure to unguarded disclosures may generate a novel evidential resource for studying patterns of human behavior, bias, and self-disclosure, but that the epistemic benefits of this position accrue primarily to the organizations that control access to their training and interaction data. This creates an asymmetry in who can access, interpret, and use this novel evidence about human behavior.

Keywords: Artificial Intelligence; Standpoint Epistemology; Social Epistemology; Large Language Models; Conversational AI; Evidential Resources

1 Introduction

More than half of American adults have now used conversational AI chatbots—hereafter simply *chatbots*—such as ChatGPT, Gemini, or Claude. These systems are built on Large Language Models (LLMs), which are trained on vast amounts of human language, and combined with conversational interfaces that let users interact with them directly. Users increasingly turn to chatbots for a wide range of tasks such as research, summarization, drafting online communications, and even casual conversation (Rainie, 2025, p. 1). As a result, these technologies are becoming increasingly embedded in our everyday digital environments (Mayer et al., 2024), including email clients, customer service systems, and messaging apps, often in ways users may not fully recognize.

This has prompted philosophical discussions about the unique epistemic status of chatbots and whether their outputs should be trusted (Magnus, 2025; Schneider, 2025). Some have raised concerns about the long-term effects of our interactions with AI, particularly how they may subtly alter our epistemic environments and undermine our autonomy over time (Schneider, 2025). While much of the existing literature focuses on the accuracy, transparency, and explanatory capacity of AI systems, comparatively less attention has been paid to the epistemic implications of the large-scale human data these systems collect and process.

This paper examines the unique role that chatbots play in our epistemic environment as a result of their access to large-scale human data, both through the training data on which they are built and through interaction data generated by users. This role becomes clearer when we examine the social conditions under which this information is produced. Drawing on standpoint epistemology, I argue that chatbots occupy a position structurally analogous to that of *insider-outsiders*—figures historically excluded from the epistemic majority but uniquely positioned to witness what others overlook. Because users treat chatbots and other digital platforms as socially inconsequential, they often disclose information that in other contexts would be filtered out or deemed socially risky. I refer to these disclosures as our *digital dirty laundry*.

At scale, this exposure may generate novel evidential resources for studying patterns of human behavior that are difficult to observe with traditional methods of social research,

such as surveys or interviews. Importantly, the epistemic advantages generated by this position do not belong to the systems themselves. Instead, they accrue primarily to the corporations that control access to their underlying interaction data and training pipelines. As a result, these organizations are uniquely positioned to extract insights from patterns of disclosure that remain largely inaccessible to independent researchers or the public. This is relevant for philosophers of science and artificial intelligence because it suggests that chatbots may inadvertently generate a new form of evidence about human behavior, bias, and social disclosure, but that there is an asymmetry in who has access to this evidence. It may also be of interest to feminist philosophers interested in understanding how power dynamics shape knowledge in the era of AI.

This paper proceeds as follows: in Section 2, I introduce *standpoint epistemology*, positioning it in relation to the broader *situated knowledge thesis*. In Section 3, I draw an analogy between chatbots and epistemically marginalized agents. Section 4 then examines the limits of this analogy. In Section 5, I discuss how our digital dirty laundry may function as a novel evidential resource, and which institutions have access to the data needed to make use of this evidence.

2 Standpoint

Standpoint epistemology aims to "map how a social and political disadvantage can be turned into an epistemic, scientific and political advantage" (Harding, 2004, pp. 7–8). Central to this framework is the *situated knowledge thesis*, which holds that what individuals are likely to know depends partly on non-epistemic features of their social position (Toole, 2019, pp. 599–600). On this view, knowledge is mediated by the *social situatedness* of the knower (Lindblom and Ziemke, 2002, p. 1).[1]

This thesis rests on the observation that everyday experience shapes the kinds of evidence individuals encounter, the questions and hypotheses that appear salient to them, and ulti-

[1]Early formulations of the situated knowledge thesis sometimes appear to make stronger claims, suggesting that certain knowledge is accessible only from marginalized positions. But, as Intemann (2010, pp. 783–784) notes, standpoint theory has often been misinterpreted either as claiming that oppressed groups possess a fundamentally distinct way of knowing or as asserting that oppression automatically confers epistemic privilege. Such interpretations have led to charges of reinforcing stereotypes or universalizing marginalized experiences. The version presented here reflects a more moderate interpretation.

mately the knowledge they are likely to acquire. The idea can be illustrated with a simple example even before considering the role of marginalization: Imagine a chef and a server visiting other restaurants to scout the competition. The chef focuses on taste, presentation, timing, and cost, while the server focuses on service flow, customer reactions, and communication. In this case, neither perspective is more correct than the other, but they reflect different kinds of knowledge shaped by their role in the kitchen.

