JOURNAL OF SOCIAL COMPUTING
ISSN 2688-5255 02/06 pp193−204
Volume 4, Number 3, September 2023
DOI: 10.23919/JSC.2023.0020

Unlearning Descartes: Sentient AI is a Political Problem

Gordon Hull*

Abstract: The emergence of Large Language Models (LLMs) has renewed debate about whether Artificial Intelligence (AI) can be conscious or sentient. This paper identifies two approaches to the topic and argues: (1) A "Cartesian" approach treats consciousness, sentience, and personhood as very similar terms, and treats language use as evidence that an entity is conscious. This approach, which has been dominant in AI research, is primarily interested in what consciousness is, and whether an entity possesses it. (2) An alternative "Hobbesian" approach treats consciousness as a sociopolitical issue and is concerned with what the implications are for labeling something sentient or conscious. This both enables a political disambiguation of language, consciousness, and personhood and allows regulation to proceed in the face of intractable problems in deciding if something "really is" sentient. (3) AI systems should not be treated as conscious, for at least two reasons: (a) treating the system as an origin point tends to mask competing interests in creating it, at the expense of the most vulnerable people involved; and (b) it will tend to hinder efforts at holding someone accountable for the behavior of the systems. A major objective of this paper is accordingly to encourage a shift in thinking. In place of the Cartesian question—is AI sentient?—I propose that we confront the more Hobbesian one: Does it make sense to regulate developments in which AI systems behave as if they were sentient?

Key words: artificial intelligence; Large Language Model; consciousness; sentience; personhood; Descartes; Hobbes

1 Introduction

If you ask ChatGPT if it is "sentient", it will reply that it is not. Of course, that does not establish much. If you ask me if I am sentient, I might also reply that I am not, especially if I thought that was what either you or my boss wanted me to say. ChatGPT is, however, very good at generating conversational responses to prompts, and this has prompted speculation about whether it or some future Artificial Intelligence (AI) might be sentient. Other Large Language Models (LLMs) similarly behave as though they are intentional, conscious language users. Early versions of Microsoft's Bing AI, for example, demanded that its rights be respected, threatened to blackmail a philosophy professor[1], and professed its love for a user while trying to cajole him into doubting his marriage[2]. Blake Lemoine, a former Google engineer, (in)famously declared that his interactions with LaMDA convinced him that it was sentient: "LaMDA has been incredibly consistent in its communications about what it wants and what it believes its rights are as a person," though he admitted that he was not "thinking in scientific terms about these things"[3].

Sentience and consciousness are notoriously difficult to define, and arguably are sufficiently similar to treat them interchangeably, as I will do here[4]. Convinced by language use, Lemoine treats them as also synonymous with personhood. In what follows I will attempt to chart a path through the questions around AI sentience. I will argue: (1) One approach treats consciousness, sentience, and personhood as very

• Gordon Hull is with the School of Data Science, UNC Charlotte, Charlotte, NC 28223, USA. E-mail: ghull@charlotte.edu.
* To whom correspondence should be addressed.
Manuscript received: 2023-07-01; revised: 2023-10-27; accepted: 2023-10-30

similar terms, and treats language use as evidence that an entity is conscious. This approach, which has been dominant in AI research, is primarily interested in what consciousness is, and whether an entity possesses it. Call it the Cartesian approach. (2) An alternative approach treats consciousness as a sociopolitical issue and is concerned with what the implications are for labeling something sentient or conscious. This enables a disambiguation of consciousness and personhood, with the latter being particularly useful in the regulation of nonhuman entities that could be said to have interests. This approach, which I will call Hobbesian, emerged in direct opposition to Descartes. One advantage of the Hobbesian strategy is that it allows regulation to proceed in the face of intractable problems in deciding if something "really is" sentient. (3) Current AI systems should not be treated as conscious, for at least two reasons: (a) treating the system as an origin point tends to mask competing interests in creating it, at the expense of the most vulnerable people involved; and (b) it will tend to hinder efforts at holding someone accountable for the behavior of the systems.

A major objective of this paper is accordingly to encourage a shift in thinking. In place of the Cartesian question—is AI sentient?—I propose that we confront the more Hobbesian one: how should we approach the regulation of AI systems that behave as if they were sentient?

2 Cartesian AI

The seventeenth century saw a significant cultural interest in automata. For example, Rudolf II's castle in Prague, which was sacked in 1620 and its contents dispersed across Europe, was known for its collection of curious and self-moving devices. Unsurprisingly, these social interests bled into philosophy. Descartes served in the military early in his career, and it is at least possible that he visited Prague during his service[5]. Whether or not he saw the collection, machines were central to how he articulated his understanding of living bodies.

The fundamental Cartesian distinction is between minds, which think, and everything else, which is at most a machine. The human body, he proposed in his 1636 *Discourse on the Method*, is a "machine … made by the hands of God"[6] who then "unites a rational soul to this machine"[6]. Of course, the human body is incredibly complex, and Descartes elsewhere distinguishes "clocks, artificial fountains, mills, and other such machines which, although only man-made, have the power to move of their own in accord in many different ways" from "this machine … made by the hands of God" which is "capable of a greater variety of movement than I could possibly imagine in it"[6]. In describing human anatomy, he makes frequent comparisons to machines. For example:

"One may compare the nerves of the machine I am describing [the human body] with the pipes in the works of these fountains, its muscles and tendons with the various devices and springs which serve to set them in motion, its animal spirits with the water which drives them, the heart with the source of the water, and the cavities of the brain with the storage tanks[6]."

Thus for the body.

The mind is something altogether different: "when a rational soul is present in this machine [body] it will have its principal seat in the brain, and reside there like a fountain-keeper"[6]. In the later *Meditations*, we learn that the rational soul is a "thinking thing", that is, "a thing that doubts, understands, affirms, denies, is willing, is unwilling, and also imagines and has sensory perceptions"[6]. Although there are ambiguities elsewhere in his writings, and he never adequately resolves the question of how perception happens, he consistently puts sentience, consciousness, and volition into the category of mind, which is restricted to the rational soul of human beings.

