Why You Are (Probably) Anthropomorphizing AI

Ali Hasan

[Draft. Please not cite without permission.]

Abstract: In this paper I argue that, given the way that AI models work and the way that ordinary human rationality works, it is very likely that people are anthropomorphizing AI, with potentially serious consequences. I start with the core idea, recently defended by Thomas Kelly (2022) among others, that bias involves a systematic departure from a genuine standard or norm. I briefly discuss how bias can take on different explicit, implicit, and "truly implicit" (Johnson 2021) forms such as bias by proxy. I then discuss biased anthropomorphism of technology, focusing on the case of Large Language Models (LLMs) like chatGPT. As with other kinds of bias, there are importantly different kinds of anthropomorphism, some of which can persist without others, and some of which can encourage others. Anthropomorphism can take rather subtle, implicit forms, that can be difficult to detect, resist, and dislodge. Attention to these kinds of anthropomorphism can help us avoid confusing importantly different kinds of evaluation, better assess the risks, and inform strategies for bias prevention and mitigation.

It's still magic, even if you know how it works.
--Terry Pratchett

1. Bias

Let's start with the general idea that bias involves a "systematic [as opposed to a simply random] departure from a genuine norm or standard of correctness" (Kelly 2022).[1] We can distinguish kinds of bias by the kind of norm involved. For example, statistical bias is an example of departure from *accuracy* norms; confirmation bias and the gambler's fallacy are examples of departure from *epistemic* norms (like the norm to be rational or to believe in accordance with one's evidence); the sunk costs fallacy, strong risk aversion, and optimism bias depart from *practical or decision-theoretic* norms (like maximizing expected value); biases of racism and sexism are departures from *moral* norms (like norms against discriminating).

We can also distinguish different biases not by the nature of the norm but by asking questions about the source of deviation from the norm. A classic, straightforward kind of bias involves cases where the very content of one's belief or judgment is in tension with the norm. One might believe that "foreigners are untrustworthy". This disposition to think differently of foreigners might be a disposition to depart from epistemic norms and/or accuracy norms of

[1] See also psychologist Jonathan Baron (2012). Kelly calls this a "norm-theoretic" account. Perhaps the account does not capture all uses of the term in ordinary language. Still, something like this sense of bias plausibly captures what is common across many different and interesting cases.

belief; it might also depart from moral, anti-discrimination norms of belief.[2] And if this belief influences action, disposing one to behave as though foreigners as untrustworthy, this would be a departure from anti-discrimination norms of action. Among these cases of bias due to patently immoral or biased content, we can distinguish between having an *accessible belief* with biased content ("explicit bias"), and having a relatively *inaccessible belief* with biased content ("implicit bias"). So having an implicit bias in this sense is compatible with sincerely believing (falsely) that one does not have it. An explicit bias is one that is consciously held or relatively accessible or available to consciousness. Note that its being consciously held doesn't imply recognizing that it is a bias or that it departs from a genuine norm.

Bias can have its source in other dispositional states, like affective states or emotions—fear, hate, disgust, anger, envy, arrogance. On one dominant view, emotions have an essential judgmental or doxastic component—e.g., fearing foreigners is partly constituted by some negative evaluation of foreigners as untrustworthy or dangerous. In any case, most seem to agree that emotions at least typically if not always involve evaluations or appraisals of some sort.[3] Such emotions can be biased in similar ways to belief.

In addition to explicit and implicit bias, there's what is sometimes called "truly implicit bias" (Johnson 2021). One can have a disposition to think and/or behave *as if* one has a biased belief (or other state with biased content) even though one doesn't. These dispositions might be due to the influence of interests rather than beliefs, as when an employer's preference to be in the company of others of a similar social or racial background influences evaluation of applicants to the job. Or they might be due to the influence of subconscious or sub-personal processing, as when confirmation bias leads to over-estimating the strength of new evidence for one's already held belief.

Some truly implicit biases involve *proxies* for categories like race and gender, as when one has negative beliefs about individuals based on their clothing, hairstyle, accent, job, or place of residence, where these are strongly correlated with a particular race or gender, without necessarily forming negative beliefs explicitly about race or gender. Or, to give an example of institutional bias by proxy: Black Americans are more likely to face tax audits. Why? Because the sorts of potential errors that they tend to make are easy for I.R.S. systems to identify, and the ones with errors that are easy to identify are the ones the I.R.S. audits more often (Tankersley 2021). In this case, tax filings with errors that are easy to find are proxy for tax filings by Black Americans.