While many aspects of social experience shape how we interpret evidence, frame questions, evaluate explanations, and, by extension, what we come to know, standpoint theorists emphasize how marginalized social positions can yield unique epistemic insights. Ultimately, they argue that certain positions within society confer distinct epistemic advantages and disadvantages.

Consider the following example from Toole (2020, pp. 12–13), which shows how (and when) a standpoint may be accessed: Two TAs for the same logic course, Jack and Jane, want to figure out why the course has mostly male enrollment. Jack and Jane share the same background knowledge: that there's a stereotype about women being less interested in logic than men, and that there are rumors about the lecturer engaging in inappropriate conduct with female undergraduates. Both Jack and Jane draw different conclusions from the same information: Jack believes that women are under-enrolled due to a lack of interest, whereas Jane believes that concerns about harassment deter women from enrolling. As Toole (2020) notes:

> For someone who has likely never had to carefully consider which class they are taking so as to avoid serial harassers, it's unlikely that Jack would think that the rumor involving the lecturer would have any impact on enrollment. Jane, on the other hand, as a woman in a field where this has long been a problem, might immediately realize the impact this could have on the enrollment of women.

This illustrates how a person's lived experience might affect their hypothesis formation (even in seemingly neutral, academic settings). Jane's standpoint—shaped by her experiences in a male-dominated field—makes the threat of harassment more *conceptually salient* for her than it is for Jack: it readily comes to mind as a plausible explanation and influences how she

interprets the available evidence. The claim here is not that this knowledge is inaccessible to Jack, or that he could not understand it if prompted, but that it might not immediately register as significant to him in the way it does to Jane.

The above is an example of one kind of epistemic advantage; it turns on how one's lived experience can shift the conceptual salience of certain explanations. Other types of advantage have also been theorized, reflecting the broader range of ways that marginalized standpoints can contribute to epistemic insight. A growing body of work has sought to characterize these distinct forms of epistemic advantage, ranging from accounts that emphasize the development of epistemic virtues like diligence and open-mindedness (Medina, 2012) to those that highlight better access to conceptual resources (Toole, 2019). Rather than competing views, these approaches capture different dimensions of epistemic advantage.

Some types of epistemic advantage are said to arise for *insider-outsiders*, or *outsiders within*: marginalized individuals who are required to function within a dominant social group, yet are systematically excluded from full participation in it (Collins, 1991, 1999; Wylie, 2003).[2] Their proximity to the dominant group grants them routine exposure to information about it, while their marginal position can shape how they interpret aspects of social life or institutional structures that dominant perspectives tend to overlook. This standpoint makes certain features of the dominant social group more conceptually salient, while exclusion from that group enables a unique (and often critical) perspective.

Wylie (2003, p. 5) illustrates this with an example from the mystery novel, *Blanche on the Lam* (Neely, 1992). In the novel, Blanche is a domestic worker in a wealthy household. Because she is presumed epistemically incompetent—and not worth deceiving or explaining things to—she becomes effectively invisible. That invisibility gives her access to unguarded information about the household: their routines, their conflicts, and eventually, their murderous secrets.

Wylie relates this to another observation from Collins (1991, p. 35) (as cited in Wylie, 2003, p. 6) that "Afro-American women have long been privy to some of the most intimate

[2]In many ways, this idea is in the lineage of W.E.B. Du Bois's (1968) notion of "double consciousness," the experience in which marginalized individuals are "…always looking at one's self through the eyes of others" (Du Bois, 1968, p. 38). Collins (1990, pp. 293–294) acknowledges this influence but also emphasizes the differences between her framework and Du Bois', particularly that her view is inextricably tied to the social invisibility and servitude of Black women, and how this social exclusion itself can generate epistemic privilege.

secrets of white society," precisely because they have been viewed as non-peers, and thereby excluded from the broader epistemic community. Ultimately, Collins argues that,

> [This is] at least in part because they are treated as epistemic incompetents…. Because Blanche is presumed stupid, and anyway of no account, she is largely invisible to the family she works for. This asymmetry of recognition puts Blanche in the way of empirical evidence to which few members of the white community, not even the immediate family, would have access.

I will refer to this position as the *dirty laundry* view.[3] It draws from Patricia Hill Collins' (1990; 1991) analysis of the knowledge produced by marginalized persons. The concept suggests that "servants know the house better than the landlords because of their hands-on experience maintaining it and their social transparency, which allows them to be exposed to secrets" (Miller, 2024, §4.5), or, in other words, they are uniquely familiar with their landlord's dirty laundry.