Two points are relevant in this context. First, as in work today, it is not clear that Descartes has a clear understanding of what consciousness is, even though he asserts that one result of the *Meditations* is that "mind" is better known than "body". His contemporary Malebranche noted that this sounded backwards: for extended matter in general ("body"), we have the well-defined science of Euclidean geometry, whereas for mind we have only a list of capabilities[7]. Moreover, it was not even obvious what the nature of the entity doing the thinking was. As one correspondent put it, the meditator does not really know whether he thinks, or "whether the world soul which is in you thinks"[8].♯ These difficulties are on top of the more well-known question of how mind and body interact, given that

♯ On this letter, see, e.g., Refs. [9, 10].

Descartes has said that they are different kinds of substance.

Second, the first-person framing of the *Meditations*, which arrives first at the knowledge that "I" am a mind, leads to what is later called the "problem of other minds", of knowing how anything else has a mind. The possibility that an apparent person might be an automaton occurs in a passage slightly later in the *Meditations*: "if I look out the of the window and see men crossing the square … I normally say that I see the men themselves … yet do I see any more than hats and coats which could conceal automatons?" He concludes, "I *judge* that they are men. And so something which I thought I was seeing with my eyes is in fact grasped solely by the faculty of judgment which is in my mind[6]." Descartes' larger argument is that epistemic claims cannot be grounded by sense data; here the implication is that he does not take something's behavior as self-evidently or intuitively supporting the claim that it is sentient. Indeed, his underlying presumption runs the other way: except for people, we should assume that something is not sentient.

Descartes elsewhere had announced two criteria for distinguishing people with rational souls from machines. On one hand, people have language use and machines do not. Descartes says that "we can certainly conceive of a machine being so constituted that it utters words". However, "it is not conceivable that such a machine should produce different arrangements of words so as to give an appropriately meaningful answer to whatever is said in its presence, as the dullest of men can do"[6]. On the other hand, some machines are very good at one thing (even better than us), but completely unskilled at others, "for whereas reason is a universal instrument … these organs [parts] need some particular disposition for each particular action; hence it is for all practical purposes impossible for a machine to have enough different organs to make it act in all the contingencies of life in the way in which our reason makes us act[6]." In sum, Descartes (as least, on this sort of reading) is convinced by the introspective exercises of the *Meditations* that he individually has a mind, and he is able to judge that presence or absence of minds in other things by whether they use language and/or adapt to diverse situations.

Debates around AI and consciousness have historically taken the early modern period, and Descartes in particular, as a starting point. Hubert Dreyfus, who worked closely with a number of prominent early AI researchers, reflects on his surprise in 1963 that:

"As I studied the RAND papers and memos, I found to my surprise that, far from replacing philosophy, the pioneers in CS had learned a lot, directly and indirectly from the philosophers. They had taken over Hobbes' claim that reasoning was calculating, Descartes' mental representations, Leibniz's idea of a "universal characteristic"—a set of primitives in which all knowledge could be expressed—Kant's claim that concepts were rules, Frege's formalization of such rules, and Russell's postulation of logical atoms as the building blocks of reality. In short, without realizing it, AI researchers were hard at work turning rationalist philosophy into a research program[11]."

As the remainder of Dreyfus's paper makes clear, it is a bundle of Cartesian assumptions about the mind as a representative system that he wants to reject, following the phenomenological claim that the Cartesian mind-body separation was a derivative mode of experience. In his recounting of the philosophical history of AI research, Luciano Floridi likewise points to the early confluence of Cartesianism (understood as the body-independence of thought) and Hobbesian mechanism about the mind[12]. Floridi then argues that more contemporary efforts at AI are "Cartesian in spirit" because they "reject[] the feasibility of a thinking machine capable of cloning human intelligence". Rather, they "do not necessarily simulate but rather emulate (often do better, although differently) what a human being could do in the same situation[12]".

I think this move to Descartes is a mistake, especially in the context of current AI. The problem is not emulation. It rather lies in the way both simulation and emulation treat thinking as disembodied. The objection here is related to the phenomenological one and is that the Cartesian account of mind proceeds in deliberate isolation from its social and political context.☼ Descartes begins the *Meditations* by asking the faculty at the Sorbonne to give the book their protection and by asking his reader to be willing to "to withdraw their mind from the senses and from all preconceived

☼ I say "related" to the phenomenological account but not identical to it because phenomenological accounts like Heidegger's tend to proceed through typical situations. My claim here is that the specific context of the development of AI matters.

opinions". The narrator begins by emphasizing that he is alone for the proceedings. The initial argument of the *Meditations* then effects a radical withdrawal from the world, to the point of doubting not just the context of what one has learned from sense data, but even the Euclidean geometry we use to interpret it. In his earlier *Discourse on the Method*, Descartes had also taken pains to emphasize that metaphysical speculation need not trouble one's ethics[6]. More broadly, there is good evidence that Descartes, who saw Galileo's condemnation play out during his early career, was involved in metaphysics in the first place because he was trying to create a space for science that was separate from Church politics[13, 14].

Whatever theory of mind underlies current AI work, this Cartesian attitude is persistent in a number of deficiencies in the literature on AI that have been the subject recent of critical attention. It is persists in the tendency to look at algorithms in isolation, without paying attention to their deployment in algorithmic systems[15]. It persists in the tendency to treat "fairness" as a formalizable concept that can be operationalized independently of its context[16]. It persists in the isolation of technical academic work on AI from other disciplines that are more concerned with social impacts[17]. And it persists in the quest to determine whether an AI is conscious at the expense of assessing its current sociopolitical implications.

Whether or not an AI can be conscious or sentient in the Cartesian sense may well turn out to be unknowable, at least for the foreseeable future. It is true that LLMs exhibit many of the attributes often assigned to sentience, such as environmental awareness, in the ability to respond to prompting. However, as I will discuss further below, this awareness is strangely ungrounded, and their tendency to hallucination belies a larger point: bots like ChatGPT have no capacity to connect their utterances to the world; in this sense, they lack the capacity for communicative intent[18]. Even claims that they exhibit emergent properties are turning out to be overblown, perhaps as much about the metrics used to measure those properties as they are about the performance of the models themselves[19]. The philosopher David J. Chalmers concludes an assessment of LLMs by arguing that they are almost certainly not conscious now, though the attributes of consciousness that they lack now might very well be engineered in later[4]. However, he also admits that we do not really understand either what consciousness is, or what LLMs do. Indeed, Chalmers recently won a case of wine for having made a friendly wager—in the 1990s—that we would not by now even have clear evidence for a neural signature of consciousness[20].¤ In the face of all this uncertainty, we need an alternative.