There are also cases involving a kind of proxy not for the social group or identity (such as race or gender) but for some negative or normative property. Suppose that an employer has an explicit affirmative action policy in its hiring practices, and that some employees of

[2] Is there really such a norm of *belief*, where it is a moral norm and not merely an epistemic norm? Perhaps not. But even if that's right, if one's belief leads one to be disposed to treat foreigners differently, we would then have a departure from a moral norm.

[3] See the discussion of evaluative and other accounts of emotion in Scarantino and de Sousa (2021).

underrepresented group G are hired, ostensibly under that policy. Suppose that other employees do not hold a biased belief against group G, but mistakenly think that the policy leads to less qualified hires. We can even suppose that the members of group G that are hired would have been hired even without there being an affirmative action policy. The result is that employees belonging to group G tend to be treated by fellow employees as less qualified than they are; the employer unwittingly encourages employees to treat Gs as less qualified. In this way, the employer or company's treatment of G as hired under an affirmative action policy can be a kind of proxy for their being (unfairly) treated as unqualified (or less qualified). How best to solve the problem—whether to stop using the policy, or change employees' attitudes about the policy, breaking the proxy's relation to what it is proxy for—is a further matter.

Note that in the example just given, the negative consequence of the employer's affirmative action policy is mediated by others' attitudes. There are arguably many instances of this—where an attitude or policy on the part of one (single or collective) agent that is not itself biased in its content can have a problematic, discriminatory or unfair consequence due to the attitudes of others in that particular context.

Here's a simple example that doesn't involve such mediation by others' attitudes. A hospital provides different medical care to men and women suffering from the same illness A, based on strong evidence that different medications work differently for each. But suppose that, unbeknownst to the medical staff, the medication given to women makes them highly susceptible to a much more serious illness B (while there is no similar risk for men). The staff believe, and act on the belief, that men and women with illness A should be given different medical treatment, without yet realizing that this treatment exposes women to significantly more risk of illness B than men.

Bias by interest, subconscious processing, and by proxy are all forms of bias that don't require the corresponding immoral, biased, or stereotyping content, whether easily accessible or not, and in this sense this bias is "truly implicit."

Biases might take human, institutional, and technological forms. If we understand bias as a systematic departure from genuine norms, and we can make sense of institutions and technology (like algorithms and automated decision systems) as behaving in ways that they shouldn't according to some genuine standard—if they depart from certain norms of accuracy and/or certain legal or moral norms—then we can make sense of institutions and technology as being biased. Much more can be said about biases in general and the distinctions between forms of bias.[4] But let us turn to human biases of technology, and specifically, biased anthropomorphism about AI.

2. Anthropomorphism about AI and Large Language Models (LLMs)

[4] See Kelly (2022) for an excellent discussion of the "norm-theoretic" account of bias. On implicit bias, see Holroyd (2017), Beeghly and Madva (2020), Mandelbaum (2016), Brownstein (2017). On algorithmic biases, see Johnson (2021), and Fazelpour and Danks (2021).

Anthropomorphism of technology involves applying some human properties to technology. So understood, anthropomorphism is not necessarily problematic. It need not conflict with or depart from a genuine norm; "unbiased anthropomorphism" doesn't seem to be a contradiction in terms. Some technology might, after all, genuinely have a specific human or human-like feature, and attributing such a feature need not be out of line with the evidence, or be practically or morally problematic. Moreover, one might erroneously apply a human property in some particular case without being disposed to apply it elsewhere; it might be a more or less random departure from an appropriate norm of not anthropomorphizing without doing so in any patterned, systematic, or regular way. As I am interested in cases in which anthropomorphism is a bias, however, I will often drop the qualifier.