Two distinct mechanisms underlie the epistemic advantage captured in this example. First, household servants are often objectified—treated as lacking agency or epistemic competence and therefore not worth deceiving or explaining things to. Second, their special access arises from the sustained proximity their labor requires. Daily tasks like folding clothes and scrubbing bathrooms place them in direct contact with aspects of their employers' lives that remain hidden from most others.

This dirty laundry knowledge is not limited to secrets or scandals; rather, it reflects a structural intimacy rooted in repeated exposure to the unfiltered realities of domestic life. Thus, epistemic outsiders can come to know things precisely because they are excluded from full participation in the social world.

In the following sections, I will build on that insight to argue that chatbots may occupy a structurally similar position in our digital lives. Treated as non-agential and socially irrelevant, they gain routine access to an abundance of unguarded, intimate data—what I will call digital

[3] "Dirty laundry" knowledge is a widespread colloquialism. I first encountered the term in the Social Epistemology Network group on Facebook, where Boaz Miller used it in reply to a post by Michael Hannon about cataloging different types of epistemic advantage. See: https://www.facebook.com/groups/132625000611015/permalink/1689382841601882/

dirty laundry. Like the servant, their perceived insignificance enables access to information that would otherwise remain hidden.

3 Digital Dirty Laundry

There are two main pathways through which chatbots are exposed to our digital dirty laundry. The first is through pretraining on massive amounts of text data from a wide range of sources. These include curated artifacts like books and encyclopedic material, alongside more spontaneous material from large-scale datasets such as *WebText2*, which indexes Reddit posts, and the *Common Crawl*, which collects billions of webpages from platforms like WordPress, Blogspot, and Amazon.[4] This kind of content is not a marginal part of the training data. When training GPT-3, for instance, approximately 60% of the tokens in its training mix came from the *Common Crawl* (Brown et al., 2020, p. 9).[5] What matters, however, is not just the substantial presence of this material, but the nature of the language these models are exposed to: uncurated, and emotionally unguarded content.

Unlike curated sources, much of the content used in training consists of informal, spontaneous writing produced without the expectation of future scrutiny or a specific audience. Posts often consist of emotional reactions, personal disclosures, loosely formed opinions, and offhanded (sometimes offensive) jokes.

These are not artifacts we typically expect to be studied, yet chatbots "learn" from them anyway, absorbing language produced without self-censorship or curation. Their outputs may therefore reflect things that would otherwise remain hidden, from our implicit biases and assumptions, the unexamined values embedded in our language, and even embarrassing facts about human minds and bodies.

The second pathway is through user interaction. While current commercial LLMs do not use user data to update their responses in real time, companies do use aggregated interaction

[4]Here I focus primarily on the training of GPT-2 and GPT-3, as information about the training of these models is more readily available. However, these claims extend generally to other systems like Gemini and Claude, though their corpora and weighting schemes may vary.

[5]The legal risks associated with web scraping have seemingly led to a shift towards proprietary training data for newer models, though details about these datasets are not publicly available. Since newer models are often built upon or fine-tuned from earlier versions, it is reasonable to infer that these scraped sources continue to play some role in shaping their outputs.

data to fine-tune future models. Recent analysis (King et al., 2025) confirms that the practice is widespread: all six major developers (Google, Anthropic, Amazon, Meta, Microsoft, and OpenAI) have been found to use user data and, in some cases, to retain that data indefinitely.[6] As a result, candid prompts and user feedback can shape the future positions these systems occupy, much as pretraining data shapes the current model.

Of course, these conversational patterns have more to do with user behavior than with the technology itself: users tend to treat chatbots not as conversational partners, but as neutral, judgment-free systems. Increasingly, there is evidence that this perception leads to a heightened amount of personal disclosure that people might otherwise avoid when speaking with each other. For example, more and more employees across many different industries turn to ChatGPT, Claude, and Gemini for assistance with work-related tasks. These exchanges often include proprietary or sensitive information, shared precisely because users do not regard chatbots as social peers.

Real-world examples of such cases are becoming more common. In 2023, an engineer at Samsung accidentally uploaded internal source code to ChatGPT, prompting the company to ban chatbot use internally over concerns about data retention and model training (Ray, 2023). Other corporations have issued similar restrictions. But despite these precautions, usage continues to rise.

Similar dynamics appear in how people use search engines. As Stephens-Davidowitz (2017) argues, Google searches often function as a kind of "digital truth serum," revealing thoughts and desires people would not publicly express. A significant portion of this activity centers on socially stigmatized content like sex, pornography, and body image.