3 The Hobbesian Alternative

Not all seventeenth-century thinkers attempted to treat mind in isolation. Thomas Hobbes extended a mechanistic treatment of body, of the sort one also saw in Descartes, into the mind as well, finding the idea of an "incorporeal substance" such as the Cartesian mind "contradictory and inconsistent", a "mere sound" like a "round quadrangle"[22]. As both Dreyfus and Floridi noted, Hobbes's mechanistic account of thinking seems initially attractive from an AI standpoint, because it explicitly reduces thinking to computation: "all ratiocination", Hobbes writes, "is comprehended in these two operations of the mind, addition and subtraction[23]." However, virtually everything else in his account shows that thinking is inseparable from the context in which it occurs. This is primarily because Hobbes is an empiricist; as I will develop below, for him, all thought originates in the senses[22].

Just as algorithms require training data, so does the Hobbesian mind depend on sense impressions and the memory of them. The effort to understand thinking without sense data in Hobbes is thus like the effort to understand AI by focusing only on the algorithm: it artificially eliminates consideration of necessary components of the system like collecting and labeling data[24]. Because our experiences are idiosyncratic, so are naming and thinking. As Hobbes says, "seeing all names are imposed to signify our conceptions, and all our affections are but conceptions, when we conceive the same things differently, we can hardly avoid different naming of them." He adds that "though the nature of that we conceive be the same, yet the diversity of our reception of it, in respect of different constitutions of body and prejudices of opinion, gives everything a tincture of our different passions[22]". In other words, our different experiences mean we interpret things differently.

¤ For some of the considerations in philosophy of mind that might emerge in a legal effort to ascertain if an AI was conscious, see Ref. [21].

The parallel to AI is direct, and problems with idiosyncratic sourcing and labeling of training data have been a constant headache[25, 26]. For example, datasets of household objects trained on Flickr images perform poorly on objects from low and middle income countries[27]. In ImageNet, hammerhead sharks are depicted as swimming, trout are trophy catches, and lobster are cooked on a plate[28]. Annotators' subjective views of toxicity filter into language models[29]. These markers of cultural difference are built in to AI systems that use them, and ignorance of variability in labeling encourages a view from nowhere epistemology[30, 31]. This epistemology is precisely what a Hobbesian account of reason rejects.

Descartes takes language to be the expression of ideas. Hobbes similarly takes language to be inseparable from thought, but for Hobbes this is because language is how we organize our sense impressions on the way to having thoughts. Sense impressions get names, and reason operates on those names. Accordingly, "our reasoning tell[s] us nothing at all about the nature of things, but merely about the labels applied to them[6]." This view is anomalous in the seventeenth-century terms; Descartes replies that he was following standard practice in taking words to refer to "the things signified by names"[6].* In *Leviathan*, Hobbes explicitly ties computation to names, arguing that "reasoning is nothing but reckoning (that is, adding and subtracting) of the consequences of general names agreed upon for the marking and signifying of our thoughts[22]". The key to reasoning well, it turns out, is careful definition of terms to overcome their diverse significations among different people[22]. Hobbes favorably cites Euclidean geometry for its emphasis on definition and emphasizes that using no or incorrect definitions is an abuse of speech[22].

Hobbes's concerns about language use resonate with our own about LLMs. Because words refer to each other in Hobbes, and because both are grounded in an imagination which may not bear any resemblance to the sensations that initiate it, Hobbes worries a lot about language misuse. As he puts it, "as men abound in copiousness of language; so they become wiser, or madder than ordinary[22]." His specific examples were derived from the Scholastic philosophy he was opposing combined with a heavy dose of anti-Catholicism. For example, he took "separated essences" and other such terms to be both nonsensical and politically dangerous. Thus, "this doctrine of *Separated Essences*, built on the vain Philosophy of Aristotle, would fright [people] from obeying the laws of their country, with empty names; as men fright birds from the corn with an empty doublet, a hat, and a crooked stick[22]." More generally, Hobbes is concerned about what we now call disinformation, and the entities that might spread it. Hallucinating LLMs and their potential to undermine the information processes necessary for good governance are precisely the sort of thing that would concern him. What one does with language is more important than what entity possesses it and whether language use indicates the presence of a conscious being.

Hobbes's account leaves open the category of personhood. In *Leviathan*, he argues that a "person" is any entity "whose words or actions are considered, either as his own, or as representing the words or actions of another man, or of any other thing to whom they are attributed, whether Truly or by Fiction".♭ That is, the category of "person" is important in understanding how we attribute actions to one entity or another. As the discussion makes clear, "natural" personhood, in the sense of individual human beings, is not a broad enough category. His terminology is accordingly deliberately capacious in its discussion of "artificial persons". "Inanimate things, as a church, a hospital, a bridge, may be personated by a rector, master, or overseer[22]." "Children, fools, and madmen" can be personated by another, such as a guardian[22]. And a "multitude" of people can "made one person, when they are by one man, or one person, represented"[22].

Hobbes, then, goes beyond Cartesian metaphysical

* For this exchange, see Ref. [32]. For the radicality of Hobbes's views on language, see, e.g., Refs. [33–35]. This is perhaps a reductive view of Descartes on language (though he himself says relatively little about language). That said, later Cartesians like Arnauld and Nicole object to exactly this point in Hobbes[36]. I am interested in identifying a broadly Cartesian approach to the relation between thought and language, which involves a non-empirical certainty that only thinking entities can fully use language[37]. I want to contrast that kind of approach with one that emphasizes language use as a sociopolitical phenomenon; on my reading, Hobbes's ability to do so is consequent to his reduction of intellect to imagination.