Though what I say here has broader implications and applications, I will focus on a particular kind of AI that has received a lot of attention lately – generative, conversational AI powered by large language models (LLMs).[5]

Anthropomorphism can take different forms. One way to distinguish them is by the sort of human-like feature that is explicitly or implicitly applied: *consciousness, feelings, thoughts, understanding, intentions, interests, emotions, happiness, sadness, character, personality, conscience,* and so on. In some cases, I might ascribe human-like physical characteristics to technology: I might believe of a voice assistant's output that it sounds human or that it sounds like a happy person, or of a robot that it looks or moves like a human; these might be unbiased (accurate) anthropomorphisms. Some kinds of anthropomorphism, even when false or biased, might be justifiable in some circumstances such as the use of humanoid robots in some health settings.[6]

There has been some discussion of explicit (accessible) and implicit (relatively inaccessible) belief ascribing human features to technology.[7] LLMs like chatGPT present a very interesting and pertinent case where some explicit and implicit forms of anthropomorphism come up. As such LLMs develop further, and as they are combined with other modalities (audio-visual input, more varied forms of output, and integration with robotic systems) explicitly or implicitly attributing consciousness, feeling, thoughts, intentions, desires, and so on could become increasingly difficult to resist. Currently, most people are unlikely to believe that today's generative AI literally have such properties, but there are exceptions (some quite tragic) and the numbers are likely to grow.[8] And while some might not believe that AI has such features,

[5] For some related discussion of LLM's misleadingly plausible but untruthful outputs, and related concerns with LLMs, but that do not explicitly discuss anthropomorphism, see Bender et al. (2021) and Sobieszek and Price (2022). For a classification of various types of harms from use of LLMs, including some brief discussion of anthropomorphism, see Weidinger et al. (2021).

[6] For a discussion of some contexts where anthropomorphism might be permissible, see Darling (2017).

[7] See Kim and Sundar (2012) for a discussion of "conscious (mindful)" and "unconscious (mindless)" anthropomorphism.

[8] See, for example, the story about the Google engineer who believes the company's AI, LaMDA, is sentient (Tiku 2022) and the story regarding a suicide after conversations with a chatbot (Pasquini 2023).

they might not disbelieve it either – they might feel confused and not know what to think. In a recent rant about AI, Snoop Dogg, in his inimitable way, seems to anthropomorphize but also to just express concern and confusion about what to think.[9]

The bias might take a truly implicit form. Here's an interesting case: Imagine that Dave has some but relatively limited experience with and information about models like chatGPT, and that he sits down to try out the latest version of it (powered by GPT-4 architecture). Dave finds that chatGPT's responses in natural language are not only grammatical but organized, specific, and apparently conceptually coherent to a high degree. It seems capable of responding to complex questions across a wide range of topics by generating more or less the sort of testimonial or information-sharing text he would expect to be produced by humans. It can also mimic the sort of text that humans write when they are trying to express themselves or their feelings.

After a while of playing with chatGPT, Dave says: "Wow. How is it able to do that? I mean, I don't believe that it is conscious or has feelings or really understands what it's saying, but it sure *seems* like it does. It must have some kind of internal structure and abilities that work *like* understanding, even though there's really no consciousness or understanding in the machine. Otherwise, how could it respond directly and coherently to all my prompts and questions?"

Dave need not believe or think that chatGPT is really understanding, conscious, has feelings, etc. He avoids applying obvious or straightforward features of persons or human minds, and so avoids anthropomorphizing in this classical or standard sense. But he might believe it can *simulate* some of these features, that it has an "internal structure and abilities that work like understanding." (That these systems are often described ambiguously as involving "deep learning" and "neural networks" doesn't help.) Dave is likely mistaken in taking chatGPT to have an internal structure like whatever internal structure we, our minds, or our brains are in when we understand. For chatGPT is in a sense very good at *mimicry*, but mimicry is not the same as *simulation.*

To see this point, it might help to distinguish between being functionally alike and being behaviorally alike. For two systems to be behaviorally alike over some range of situations is for them to produce the same outputs given the same inputs in those situations. But for them to be functionally alike, not only must they generate the same (or similar) outputs if given the same inputs, but they must do so via internal mechanisms or processes with the same (or similar) causal structure. This is why I said that mimicry is not the same as simulation: that chatGPT has a way to produce the same (or similar) outputs given textual input as humans do

[9] "Well I got a motherfucking AI right now that they did made for me. This n***** could talk to me. I'm like, man this thing can hold a real conversation? Like real for real? Like it's blowing my mind because I watched movies on this as a kid years ago. When I see this shit I'm like what is going on? And I heard the dude, the old dude that created AI saying, 'This is not safe, 'cause the AIs got their own minds, and these motherfuckers gonna start doing their own shit. I'm like, are we in a fucking move right now, or what? The fuck man? So do I need to invest in AI so I can have one with me? Or like, do ya'll know? Shit, what the fuck? I'm lost, I don't know." https://arstechnica.com/information-technology/2023/05/snoop-dogg-on-ai-risk-sh-what-the-f/

does not show that that way is the same as a human's. The system could be functionally or internally very different despite having been found to be behaviorally similar. And because of this, its future performance could diverge significantly from the typical outputs of genuine understanding.