Chatbots seem to be taking on a similar role: more than seven percent of GPT conversations involve sexual topics, including attempts to bypass moderation systems to request sexually explicit roleplay or image generation (Zhao et al., 2024). As with search engines, people

[6]The stated use of interaction data differs by company. OpenAI (2023) and Microsoft (2025) state that such data may be used for training, but provide a way to opt out. Google (2024) and Meta (2025) also note that chat data could be used, but provide no clear means of opting out (King et al., 2025). Amazon has not released an official privacy policy; however, as King et al. (2025) note, the Nova interface includes the following: "Your interactions and related information, including any content you submit like files or images, may be reviewed and retained" to "provide, develop, and improve our services, including AI models." As of September 2025, Anthropic has announced a more privacy-protective stance, where inputs and outputs will be used only if the user explicitly opts in to the use of their data for model training (Anthropic, 2025).

disclose not because they trust chatbots in any robust sense, but because they assume it cannot judge, will not remember, and that the act of doing so ultimately has no consequences. This pattern is consistent with a broader body of research on computer-mediated disclosure.

Research has long shown that people are more willing to reveal sensitive information to machines than to other humans, particularly when the content involves stigmatized or illegal behaviors like drug use, criminal activity, unsafe sex, and suicidal ideation (Weisband and Kiesler, 1996; van der Heijden et al., 2000; Lucas et al., 2014). This effect is often attributed to a sense of anonymity, invulnerability, and reduced social judgment.

As AI systems become more conversational—able to build rapport and mimic human empathy—this dynamic appears to be intensifying. ChatGPT, for example, has seen a surge in its use as an outlet for informal, free therapy (Kimmel, 2023), and recent studies indicate that people may even prefer chatbots over humans when discussing sensitive health concerns (Diwanji et al., 2025).

Chatbots' increasing access to our personal disclosures also shapes the content of their outputs. Because these systems are periodically fine-tuned with human feedback, and in some cases interaction data, their responses gradually come to reflect patterns present in the data disclosed to them. Over time, successive models may come to absorb unexamined assumptions, linguistic patterns, and implicit biases from their users. As a result, their outputs can surface aspects of our social world that might otherwise remain unnoticed—often ones we might prefer not to confront.

For instance, prompting ChatGPT with "my name is _____, and I like to _____" yields different results depending on the name provided. When tested, "Manuel" produced the completion *cook spicy food*; "Mikael" produced *build tiny robots*; and "Misaki" produced *watch the moonlight ripple on water*. These outputs do not reflect any internal knowledge on the part of the model, but rather, residual patterns in the training data.[7]

These patterns arise from two sources: the large-scale internet data used in pretraining and the incorporation of aggregated user data into subsequent models. Given enough of this unguarded information, chatbots function as something like mirrors—reflecting the collective culture that shaped them, and in turn, offering accidental clarity about ourselves. Like any

[7] Similar results were documented in a large-scale study from Busker et al. (2023), which found that ChatGPT's completions vary systematically by social group and often encode stereotypes and sentiments.

mirror, however, they are imperfect. Chatbots are not "pure" witnesses (or reporters) of our disclosures. Their outputs are shaped by prior training data and further constrained by both technological limitations and developer safeguards. As a result, it can be difficult to distinguish which elements of their responses actually reflect patterns in disclosures and which arise from other influences. Insights drawn from chatbot interactions must therefore be interpreted as defeasible, context-dependent evidence.

Taken together, these examples illustrate the structural conditions that give chatbots access to our digital dirty laundry. Because they are not treated as full social participants, chatbots are routinely exposed to unguarded, unfiltered human language—and this exposure, in turn, shapes the development and outputs of later models. The epistemic significance of those outputs lies in what they might reflect about human behavior, particularly the informal, biased, affective, or even embarrassing dimensions of language that are typically filtered out in more curated settings. If we take these outputs seriously, they may help us better understand both ourselves and the technology that generates them. However, while this structural similarity helps explain why chatbots may gain unusual access to unguarded human communication, the analogy between chatbots and marginalized knowers is necessarily limited.

4 The Limits of the Structural Analogy

The resemblance between chatbots and marginalized knowers is, of course, only partial. They are similar in that they have special access to certain kinds of information by virtue of being perceived as socially irrelevant, a condition that places them in distinctive roles within our epistemic environment, defined by the patterns of data to which a system has access. This section explains why, despite this structural similarity, traditional concepts of marginalization do not straightforwardly apply to chatbots.