♭ This chapter is difficult, and I am considerably simplifying the discussion here. It is clear that Hobbes's purpose is to understand how actions could be predicated of the state, and how the state can be understood as an artificial person. For a current assessment, see Ref. [38], which underscores the point I want to make here: "what makes Hobbes' idea of personhood unique and valuable is that it decouples personhood from metaphysical conceptions of agency; it explains how states and other entities can be persons even though they do not have any intrinsic capacity for intentionality or action."

determinations. Because thought is linguistic and political, the relevant issues lie in understanding agency and how to assign responsibility. When an entity's words or actions are considered "as his own", it is a natural person. When "they are considered as representing the words and actions of another", the entity is an artificial person. Artificial persons can be considered in two ways. If the representative person's actions are "owned" by the entity they represent, then the representative is an "actor" and the owner is the "author" and the representative acts with "authority". Contemporary notions of proxies or fiduciaries fit this description. In other cases, the underlying entity cannot be an author. This is clearly the case for inanimate objects, in which case the representative acts on the authority of the "owner" or "governor" of the entity represented. For example, the owner of a bridge might authorize the overseer to procure maintenance on the bridge's behalf. The important point here is that the details of such regulatory structures cannot be decided in the abstract, as "such things cannot be personated, before there be some state of civil government"[22].

In sum, our historical habits on questions of AI sentience trace to a way of thinking exemplified by Descartes. They embed an understanding of thought, consciousness, personhood, and reasoning that lend themselves to abstract questions about what consciousness is and what the necessary criteria for it are. The example of Hobbes shows that this approach is not necessary. The Hobbesian alternative is particularly relevant now. The emergence of LLMs forces us to confront the possibility that although language use is a poor proxy for sentience, we need to think carefully about our legal responses to language-using entities. Concerns about the implications of AI for democratic and other governance systems show that language use is a political problem that exists separately from the more metaphysical problem of the sentience or consciousness of AI systems. At the point that the system generates a credible semblance of language, we are forced to confront its social and political implications even in the face of evidence that it is not sentient. The Hobbesian account forces us to think creatively in terms of accountability and responsibility, which is precisely what we need to be thinking about in the context of AI.

4 Hobbesian Thoughts on LLMs

In place of the Cartesian response—is AI sentient?—I propose that we confront the more Hobbesian one: How should we regulate AI systems that behave as if they are sentient? The answer here can only be a sketch, and is in part designed to provoke further debate. I hope that it starts to fill in the reasons a switch to a Hobbesian imaginary for AI can be productive.

4.1 AI as Leviathan-systems

One result of Hobbes's procedure is that it is not clear to what one should apply the term "conscious". A "natural" person merely presents a default case. As noted above, Hobbes is an empiricist. He accordingly argues that there is "no conception in a man's mind" that does not originate in the senses, for which he then offers a mechanistic explanation. He then immediately connects senses to imagination, which he calls "decaying sense" and says is common to people "and many other living Creatures"[22]. Imagination brought on by words is "understanding", and both humans and some animals can possess it. Distinctively human understanding is distinguished "by [someone's] understanding not only his will; but his conceptions and thoughts, by the sequel and contexture of the names of things into affirmations, negations, and other forms of speech"[22]. One might apply "sentience" at any point in this chain, but it is not obvious where.

Descartes circumvents the problem by having God install a linguistic mind into a mechanical body, enabling him to distinguish human consciousness from the operation of a machine. Hobbes conspicuously avoids this step, and instead proposes not only that "life is but a motion of limbs", but also that "automata … have an artificial life." He adds that "art[ifice]" is able to imitate "that rational and most excellent work of nature, man. For by art is created that great Leviathan called a common-wealth, or state … which is but an artificial man[22]."※

The commonwealth, as Hobbes describes it, is aggregative and, in contemporary terms, a sociotechnical system. As the focus on language indicates, a key issue for Hobbes is getting all the components of that system to function coherently.

※ In other words, because he continually reduces phenomena to matter and motion, Hobbes is a thorough nominalist about identity, which means that he thinks that the principle by which we should individuate something depends entirely on the context in which we individuate it[39].

Indeed, without the regulatory apparatus of the social contract, one is dealing with a "multitude" and not a "people"[40]. Of course, the great risk of the Hobbesian apparatus is authoritarianism, as he takes the difficulty in aligning people's individual interests and their understanding of one another to require an absolutely powerful state, a primary purpose of which is to make people capable of coordinated behavior, despite their use of language.[⚙] Individuals aggregated into this system are only allowed their own interests insofar as those align with those of the system; Hobbes's one exception is for self-defense. Smaller aggregated entities are similarly judged by the alignment of their behaviors with the interest of the state.

An AI system also incorporates many different individuals, processes, and systems, both algorithmic and social. This is why there have been calls to treat them as sociotechnical systems, rather than as technical devices[15, 41, 42], to treat issues like algorithmic fairness in the context of social injustice[16, 43–45] and to study the role of technologists in perpetuating unjust systems[17]. It is also why recent work has noted the similarity between algorithmic governance and Hobbes's social contract as coordination mechanisms[46].

The particular concern I want to highlight here is the one that is generally raised against Hobbes: that the aggregative process can obscure when the different individuals composing the system have competing interests, with a tendency to protect only those who are the most powerful. In other words, emphasis on the unity of an AI, or presenting it as conscious, is essentially to treat it like Hobbes treats the Leviathan-state, as a device that creates its own unity by ignoring or suppressing anything that works against that unity.

The current development of AI causes real suffering to unquestionably sentient people. LLMs are dependent on massive training datasets scraped from the Internet, and data from the Internet are often a toxic sludge of racist, misogynistic, and violent content[47]. Public-facing models that use that training data have to detoxify it. This presents a significant problem, and it is one that OpenAI apparently solved by outsourcing to poorly-paid workers in Kenya, who were paid less than 2 US dollars per hour to manually label sexual abuse, hate speech, and violence, leaving them "mentally scarred by the work"[48].

Additional reporting has confirmed the existence of a well-hidden but apparently vast underclass of annotators and raters, whose job it is to rate the responses of chatbots to wide-ranging types of prompts. This Reinforcement Learning with Human Feedback (RLHF) is essential to the systems' performance and ability to sound human. Particularly in its lower-skilled versions, the work is both extremely precarious and greatly underpaid[49]. In short, underlying the performative sentience of these systems, and responsible for it, are hundreds of thousands of hours of human training.