This doesn't involve assuming that all the systems required for understanding must be like the human system. We can leave open the possibility that animal and alien beings that are capable of understanding language are internally, materially, and perhaps even functionally different. Nor am I assuming that there is no way that a future AI can achieve genuine understanding. The main point is, rather, that there is no good reason for us to think that all systems that are behaviorally like systems capable of genuine understanding will be functionally like genuine understanding.

(However, we can understand why Dave might feel dissatisfied with this. If chatGPT understands none of it, how can it respond in such a specific, structured, and apparently meaning-responsive way? More on this in the next section.)

Unlike racism and sexism, in the case of anthropomorphism the direct targets of bias are machines and not humans. While the danger in the case of bias against humans by proxy is that our actions still treat someone who is in a protected category in ways that lead to disparate or discriminatory treatment, the danger in the case of anthropomorphism is that our actions treat something that is *not* human in certain ways as we would treat a human interlocutor, with potentially serious consequences to us and other people. While we might not worry about the effect on the chatbots' possible rights or wellbeing (at least not at the current stage of AI development, and arguably not any time soon), we might very well worry about the effect on us, how we understand our interaction with it, and whether, when, and to what extent we should trust it and rely on it.

These distinctions between anthropomorphisms help explain how people can be anthropomorphizing in one sense but not another at the same time: they don't believe that the output is of a human or a conscious mind, and so are not tempted *in that way* to believe that the output has other features characteristic of human output (e.g., being trustworthy or accurate). They believe, correctly, that the output has certain features that are superficially human-like (being *apparently* coherent, fluent, responsive in sensical and complex ways, specific, etc.), and they then take these cues or signs as good indicators of other features (e.g., having states of unconscious understanding, states that are structurally or functionally like understanding or like other mental states and processes). This could lead them in turn to treat some technology as more trustworthy or accurate than it is, a kind of automation bias, or perhaps more untrustworthy or manipulative than it is, a kind of automation phobia.

Note how the point is not just that it's possible to apply some human-like properties without applying others—e.g., applying "understanding" but not "consciousness", or applying "reasoning" but not "emotion". The point is rather that, in some cases, applying one property supports or encourages acting as if the other property applies, and so has some of the same consequences. Treating an AI as though it simulates understanding, tracks the meanings of words and the logic of our inferences, etc., strongly encourages acting, in many ways, as if it really does understand. David's saying "even if I'm wrong about whether it really simulates understanding, at least I don't think it really does understand" isn't quite as harmless and unproblematic as it sounds.

Whether some case of anthropomorphism is indeed a bias will depend on whether and to what extent it disposes one to diverge from appropriate norms. My main point in discussing anthropomorphism is not to insist that there exists widespread, problematic, biased anthropomorphism, but to introduce some interesting varieties of it, and raise it as a serious risk (including risk of automation bias and automation phobia). This should help us think of possible risk management and mitigation strategies—strategies that go beyond transparency about the general type of technology used and general directives not to anthropomorphize, which are unlikely to help much given the possibility of implicit and truly implicit bias.

3. Explaining Coherence, and the Hijacking of Rationality

Like other cases of implicit and truly implicit bias, implicit and truly implicit anthropomorphism tends not to be easy to dislodge, correct, or block from practical deliberations. If it is due to a single and isolated mistaken belief that is easy to correct—say, the belief that AI merely reproduces copied human text when in fact it generates text—then the fix might be easy. Once the false belief is corrected, down-stream errors can be avoided. Here's an analogy: once I've corrected my false belief about what some symbol on the dashboard display of my car means, I avoid being systematically wrong about how my car is fairing and whether I need to take it to the mechanic. But, unfortunately, sometimes the problematic beliefs are not so easy to correct. There could be pressures—in some ways *very rational pressures*—to keep the proxy belief. Let's discuss some of these pressures in more detail.