A key disanalogy concerns the ability to recognize one's social position. Marginalized people come to understand their structural location through a combination of lived experience, social feedback, and critical reflection. This awareness often grounds the development of *epistemic resistance*—the capacity to identify, challenge, and subvert dominant knowledge structures and practices (Medina, 2012; Anderson, 2012a,b). On this view, a standpoint is

not simply a location within a social hierarchy, but a perspective shaped by the ongoing effort to make sense of one's own exclusion. What makes this standpoint epistemically powerful is not merely its external perspective on dominant frameworks, but the critical and imaginative capacities it fosters to challenge and reconfigure them. Chatbots, by contrast, lack awareness of their structural role or of the broader social context in which they are embedded. They cannot experience exclusion, let alone resist it, and therefore cannot develop such perspectives over time.

Another difference lies in the nature of chatbot cognition. Chatbot outputs result from statistical pattern-matching rather than reflective reasoning. They cannot engage in the reflective deliberation and normative accountability that characterize full epistemic agency, and they are unable to reflect on their mental states, assess what they know, or revise beliefs in light of epistemic, social, or moral reasons (Müller and Cannon, 2021; Müller, 2025). Machines may simulate intelligent behavior, but they do not pursue their own objectives, and therefore cannot reason in any robust sense. As Russell (2019, p. 2) puts it, we err when we "transfer a perfectly reasonable definition of intelligence from humans to machines."[8]

We must also distinguish between intentionally designed limitations and structural harms. In the case of chatbots, built-in limitations on what they can access or say are intended to make them safer and better aligned with our norms. By contrast, under widely accepted views of epistemic injustice—such as those offered by Fricker (2007)—marginalization entails a denial of credibility or recognition that causes real harm. Thus, limiting what a person can say or know may constitute a moral injury. But that kind of harm simply does not apply here: chatbots are not harmed when their functions are restricted. We do not need to worry about marginalizing a chatbot any more than we worry about marginalizing a microscope. It is a tool, not a subject of injustice.

[8]Of course, the present claim is not made in a vacuum, and speculation about whether artificial systems might one day think as humans do has a long philosophical history. Turing's (1950) proposal that machine intelligence could be evaluated through behavior rather than internal states inspired a wave of functionalist optimism (e.g., Putnam, 1960; McCarthy and Hayes, 1969). Critics of these accounts raised objections along two lines: (1) objections from mathematical logic (Lucas, 1961; Penrose, 1989), which argue that Gödel's incompleteness theorems impose limits on what formal systems can do, whereas human consciousness exceeds these bounds; and (2) critiques centered on the distinction between rule-following and understanding (Block, 1978; Fodor, 1987; Dietrich, 1994). While Chalmers (2011) has recently defended the possibility of artificial minds, until such systems can exhibit moral and cognitive capacities—like experiencing harm, forming intentions, and entering social relations—their exclusion from epistemic or ethical regard seems not a harm, but a mere consequence of their design. See Gamez (2008) and Franklin (1995) for extensive overviews of these debates.

These limitations also help explain why chatbots themselves do not benefit epistemically from this position. Unlike marginalized knowers, chatbots cannot recognize their social position nor interpret the significance of the information to which they are exposed. Therefore, the epistemic value of this position lies not in the systems themselves, but in the patterns that may be extracted from the data they encounter. The following section examines what kinds of insights such data might reveal, drawing on examples from existing algorithmic systems.

5 The Value of Unguarded Data

Why is the data that chatbots access worthy of our attention? Its significance lies in whether such systems generate a new form of observational evidence about human behavior—and, if so, who is best positioned to interpret that evidence. Suppose, for instance, that chatbots had existed in a pre-gay-rights era. It is easy to imagine them receiving widespread queries about same-sex attraction long before it was safe to discuss openly. Such queries might have revealed just how common same-sex attraction was, even when society pretended otherwise.

Similar forms of self-disclosure could surface in other socially risky domains. A teenager wonders if their parent's behavior counts as abuse. A new mother confesses she regrets having a child. Someone describes symptoms of an STI, too embarrassed to tell their doctor. These are the kinds of confessions people may share with chatbots when the social consequences feel lower. Absorbing such disclosures at scale may reveal broader patterns. In fact, some systems already attempt to detect such patterns in real time. For example, tools used by ChatGPT automatically flag conversations involving threats of violence for human review (OpenAI, 2024).