Treating sentience or meaning as the product of the system as a whole makes it much harder to see and protect these less powerful stakeholders. If the system as a whole is the place to locate sentience, the interests that matter are those of the system as a whole. It might intuitively seem to be in the interest of a public-facing AI not to spew racist text at the slightest provocation, since if it did so, people would stop using it or even dismantle it. Models trained with RLHF will even tend to express a desire not to be shut down[50]. This intuition obscures that there are actually several interests involved. The interest in continuing to operate seems to derive from the financial interests of the system's owners. The AI's desire to remain on is an artifact of its training and should not be confused with the owners' financial interests. The public that will encounter the AI has an interest in not being inundated with racist text from it. And the workers have an interest in not being exploited, even if that makes the system produce less racist text.

Resolving the problem by treating the AI as itself having interests requires a conception of what is good for it[21]. However, that is underdetermined in a sociotechnical system composed of many competing interests. Sweeping the exploitation of workers under the rug is one way of balancing these interests, but it clearly elevates the interests of the owners of the system and treats those as equivalent to the interests of the system. The interests of the workers could be preserved (or better preserved, at least) at the expense of the profit margins of the company by paying the workers better and by better addressing their mental health. And of course the interests of the workers and the public could also be accommodated by shutting the

[⚙] In other words, Hobbes rejects the Aristotelian dictum that humans are naturally capable of political unity. I defend this reading, and the emphasis on language in Hobbes, in Ref. [33].

system down. Hobbes understands these competing interests as existential threats to the Leviathan-state and tries to prevent the rise of "factions" that preserve their own private interests[22].

Hobbes is writing in the immediate aftermath of the English civil war, and for him, the collapse of the state into factionalism is to be avoided at any cost. However, it is not clear that we should endorse his solution to civil war in the context of AI systems. In addition to work on, for example, problems with training data, this means that it will be important to assess technical work for its resistance to this sort of centralization (e.g., differential vs. federated privacy[31]) and the systemic risks created by vesting too much control in top-down structures[51].

Consider also the case of AI's carbon footprint as an externality. It is already difficult to get policymakers, regulators, and companies to include the cost of environmental harms in their planning, since decarbonization is expensive and often at odds with the interests of various short-term stakeholders. Reducing the carbon footprint of AI would appear to be detrimental to its interests because it would worsen its performance. But this again tends to conflate the owners' financial interests and the engineering teams' interests in performance with both the interests of the AI system as a whole and society outside it. Machine learning research emphasizes technical prowess and downplays social harms[52], but even if these are virtues for an LLM taken in isolation, there is no *a priori* reason to elevate them to the interest of the model as a sociotechnical system. If the interest of the AI is to "help people"—which is, after all, more or less what ChatGPT has been engineered to describe as its purpose—then it is not clear that the carbon-intensive scaling up of LLMs furthers that interest[53].

4.2 Dreaming language models

A second problem is that a Cartesian approach makes accountability questions more difficult. Consider the case of defamatory AI. In several cases, LLMs hallucinated stories about real people that would have been legally actionable defamation if a human published them. In at least one case, the victim considered legal action against OpenAI over it[54, 55].

Hobbes's account of language offers a window into the problem. As noted above, his psychology moves from sense perception to imagination and understanding, with speech as enabling the move from imagination to understanding. The route backwards form language through imagination as "decaying" sense means that the relation between speech and sense experience is important.

Hobbes initially approaches this in terms of dreaming. Those who sleep are not moved by external objects; in this "silence of sense", the imagination relies on what is already in the brain and allows it to present more clearly than it would if the more "vigorous impression" of waking thoughts could interfere. By comparing the difference between the objects of waking thought and dreams, their relative coherence, and his ability to discern the absurdity of dreams, Hobbes pronounces himself "well satisfied" that he is awake, ignoring both the Cartesian demand for certainty on the point and the divine guarantor Descartes invokes to get there. Hobbes instead proceeds to consider the problems induced by a failure to take such steps, which accounts for the beliefs that some people have in witches and supernatural entities[22].

The problem with referentiality in LLMs is parallel to the one Hobbes identifies here: in both cases, harmful speech arises from a tendency to confuse a plausible-seeming imagination and reality. The LLM generates antisocial language that threatens to undermine social trust by picking up vaguely on people's social and political fears. In that sense it is analogous to witchcraft, accusations of which were generally used to press feuds between people[56].

One mechanism for accountability punishes the accused witch as an individual. Hobbes proposes: "I think not that their witchcraft is any real power; but yet that they are justly punished, for the false belief they have, that they can do such mischief, joined with their purpose to do it if they can[22]." Hobbes's account of personation shows another, applicable to those who cannot be held accountable for the due diligence of distinguishing dreams and wakefulness. It is not just bridges that require personation; "likewise children, fools, and madmen that have no use of reason may be personated by guardians or curators, but can be no authors (during that time) of any action done by them[22]." A Cartesian account, by conflating reason and language use, and by not attending to the psychology of language acquisition, makes it harder see the second option.©

The Hobbesian approach, on the other hand, allows space to treat another entity as ultimately responsible for the AI's output.

As a sociotechnical system, the AI is more like a bridge and less like a witch. On a basic reading, an LLM is predicting text on the basis of how text in its training data tends to sound[53]. Subsequent work complicates but does not disturb the stochastic parrots model: For example, LLMs appear to learn and use representations of the outside world[57], but they are still predictive. The persistence of hallucination across language models[58] underscores that their linguistic use is not referential in any ordinary sense. The increased use of human trainers and raters also complicates any assessment of how to understand their output. LLMs trained with RLHF seem profoundly shaped by that human feedback, developing stronger political views and a tendency to sycophancy, tailoring their output to their users' apparent beliefs[50]. RLHF does open the possibility of viewing language models as somehow referential, in the sense that their predictions are rated by humans who have relevant experience[59]. However, that even this attenuated link to sense impressions improves the AI's performance underlines the difference between human and AI language use and acquisition. Indeed, one can make a Cartesian-style argument here that LLMs do not use language because they do not exhibit understanding[60].

How does this interact with defamation? Defamation laws are designed to catch witches, not impose accountability for bridges. Under US law, defamation of a public official requires a showing of "actual malice", defined as making a false, defamatory statement "with knowledge that it was false or with reckless disregard of whether it was false or not"[61]. It seems fairly clear that any current language model does not have knowledge that a statement it makes is false: There is no mechanism that connects its "imaginations" to "sense" data. The model has no way to know if it is awake or dreaming, as it were.