A high degree of coherence of linguistic representation, or the appearance thereof, is arguably a very good heuristic, indeed practically indispensable, for identifying human or intelligent output. Language depends on conventional signs, where the meaning we attach to the items in language are not a matter of some natural disposition of the signs themselves but a matter of arbitrary conventions for the use of symbols or syntax. When these bits of syntax come together in ways that we can readily read, understand, and make coherent sense of, this calls out for explanation. And the most natural explanation is that someone who *understands* the language, more or less as we do, is the producer or at least the ultimate source of the text. We can easily make sense of a simple program that, after being fed some such text, is tasked with finding obvious spelling mistakes and fixing them, and similarly with the more complex task of finding relatively straightforward mistakes of grammar. It is, after all, comparatively easy to see how one might program a machine to find such errors, many of which diverge from canonical syntactical forms. But our friend Dave is right to think it's quite another matter to explain the performance of chatGPT.

Our trust in human testimony depends heavily on the very rich communicative experiences we have had with them, experiences that can best be explained by ascribing certain experiential, cognitive, and affective states and capacities to other humans. Other humans also look like us, and that counts for something, but their being like us in capacities of communication is independently, epistemically very significant. If we found a rock on another planet with marks on it, and after analysis found that when we interpreted the marks a certain consistent way they seem to tell a detailed and coherent story, that would, other things being equal, be some significant evidence that the story was produced by a being with understanding.

So, we normally have good reasons to take the apparent coherence of a specific and complex text displayed in the syntax of a natural language as indicative of a human or genuinely intelligent, understanding source. This is of a piece with deeply engrained and quite rational disposition to take coherence as something that *needs to be explained*—the apparent coherence of or coordination between our different senses, the coherence between what we intend to do and what we observe ourselves doing, the coherence between our perceptions and memories, the coherence of independent observations and studies in the sciences, and so on. Recognizing a high degree of coherence between different bits of data within or across different modalities (sensory, mnemonic, introspective, observational, linguistic, etc.), when they could easily fail to stand in such coherent, patterned or structured relations, calls out for explanation.

Of course, there are explanations compatible with a relatively high degree of falsehood in the text. The source might be a bullshitter who doesn't care particularly about truth or honesty as such, but is for some reason interested in getting you to believe their message. Or it might be intended as a piece of fiction, to delight and amuse rather than deceive. In the context of watching a play that we know is not real, we can understand the actors as intending to make-believe, and we allow ourselves to make-believe as well. No harm here, since we can clearly distinguish the contexts and not keep believing, as a very young child might, of the reality of what they had just seen. In the context of someone we know is a bullshitter, or of a social context that we are convinced involves a heavy dose of fake news and unhinged conspiratorial thinking, we have no difficulty dismissing it all. But in the context of LLMs, someone like Dave is not tempted to think such defeaters are present: there is no intention to make-believe on the part of the AI, no wild, politically motivated conspiratorial thinking—though LLMs are ripe for abuse and misuse by individuals and groups with such intentions or motivations! More importantly, these explanations all assume that the producers of the text (the bullshitter, the actor, the conspiracy theorist) *understand* the language, and that understanding is part of the explanation of their ability to use it for their intended purposes.

So, our friend Dave lacks an alternative explanation for what he is experiencing. But he might accept, on the AI expert's authority, that there is a complex explanation that involves no genuine or real understanding at all. Still, in the absence of an alternative explanation, it will be very difficult for Dave to resist believing that chatGPT:

- simulates understanding
- simulates conversational competence

- has a detailed model of the world that our language is about
- tracks the meaning or reference of our words and sentences
- tracks the inferences and reasoning involved in our use of language
- etc.

We do have an explanation that, at least at a high level of generality, does in fact account for the surprising success of LLMs. I won't provide a detailed explanation here. But we don't actually need to get into much technical detail to get the general idea. We can even imagine that Dave asked chatGPT to offer some explanation, and chatGPT obliges: "As an AI language model, I'm trained on vast amounts of text data, learning patterns and structures in language. I don't 'understand' language like humans do; instead, I predict the most likely words or phrases to follow based on context. My responses are generated by utilizing these learned patterns, giving the illusion of understanding, even though I lack true comprehension.'[10]