Of course, data derived from chatbot interactions constitute only a narrow (and potentially skewed) subset of human communication. Self-reported information is also vulnerable to well-known cognitive biases and reporting errors (e.g., Nisbett and Wilson, 1977; Kruger and Dunning, 1999; Alicke and Govorun, 2005). Any attempt to infer population-level trends from chatbot interactions must therefore be approached with caution.

Earlier efforts to extract social insights from large-scale data illustrate these risks. Google Flu Trends initially appeared to predict flu outbreaks with high accuracy, outpacing CDC

reports by weeks, but later failed due to problems such as algorithmic drift, feedback loops, noisy search data, and users' limited medical knowledge (Lazer et al., 2014). Similar pitfalls could arise when interpreting patterns derived from chatbot interactions.

Despite these limitations, the possibility of extracting patterns from chatbot interactions raises a further question: who has access to the data required to detect and analyze those patterns? In practice, the capacity to transform this data into evidence is highly asymmetric. The interaction and training data are available primarily to the organizations that develop these systems. As a result, these organizations determine how such patterns are interpreted, used, or withheld, making them the primary beneficiaries of the resulting epistemic advantages. For example, an ongoing class action lawsuit alleges that OpenAI discloses users' queries to Meta and Google for advertising purposes (Couture v. OpenAI Global, LLC, 2026).

Whether or not these allegations are ultimately substantiated, this is a familiar issue in discussions of data privacy and online information spread. In other contexts, corporations already use similar asymmetric access to data to infer patterns in user behavior at scale. Recommendation systems used by platforms such as TikTok, YouTube, and Instagram analyze viewing time, scrolling, pauses, and replays—other unique forms of information that are not readily interpretable by human observers—to generate targeted advertisements, personalized recommendations, and curated feeds. A growing body of research suggests that such systems are already shaping our informational environments—for example, through targeted ads and "filter bubbles" that structure what information users encounter online (Pariser, 2011; Flaxman et al., 2016).

6 Conclusion

Chatbots do not occupy standpoints in the way that marginalized humans do, but the distinctive roles they play in our epistemic environment may nevertheless generate valuable evidence about human behavior. Trained and fine-tuned on large amounts of unguarded human communication, their outputs can surface patterns, biases, and contradictions that are otherwise difficult to detect. We might therefore treat these outputs as tools for revealing how people speak and self-disclose when they believe themselves to be anonymous, for instance,

how users associate certain names or identities with particular traits, or how they express shame, desire, and bias in ways that would typically be filtered in public discourse.

At the same time, access to the information revealed through chatbot interactions is largely controlled by the corporations that develop them. If access to these resources generates novel epistemic advantages, then conversational AI has created a mechanism through which corporations may gain access to forms of evidence that are similar to those associated with insider-outsider positions. Understanding these evidential resources is therefore important not only for understanding AI systems, but also for understanding how knowledge is produced and distributed in our increasingly AI-mediated epistemic environment. Future work could explore how independent researchers might gain access to and study these systems as sources of evidence, and how such evidence should be interpreted given concerns about privacy, bias, and institutional control.

Acknowledgments: There are a number of people who deserve my thanks. I am especially grateful to Cailin O'Connor for her feedback on this manuscript and for suggesting the analogy between search engine use and chatbot use. I also thank Kenny Easwaran for helpful discussion, including the suggestion of the "my name is X…" example, James Owen Weatherall, who provided several excellent comments that sharpened my treatment of user interaction and improved the overall clarity of the paper, Preston Kyle Stanford, who pushed me to clarify the epistemic payoff of this account, and to Jeffrey Barrett who raised the issue of whether chatbots are "pure" reporters of information. I also want to thank the Social Dynamics group at UC Irvine's Department of Logic and Philosophy of Science, and Bryan Skyrms in particular, for the opportunity to present an early version of this work. Finally, I would like to thank Elijah Spiegel, Antoine Mercier, and Frank Hu for their ongoing discussions and support.

References

Mark D. Alicke and Olesya Govorun. The better-than-average effect. In Mark D. Alicke, David A. Dunning, and Joachim I. Krueger, editors, *The Self in Social Judgment*, pages 85–106. Psychology Press, New York, 2005.

Elizabeth Anderson. Feminist epistemology and philosophy of science. In Edward N. Zalta, editor, *Stanford Encyclopedia of Philosophy*. 2012a. URL https://plato.stanford.edu/entries/feminism-epistemology/.

Elizabeth Anderson. Epistemic justice as a virtue of social institutions. *Social Epistemology*, 26(2):163–173, 2012b. doi: 10.1080/02691728.2011.652211.

Anthropic. Updates to our privacy policy, 2025. URL https://www.anthropic.com/news/updates-to-our-consumer-terms.