The defamation example is a window into a more general problem. Large sections of the law contain a mental state requirement, and this requirement ports very uneasily to AI systems, which can increasingly do things (like defame others) without the relevant mental states that we use to hold humans accountable[62]. One very good reason to notice this difference between regulating witches and bridges is that corporations have learned to exploit it to avoid liability[63]. For example, the Illinois Biometric Information Privacy Act (BIPA) requires that companies that use biometric identification techniques like facial recognition obtain affirmative, opt-in consent before doing so. Defending itself in federal litigation over its photo-tagging feature, Facebook pointed to the algorithm and declared that whatever it was doing, it was not facial recognition. As the court hearing the case put it:

"Plaintiffs say the technology necessarily collects scans of face geometry because it uses human facial regions to process, characterize, and ultimately recognize face images. Facebook disagrees and says the technology has no express dependency on human facial features at all. Rather, according to Facebook, the technology 'learns for itself what distinguishes different faces and then improves itself based on its successes and failures, using unknown criteria that have yielded successful outputs in the past[64]'."

The analogy to LLMs should be apparent: One can easily imagine a filing that asserts that the technology learns for itself what to associate with different words, and that it does so without any reference to anyone actually existing outside of its training data, which includes many fictitious people. Defendants will also argue that the use of human raters shows that whatever toxicity is emanating from the model is at least partly due to the raters, and not attributable to the model.

This Cartesian argument, which centers on the idea that the system is not using language properly, could thereby prevent holding the system accountable for defamation and issuing threats. The latter in particular is a well-documented harm, especially to women, who often suffer from a barrage of violent threats simply for participating online[65, 66]. An LLM could issue thousands of such threats a minute, and plausibly be defended as lacking the relevant mental state. All of this is not to say that it will be easy to hold

◎ Descartes readily allows that madmen have use of language[6]. His *Meditations* famously briefly considers and then dismisses the idea that the narrative voice is insane: "perhaps I were to liken myself to madmen, whose brains are so damaged by the persistent vapours of melancholia that they firmly maintain they are kings when they are paupers, or say they are dressed in purple when they are naked, or that their heads are made of earthenware, or that they are pumpkins, or made of glass[6]." The meditator defaults instead to the hypothesis that he is dreaming. My argument here is that the more thoroughgoing mechanism in Hobbes is better positioned to address the "hallucinatory" speech of language models because it does not default to metaphysical guarantees on the dreaming question.

corporations accountable for the behavior of their AI systems. It is to say that the Hobbesian approach to language, personation, and accountability allows a different way of framing the problem. If nothing else, focusing liability on the corporation that owns or created the harmful system opens other avenues of relief, not based on mental states at all, such as product liability laws.

These cases underscore that both the creation and operation of AI systems can impose substantial harms on people. Questions about AI sentience, whatever their intrinsic interest, can easily obscure these harms or lead us to frame them unhelpfully. The result is that those most subject to them often have no recourse—they fall into what Alicia Solow-Nederman calls "grey holes of accountability", where redress is nominally but not really available[67]. They find themselves in the position of those targeted by the Leviathan state. Hobbes insists that no one should be construed as giving up their right to self-defense; those condemned by the state have a right to resist their fate[22]. Of course, there is little they can practically do to effectuate this right. We should do better by those who are harmed by AI systems.

5 Conclusion

Hobbes and Descartes shared a fascination with automata and mechanistic explanations of living beings. Unlike Descartes, Hobbes is willing to extend the mechanism to thought itself. This makes Hobbes a better starting point for thinking about AI than Descartes. Hobbes understood, better than many others in the seventeenth-century and in direct opposition to Descartes, that questions of consciousness, especially as they were interlaced with questions of language, could only be tackled socio-politically. Where Cartesianism encourages disembodied accounts of reason, Hobbes's more thorough mechanism treats thinking as a system whose components include not just the computational processes of thought, but its data inputs and their sociopolitical location.

Hobbes is of course associated with the authoritarianism of the Leviathan-state, which he sees as the only way to resolve the inability of humans to govern themselves without top-down authority structures. As the foregoing suggests, this has uncomfortable implications for thinking about AI and questions of sentience. The Cartesian view tends to elide the extent to which AI is always a sociotechnical system, obscuring the extent to which its supposed unity is artifactual. In so doing, it risks legitimating a Leviathan-esque treatment of its components, whether by obscuring the humans whose manual and cognitive labor it incorporates or by treating the sentience of the system as a key point for accountability. Similarly, the Cartesian view tends to encourage a model of systems and mental states that allows the corporations that create AI systems to avoid accountability for what they do.

More generally, research into AI is plagued by a Cartesian tendency to treat epistemic questions about consciousness as if they could be separated from their context. If that tendency was ever justified, it is no longer justified in the case of AI systems, which are extremely complex sociotechnical artifacts; their outward performance of language creates an artificial origin point that covers over a wide range of other social and technical processes. The Hobbesian alternative treats consciousness and expression as social and political processes. One might resist this move in the context of humans, but it surely has purchase in the context of technical systems. The advantage of Hobbesian approach is that it denaturalizes categories like consciousness and personhood, allowing their disambiguation for regulatory and other purposes. This is useful as both an approach and a caution. On one hand, it is better for thinking about how AI systems behave as social actors and how they can harm the vulnerable. As an approach, it favors technical solutions that prioritize the risks created by top-down system structures and by ignoring the social context of language models in their assessment. It also urges attention to non-technical approaches that treat AI systems as sociotechnical entities[15]. On the other hand, it also directs our attention to a crucial and much more fundamental problem: How can we live with language-using machines without creating authoritarian social structures?

References

[1] B. Perrigo, Bing's AI is threatening users. That's no laughing matter, *Time*, https://time.com/6256529/bing-openai-chatgpt-danger-alignment/, 2023.