That's a decent start.[11] We can help Dave by adding: Roughly speaking, ChatGPT is sort of like "auto-correct on steroids."[12] It is trained to make text predictions by feeding it a lot of data (45 gigabytes worth—equivalent to a few million pages of text), and adjusting its internal rules, the weights and balances in a very complex internal structure (neural network), to predict how some given text will continue. These weights and balances can be represented by a very complex mathematical function with literally billions of parameters or variables (GPT 4 has 1 trillion parameters). It starts off being very bad at predicting text continuation, but it improves its accuracy by repeatedly tweaking its weights in light of its previous performance. The model can thus be understood as representing the likelihood, given the training data, that a string of text is followed by some other string of text. The data is unfathomably large, and there's enough variation in it to develop a very complex and sensitive model, one that is likely to yield a string of symbols similar to one that a human would produce. Importantly, the sheer size of the model also makes likely that there are many wildly different functions or structures possible that achieve apparently similar levels of performance.

This still leaves lots of interesting details out. But this sort of explanation might help Dave understand, at least in the abstract or in a general sort of way, how chatGPT works. However, this does not translate to an explanation of the specific behavior and capacities of models like chatGPT—for example, it does not explain how chatGPT responds to a series of explicit and nuanced instructions as though it understands exactly what you asked it to do, including responding to queries about what certain words and phrases mean. Some of the capacities chatGPT acquired were very surprising even to its developers. To the extent that there is an explanation, it is not an explanation of any particular output or relatively specific features of output. We simply cannot grasp trillions of parameters and see how they work to generate this

[10] This was actually generated by chatGPT (GPT4).

[11] The explanation is misleading in some respects. For example, LLMs like chatGPT don't really work on words and phrases to predict other words and phrases; rather, the basic 'tokens' tend to be sub-words or parts of words.

[12] This comparison was made by Gary Marcus, and repeated in many news articles. See, for example, Kuhlar (2023).

sort of output. And while attempts to arrive at approximate, partial explanations of how it functions are underway in the field of explainable AI or XAI (often using other complex models), it's too early to tell how successful these will be, and these forthcoming models of how LLMs work might not lend much support to the thought that they are at all functionally like reasoning and understanding the meaning and reference of words.[13]

In a brief discussion of GPT-3 on *Daily Nous*, Henry Shevlin shares an apt quote from writer Terry Pratchett to describe our situation: "It's still magic even if you know how it's done." We do know, at least in a general sort of way, how LLMs work, but it still seems magical. Shevlin elaborates on GPT's "mesmeric anthropomorphic effects":

> Earlier artefacts like Siri and Alexa don't feel human, or even particularly intelligent, but in those not infrequent intervals when GPT-3 maintains its façade of humanlike conversation, it really *feels* like a person with its own goals, beliefs, and even interests. It positively *demands* understanding as an intentional system. [Shevlin 2020]

Our inability to really grasp, in a concrete way, how understanding and intentionality can be mimicked in this way encourages the interpretation that it does in fact understand. It may be difficult to resist having an implicit belief that the model understands, at least in some cases and as AI continues to develop and get used increasingly in social settings. At some point, one might slip from merely acting as if one's household robot understands, thinks, or feels, to believing implicitly that it does.

But even if we don't go that far, and even we don't confront such cases now, it is hard to resist believing, explicitly or implicitly, that chatGPT works via some *non-human, non-conscious states and processes that are a kind of quasi-understanding, something that functions much like understanding, or that tracks what we understand*. Or we might think that it works *by building a causal model of the world, representing objects, events, and properties of the world and relations between them.* But in fact, we don't have a good reason to think that LLMs have this kind of quasi-understanding, or a model of the world that ordinary language is about. To think that it does is to anthropomorphize chatGPT by proxy.

Dave can resist the slide from believing that chatGPT mimics understanding and rationality to believing that it simulates it. But it may be very difficult to resist. The sort of apparent clarity, fluency, and the coherent structure of the text can be seductive.[14] When I respond to someone in language, I respond by understanding what they are saying or asking, thinking about it, and responding to that – I don't consider the likelihood of that string of signs being followed by other kinds of signs. If the AI is responding to a question, it is not responding to the question we ask it, but to "what is the statistically most likely next string of symbols given that one?"[15] But

[13] For some very helpful discussions of XAI, see Fleisher (2022) and Mittelstadt et al. (2019).

[14] See Nguyen (2021) on "The Seductions of Clarity." Though Nguyen focuses on seductively clear systems like conspiracy theories and quantified value systems of bureaucracies, the text of LLMs can also give rise to feelings associated with understanding, feelings of clarity and fluency, that can be misleading.