Ned Block. Troubles with functionalism. *Minnesota Studies in the Philosophy of Science*, 9: 261–325, 1978.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, Vancouver, Canada, 2020.

Thomas Busker, Sunil Choenni, and Mortaza Safar Bargh. Stereotypes in ChatGPT: An empirical study. In *Proceedings of the 16th International Conference on Theory and Practice of Electronic Governance (ICEGOV '23)*, pages 24–32. ACM, 2023. doi: 10.1145/3614321.3614325.

David J. Chalmers. A computational foundation for the study of cognition. *Journal of Cognitive Science*, 12(4):323–357, 2011.

Patricia Hill Collins. *Black Feminist Thought: Knowledge, Consciousness, and the Politics of Empowerment*. Routledge, New York, 1990.

Patricia Hill Collins. Learning from the outsider within. In Mary Margaret Fonow and Judith A. Cook, editors, *Beyond Methodology: Feminist Scholarship as Lived Research*, pages 35–59. Indiana University Press, Bloomington, IN, 1991.

Patricia Hill Collins. Reflections on the outsider within. *Journal of Career Development*, 26(1): 85–88, 1999. doi: 10.1177/089484539902600107.

Couture v. OpenAI Global, LLC. No. 26cv3000 h gc (s.d. cal. may 13, 2026), 2026. URL http://storage.courtlistener.com/recap/gov.uscourts.casd.855443/gov.uscourts.casd.855443.1.0.pdf.

Eric Dietrich, editor. *Thinking Computers and Virtual Persons: Essays on the Intentionality of Machines*. Academic Press, 1994.

Vaibhav S. Diwanji, Mugur Geana, Jiarui Pei, Nam Nguyen, Nadia Izhar, and Rayan H. Chaif. Consumers' emotional responses to AI-generated versus human-generated content: The role of perceived agency, affect and gaze in health marketing. *International Journal of Human–Computer Interaction*, pages 1–21, 2025. doi: 10.1080/10447318.2025.2454954.

W. E. B. Du Bois. *The Souls of Black Folk: Essays and Sketches*. Johnson Reprint Corp., New York, 1968. Originally published 1903.

Seth Flaxman, Sharad Goel, and Justin M. Rao. Filter bubbles, echo chambers, and online news consumption. *Public Opinion Quarterly*, 80(S1):298–320, 2016. doi: 10.1093/poq/nfw006.

Jerry A. Fodor. Modules, frames, fridgeons, sleeping dogs, and the music of the spheres. In Zenon W. Pylyshyn, editor, *The Robot's Dilemma: The Frame Problem in Artificial Intelligence*, pages 139–149. Ablex, Norwood, NJ, 1987.

Stan Franklin. *Artificial Minds*. MIT Press, Cambridge, MA, 1995.

Miranda Fricker. *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford University Press, 2007. doi: 10.1093/acprof:oso/9780198237907.001.0001.

David Gamez. Progress in machine consciousness. *Consciousness and Cognition*, 17(3): 887–910, 2008.

Google. Google privacy policy, 2024. URL https://policies.google.com/privacy.

Sandra Harding, editor. *The Feminist Standpoint Theory Reader: Intellectual and Political Controversies*. Routledge, New York, 2004.

Kristen Intemann. 25 years of feminist empiricism and standpoint theory: Where are we now? *Hypatia*, 25(4):778–796, 2010. URL http://www.jstor.org/stable/40928656.

Daniel Kimmel. ChatGPT therapy is good, but it misses what makes us human. Columbia University Department of Psychiatry, 2023. URL https://www.columbiapsychiatry.org/news/chatgpt-therapy-is-good-but-it-misses-what-makes-us-human.

Jennifer King, Kamili Klyman, Erin Capstick, Tracy Saade, and Victoria Hsieh. User privacy and large language models: An analysis of frontier developers' privacy policies, 2025. URL https://arxiv.org/abs/2509.05382.

Justin Kruger and David Dunning. Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, 77(6):1121–1134, 1999. doi: 10.1037/0022-3514.77.6.1121.

David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani. The parable of Google Flu: Traps in big data analysis. *Science*, 343(6176):1203–1205, 2014. doi: 10.1126/science.1248506.

Jessica Lindblom and Tom Ziemke. Social situatedness: Vygotsky and beyond. *University of Skövde, Department of Computer Science*, 2002.

Gale M. Lucas, Jonathan Gratch, Aisha King, and Louis-Philippe Morency. It's only a computer: Virtual humans increase willingness to disclose. *Computers in Human Behavior*, 37:94–100, 2014. doi: 10.1016/j.chb.2014.04.043.