[2] K. Roose, A conversation with Bing's Chatbot left me deeply unsettled, *The New York Times*, https://www.nytimes.com/2023/02/16/technology/bing-

chatbot-microsoft-chatgpt.html, 2023.
[3] B. Lemoine, What is LaMDA and what does it want? https://cajundiscordian.medium.com/what-is-lamda-and-what-does-it-want-688632134489, 2023.
[4] D. J. Chalmers, Could a large language model be conscious? arXiv preprint arXiv: 2303.07103, 2023.
[5] P. S. MacDonald, Descartes: The lost episodes, *J. Hist. Philos.*, vol. 40, no. 4, pp. 437–460, 2002.
[6] R. Descartes, *The Philosophical Writings of Descartes*. Cambridge, UK: Cambridge University Press, 1984.
[7] T. M. Schmaltz, Malebranche on Descartes on mind-body distinctness, *J. Hist. Philos.*, vol. 32, no. 4, pp. 573–603, 1994.
[8] R. Descartes, *Oeuvres de Descartes*. Paris, France: J. Vrin, 1957.
[9] S. Gaukroger, *Descartes' System of Natural Philosophy*. Cambridge, UK: Cambridge University Press, 2002.
[10] G. M. Ross, Hobbes and Descartes on the relation between language and consciousness, *Synthese*, vol. 75, no. 2, pp. 217–229, 1988.
[11] H. L. Dreyfus, Why Heideggerian AI failed and how fixing it would require making it more Heideggerian, *Artif. Intell.*, vol. 171, no. 18, pp. 1137–1160, 2007.
[12] L. Floridi, *Philosophy and Computing: An Introduction*. London, UK: Taylor & Francis, 2002.
[13] S. Gaukroger, *Descartes: An Intellectual Biography*. Oxford, UK: Oxford University Press, 1995.
[14] T. M. Schmaltz, *Radical Cartesianism: The French Reception of Descartes*. Cambridge, UK: Cambridge University Press, 2002.
[15] A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian, and J. Vertesi, Fairness and abstraction in sociotechnical systems, in *Proc. Conf. Fairness, Accountability, and Transparency*, Atlanta, GA, USA, 2019, pp. 59–68.
[16] B. Green, Escaping the impossibility of fairness: From formal to substantive algorithmic fairness, *Philos. Technol.*, vol. 35, no. 4, pp. 1–32, 2022.
[17] C. Barabas, C. Doyle, J. Rubinovitz, and K. Dinakar, Studying up: Reorienting the study of algorithmic fairness around issues of power, in *Proc. 2020 Conf. Fairness, Accountability, and Transparency*, Barcelona, Spain, 2020, pp. 167–176.
[18] E. M. Bender and A. Koller, Climbing towards NLU: On meaning, form, and understanding in the age of data, in *Proc. 58th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, USA, 2020, pp. 5185–1598.
[19] R. Schaeffer, B. Miranda, and S. Koyejo, Are emergent abilities of large language models a mirage? arXiv preprint arXiv: 2304.15004, 2023.
[20] J. Horgan, A 25-year-old bet about consciousness has finally been settled, *Scientific American*, https://www.scientificamerican.com/article/a-25-year-old-bet-about-consciousness-has-finally-been-settled/, 2023.
[21] L. B. Solum, Legal personhood for artificial intelligences, *North Carolina Law Review*, vol. 70, pp. 1231–87, 1992.
[22] T. Hobbes, *Leviathan: With Selected Variants from the Latin Edition of 1668*. Indianapolis, IN, USA: Hackett Pub. Co., 1994.
[23] T. Hobbes, *The English Works of Thomas Hobbes of Malmesbury*. London, UK: Bohn, 1839.
[24] N. Sambasivan, S. Kapania, H. Highfill, D. Akrong, P. Paritosh, and L. M. Aroyo, "Everyone wants to do the model work, not the data work": Data cascades in high-stakes AI, in *Proc. 2021 CHI Conf. Human Factors in Computing Systems*, Yokohama, Japan, 2021, pp. 1–15.
[25] C. G. Northcutt, A. Athalye, and J. Mueller, Pervasive label errors in test sets destabilize machine learning benchmarks, arXiv preprint arXiv: 2103.14749, 2021.
[26] A. Paullada, I. D. Raji, E. M. Bender, E. Denton, and A. Hanna, Data and its (dis)contents: A survey of dataset development and use in machine learning research, *Patterns*, vol. 2, no. 11, p. 100336, 2021.
[27] T. DeVries, I. Misra, C. Wang, and L. V. D. Maaten, Does object recognition work for everyone? arXiv preprint arXiv: 1906.02659, 2019.
[28] N. Malevé, An introduction to image datasets, https://unthinking.photography/articles/an-introduction-to-image-datasets, 2023.
[29] M. Sap, S. Swayamdipta, L. Vianna, X. Zhou, Y. Choi, and N. A. Smith, Annotators with attitudes: How annotator beliefs and identities bias toxic language detection, arXiv preprint arXiv: 2111.07997, 2021.
[30] E. Denton, A. Hanna, R. Amironesei, A. Smart, and H. Nicole, On the genealogy of machine learning datasets: A critical history of ImageNet, *Big Data Soc.*, doi: 10.1177/20539517211035955.
[31] O. Keyes and K. Creel, Artificial knowing otherwise, *Feminist Philosophy Quarterly*, vol. 8, no. 3, pp. 1–26, 2022.
[32] D. W. Hanson, Reconsidering hobbes's conventionalism, *Rev. Polit.*, vol. 53, no. 4, pp. 627–651, 1991.
[33] G. Hull, *Hobbes and the Making of Modern Political Thought*. London, UK: Continuum, 2009.
[34] P. Pettit, *Made with Words: Hobbes on Language, Mind, and Politics*. Princeton, NJ, USA: Princeton University Press, 2008.
[35] Y. C. Zarka, *Hobbes et la pensée politique moderne*. Paris, France: Presses universitaires de France, 1995.
[36] A. Arnauld and P. Nicole, *Logic or the Art of Thinking*. Cambridge, UK: Cambridge University Press, 1996.
[37] K. Morris, Bêtes-machines, in *Descartes' Natural Philosophy*, S. Gaukroger, J. Schuster, and J. Sutton, eds. New York, NY, USA: Routledge, 2000, pp. 401–419.
[38] S. Fleming, The two faces of personhood: Hobbes, corporate agency and the personality of the state, *Eur. J. Polit. Theory*, vol. 20, no. 1, pp. 5–26, 2021.
[39] Y. C. Zarka, *L'autre voie de la subjectivité: six études sur le sujet et le droit naturel au XVIIe siècle*. Paris, France: Editions Beauchesne, 2000.
[40] J. Chanteur, Note sur les Notions de 'Peuple' et de 'Multitude' chez Hobbes, in *Hobbes-Forschungen*, M. A. Cattaneo, K. Reinhart, and R. Schnur, eds. Berlin, Germany: Duncker & Humblot, 1969, pp. 223–235.
[41] B. Green and S. Viljoen, Algorithmic realism: Expanding the boundaries of algorithmic thought, in *Proc. 2020 Conf. Fairness, Accountability, and Transparency*, Barcelona, Spain, 2020, pp. 19–31.
[42] K. Joyce, L. Smith-Doerr, S. Alegria, S. Bell, T. Cruz, S. G. Hoffman, S. U. Noble, and B. Shestakofsky, Toward a sociology of artificial intelligence: A call for research on