[15] See Shanahan (2022) for a discussion of this point.

even this way of putting it is strongly anthropomorphic. It is arguably not responding to any questions, at least not in any sense of “responding to questions” that we do in normal language—not even to questions about the statistical likelihood of different events. In this respect, it’s not even like a person who has internalized all sorts of patterns of syntax in a language without understanding the language.[16]

It is very difficult to resist applying words like “know”, “think”, “understand”, “reason”, “believe”, “feel”, “want” etc. to describe AI. Indeed, it is a practically indispensable means of making effective use of the technology, writing effective prompts, anticipating its behavior, describing its errors, and communicating to others about these systems. We can consciously deny that we use any of these literally or in the same sense as used with humans, but what understanding of these terms are we applying? There’s a danger here again of anthropomorphizing, a danger from which AI practitioners and developers are not entirely immune.

Shanahan puts the central problem nicely:

> Insofar as everyone implicitly understands that these turns of phrase are just convenient shorthands…it does no harm to use them. However, in the case of LLMs, such is their power, things can get a little blurry. When an LLM can be made to improve its performance on reasoning tasks simply by being told to “think step by step” (Kojima et al. 2022) (to pick just one remarkable discovery), the temptation to see it as having human-like characteristics is almost overwhelming.
>
> ….such systems are simultaneously so very different from humans in their construction, yet (often but not always) so human-like in behavior, that we need to pay careful attention to how they work before we speak of them in language suggestive of human capabilities and patterns of behavior. [Shanahan 2022: 3]

The core of the problem is that many who avoid believing that AI are sentient, sapient, or sentimental, will be inclined to form beliefs and associations that are proxy for these sorts of features, and so treat AI in significant ways as though they have these features. In the face of interacting with sophisticated LLMs, it will be difficult for the ordinary subject, or indeed anyone, to resist the belief that it is tracking or modeling our linguistic understanding, coherence, and inference.

4. Concluding Thoughts

[16] Compare the person in the Chinese Room in Searle’s (1980) famous thought experiment, who responds to the input of Chinese text he doesn’t understand by output of Chinese text he doesn’t understand in accordance with rules that he does understand, rules telling him that when given such-and-such symbols he should respond with such-and-such other symbols. Internalizing or memorizing all the rules that he does understand for patterns of input and output won’t help him understand Chinese.

The dangers of LLMs are similar, in certain ways, to the dangers of echo chambers and conspiracy theories, and the spread of misinformation on social media. It would arguably be a mistake to think that the spread of misinformation is due *primarily* to gross incompetence and stupidity, irrationality, willful disregard for the importance of truth, or even the influence of systematic subconscious biases. There is no need to invoke the idea that victims of conspiracy theories or members of echo chambers are irrational or intellectually vicious. Rather, the main problem is that the informational environment has been polluted with so much informational sewage. Various mechanisms and structures limit and filter the source and nature of the subject's information, with the result that information most readily available to the subject is highly coherent. Conspiracy theories and echo chambers can work more effectively on individuals if, rather than preying on their irrationality, they hijack their rational capacities, including cognitive methods that are normally reliable, and deeply engrained. They can lead us to accept misinformation by hijacking core features of ordinary rationality.[17]

Similarly, a significant problem with AI-generated text is that there are *rational* pressures to anthropomorphize. As we have seen, trusting the AI-generated text might not be rational *all things considered*—at least for those of us who have defeaters or good reasons not to trust it, reasons to think that in this case we should take the apparent coherence with a huge grain of salt, and not think the AI has states that are even approximately functionally like our own states. But I fear that few people will recognize these sorts of defeaters, some of which can be subtle (like the difference between mimicry and simulation), and that fewer still will remember them while deeply engaged with an advanced LLM.

There is thus a significant risk that, whether or not we explicitly and strongly anthropomorphize AIs, we will anthropomorphize them in a more subtle way, believing them to have properties that are in fact strongly correlated with having consciousness, understanding, and various other sorts of mental states. This may ultimately encourage more explicit or stronger forms of anthropomorphism, especially as the LLMs get larger and more sophisticated.