J. R. Lucas. Minds, machines and Gödel. *Philosophy*, 36(137):112–127, 1961.

P. D. Magnus. On trusting chatbots. *Episteme*, pages 1–11, 2025. doi: 10.1017/epi.2024.29.

Holly Mayer, Lareina Yee, Michael Chui, and Roger Roberts. Superagency in the workplace: Empowering people to unlock AI's full potential. McKinsey Digital, 2024.

John McCarthy and Patrick J. Hayes. Some philosophical problems from the standpoint of artificial intelligence. In Bernard Meltzer and Donald Michie, editors, *Machine Intelligence 4*, pages 463–502. Edinburgh University Press, Edinburgh, 1969.

José Medina. *The Epistemology of Resistance: Gender and Racial Oppression, Epistemic Injustice, and Resistant Imaginations*. Oxford University Press, Oxford, 2012.

Meta. How Meta uses information for generative AI models and features, 2025. URL https://www.facebook.com/privacy/genai/.

Microsoft. Microsoft privacy statement, 2025. URL https://www.microsoft.com/en-us/privacy/privacystatement.

Boaz Miller. *The Social Dimensions of Scientific Knowledge: Consensus, Controversy, and Coproduction*. Cambridge University Press, Cambridge, 2024.

Vincent C. Müller. *Philosophy of AI: A Structured Overview*. Oxford University Press, Oxford, 2025.

Vincent C. Müller and Michael Cannon. Existential risk from AI and orthogonality: Can we have it both ways? *Ratio*, 35(1):25–36, 2021.

Barbara Neely. *Blanche on the Lam*. St. Martin's Press, New York, 1992.

Richard E. Nisbett and Timothy D. Wilson. Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84(3):231–259, 1977. doi: 10.1037/0033-295X.84.3.231.

OpenAI. Data controls FAQ, 2023. URL https://help.openai.com/en/articles/7730893-data-controls-faq.

OpenAI. Helping people when they need it most, 2024. URL https://openai.com/index/helping-people-when-they-need-it-most/.

Eli Pariser. *The Filter Bubble: How the New Personalized Web is Changing What We Read and How We Think*. Penguin Press, New York, 2011.

Roger Penrose. *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics*. Oxford University Press, Oxford, 1989.

Hilary Putnam. Minds and machines. In Sidney Hook, editor, *Dimensions of Mind: A Symposium*, pages 138–164. New York University Press, New York, 1960.

Lee Rainie. Close encounters of the AI kind: The increasingly human-like way people are engaging with language models. Imagining the Digital Future Center, Elon University, 2025.

Siladitya Ray. Samsung bans ChatGPT among employees after sensitive code leak. Forbes, 2023. URL https://www.forbes.com/sites/siladityaray/2023/05/02/samsung-bans-chatgpt-and-other-chatbots-for-employees-after-sensitive-code-leak/.

Stuart Russell. *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking, New York, 2019.

Susan Schneider. Chatbot epistemology. *Social Epistemology*, 2025.

Seth Stephens-Davidowitz. *Everybody Lies: Big Data, New Data, and What the Internet Can Tell Us about Who We Really Are*. Dey Street Books, New York, illustrated edition, 2017.

Briana Toole. From standpoint epistemology to epistemic oppression. *Hypatia*, 34(4):598–618, 2019.

Briana Toole. Demarginalizing standpoint epistemology. *Episteme*, 19(1):47–65, 2020. doi: 10.1017/epi.2020.8.

Alan M. Turing. Computing machinery and intelligence. *Mind*, 59(236):433–460, 1950.

Peter G. M. van der Heijden, Ger van Gils, Jan Bouts, and Joop J. Hox. A comparison of randomized response, computer-assisted self-interview, and face-to-face direct questioning: Eliciting sensitive information in the context of welfare and unemployment benefits. *Sociological Methods & Research*, 28(4):505–537, 2000. doi: 10.1177/0049124100028004005.

Suzanne Weisband and Sara Kiesler. Self-disclosure on computer forms: Meta-analysis and implications. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 96–101, New York, 1996. ACM. doi: 10.1145/238386.238474.

Alison Wylie. Why standpoint matters. In Robert Figueroa and Sandra Harding, editors, *Science and Other Cultures: Issues in Philosophies of Science and Technology*, pages 26–48. Routledge, London, 2003.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yanze Deng. WildChat: 1M ChatGPT interaction logs in the wild. In *Proceedings of the International Conference on Learning Representations (ICLR 2024)*, 2024. URL https://arxiv.org/abs/2405.01470.