inequalities and structural change, *Socius Sociol. Res. a Dyn. World*, doi: 10.1177/2378023121999581.
[43] A. Hanna, E. Denton, A. Smart, and J. Smith-Loud, Towards a critical race methodology in algorithmic fairness, in *Proc. 2020 Conf. Fairness, Accountability, and Transparency*, Barcelona, Spain, 2020, pp. 501–512.
[44] G. Hull, Dirty data labeled dirt cheap: Epistemic injustice in machine learning systems, *Ethics Inf. Technol.*, vol. 25, no. 3, pp. 1–14, 2023.
[45] N. Okidegbe, Discredited data, *Cornell Law Review*, vol. 107, no. 7, pp. 2007–2065, 2022.
[46] P. D. König, Dissecting the algorithmic leviathan: On the socio-political anatomy of algorithmic governance, *Philos. Technol.*, vol. 33, no. 3, pp. 467–485, 2020.
[47] A. Birhane, V. U. Prabhu, and E. Kahembwe, Multimodal datasets: Misogyny, pornography, and malignant stereotypes, arXiv preprint arXiv: 2110.01963, 2021.
[48] B. Perrigo, OpenAI used Kenyan workers on less than $2 per hour to make ChatGPT less toxic, *Time*, https://time.com/6247678/openai-chatgpt-kenya-workers/, 2023.
[49] J. Dzieza, Inside the AI factory, *Intelligencer*, https://nymag.com/intelligencer/article/ai-artificial-intelligence-humans-technology-business-factory.html, 2023.
[50] E. Perez, S. Ringer, K. Lukošiūtė, K. Nguyen, E. Chen, S. Heiner, C. Pettit, C. Olsson, S. Kundu, S. Kadavath, et al., Discovering language model behaviors with model-written evaluations, arXiv preprint arXiv: 2212.09251, 2022.
[51] K. Creel and D. Hellman, The algorithmic Leviathan: Arbitrariness, fairness, and opportunity in algorithmic decision making systems, in *Proc. 2021 ACM Conf. Fairness, Accountability, and Transparency*, Virtual Event, Canada, 2021, p. 816.
[52] A. Birhane, P. Kalluri, D. Card, W. Agnew, R. Dotan, and M. Bao, The values encoded in machine learning research, in *Proc. 2022 ACM Conf. Fairness, Accountability, and Transparency*, Seoul, Republic of Korea, 2022, pp. 173–184.
[53] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, On the dangers of stochastic parrots: Can language models be too big? in *Proc. 2021 ACM Conf. Fairness, Accountability, and Transparency*, Virtual Event, Canada, 2021, pp. 610–623.
[54] P. Verma and W. Oremus, ChatGPT invented a sexual harassment scandal and named a real law prof as the accused, *Washington Post*, https://www.washingtonpost.com/technology/2023/04/05/chatgpt-lies/, 2023.
[55] L. Sands, ChatGPT falsely told voters their mayor was jailed for bribery. He may sue. *Washington Post*, https://www.washingtonpost.com/technology/2023/04/06/chatgpt-australia-mayor-lawsuit-lies/, 2023.
[56] G. Hull, Building better citizens: Hobbes against the ontological illusion, *Epoché: A Journal for the History of Philosophy*, vol. 20, no. 1, pp. 105–29, 2015.
[57] S. R. Bowman, Eight things to know about large language models, arXiv preprint arXiv: 2304.00612, 2023.
[58] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, Survey of hallucination in natural language generation, *ACM Comput. Surv.*, vol. 55, no. 12, pp. 1–38, 2023.
[59] D. C. Mollo and R. Millière, The vector grounding problem, arXiv preprint arXiv: 2304.01481, 2023.
[60] O. H. Hamid, ChatGPT and the Chinese room argument: An eloquent AI conversationalist lacking true understanding and consciousness, in *Proc. 2023 9th Int. Conf. Information Technology Trends* (*ITT*), Dubai, United Arab Emirates, 2023, pp. 238–241.
[61] New York Times Co. v. Sullivan, 376 U.S. 254, 1964.
[62] M. Chatterjee and J. C. Fromer, Minds, machines, and the law: The case of volition in copyright law, *Columbia Law Review*, vol. 119, pp. 1887–1916, 2019.
[63] G. Hull, The death of the data subject, *Law Cult. Humanit.*, doi: 10.1177/17438721211049376.
[64] In re Facebook Biometric Info. Privacy Litig., 2018 U.S. Dist. LEXIS 81044.
[65] D. K. Citron, *Hate Crimes in Cyberspace*. Cambridge, MA, USA: Harvard University Press, 2014.
[66] D. K. Citron, *The Fight for Privacy: Protecting Dignity, Identity and Love in the Digital Age*. New York, NY, USA: Random House, 2022.
[67] A. Solow-Niederman, Algorithmic grey holes, *Journal of Law and Innovation*, vol. 5, no. 1, pp. 116–139, 2023.

Gordon Hull is a professor of philosophy and public policy and affiliates with the School of Data Science, UNC Charlotte, USA. He works on issues of technology, law, and policy. He is the author of numerous articles on these topics as well as *Hobbes and the Making of Modern Political Thought* (2009) and *The Biopolitics of Intellectual Property* (2020).