In her illuminating article exploring the relation between machine and human cognitive bias, Johnson says that the "insight that biases (cognitive and algorithmic) might operate using proxy attributes has important implications for mitigation techniques in both domains" (2020: 16). The same is true in the case of (biased) anthropomorphism and automation bias. The concept is helpful in making clear that avoiding anthropomorphism of the most obvious sorts still allows for a problematic interpretation and treatment of AI, treatment that in some respects is very similar to how someone who does explicitly anthropomorphize AI might treat it.

Transparency of AI use and abstract knowledge of how it works is unlikely to be effective in mitigating (biased instances of) anthropomorphism. We may need to develop new heuristics and tools to encourage and support vigilance on the part of users, and not merely assume that

[17] See Levy (2022) for a book-length defense of such a view. For related discussion of echo chambers (epistemic bubbles, etc.), see Nguyen (2020) and Elzinga (2020).

we'll adjust appropriately over time.[18] We need to find ways to assess and minimize risk of anthropomorphism and other biases involving technology, and develop systems of AI governance and regulatory guardrails informed by an understanding of these risks. My focus here was not to discuss such possible tools and methods, but to bring attention to some significant risks in this area that are easy to overlook, and to discuss important distinctions between forms of anthropomorphism that can help us identify, categorize, and assess these risks.

References

Baron, Jonathan (2012). "The Point of Normative Models in Judgment and Decision Making." *Frontiers of Psychology* 3:577. https://doi.org/10.3389/fpsyg.2012.00577

Beeghly, Erin & Madva, Alex (eds.) (2020). *An Introduction to Implicit Bias: Knowledge, Justice, and the Social Mind.* Routledge.

Bender, Emily M, Timnit Gebru, Angelina McMillan-Major, & Shmargaret Shmitchell (2021). "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 610–623. https://doi.org/10.1145/3442188.3445922

Brownstein, Michael (2019). "Implicit Bias" *The Stanford Encyclopedia of Philosophy* (Fall 2019 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/fall2019/entries/implicit-bias/>.

Fazelpour, Sina & Danks, David (2021). "Algorithmic Bias: Senses, Sources, Solutions." Philosophy Compass16 (8):e12760.

Fleisher, Will (2022). "Understanding, Idealization, and Explainable AI." *Episteme* 19 (4):534-560.

[18] Compare Nguyen's discussion of seductively clear systems like conspiracy theories and quantified value systems of bureaucracies:

> In fighting the seductions of clarity, we need to develop new counter-heuristics…. The sense of clarity is something like cognitive sugar. Once upon a time, using our sense of clarify as a signal to terminate our inquiries might have been a good and useful heuristic. But now we live in an environment where we are surrounded by seductive clarity, much of it designed to exploit our heuristics. We now need to train ourselves to become suspicious of ideas and systems that go down just a little too sweetly – that are pleasurable and effortless and explain everything wonderfully. Systems of thought that feel too clear should make us step up our investigative efforts instead of ending them. [2021: 251]

Holroyd, Jules (2017). “Responsibility for implicit bias.” Philosophy Compass 12 (3).

Johnson, Gabbrielle M. (2021). “Algorithmic bias: on the implicit biases of social technology.” *Synthese* 198 (10):9941-9961.

Kelly, Thomas (2022). *Bias: A Philosophical Study.* Oxford University Press.

Kim and Sundar (2012). “Anthropomorphism of Computers: Is it Mindful or Mindless?” *Computers in Human Behavior* 28: 214-151.

Kojima, Takeshi, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, & Yusuke Iwasawa (2022). "Large language models are zero-shot reasoners." *arXiv preprint arXiv:2205.11916*.

Kulhar, Dhruv (2023). “Can A.I. Treat Mental Illness.” *The New Yorker.* Feb. 27, 2023. https://www.newyorker.com/magazine/2023/03/06/can-ai-treat-mental-illness

Levy, Neil (2021). *Bad Beliefs: Why They Happen to Good People*. Oxford University Press.

Mandelbaum, Eric (2016). “Attitude, Inference, Association: On the Propositional Structure of Implicit Bias.” *Nous* 50 (3):629-658.

Mittelstadt, Brent, Christ Russell & Sandra Wachter (2019). “Explaining Explanations in AI” In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* ’19).* Association for Computing Machinery. https://doi.org/10.1145/3287560.3287574

Nguyen, C. Thi (2020). “Echo chambers and epistemic bubbles.” *Episteme* 17 (2):141-161.

Nguyen, C. Thi (2021). “The seductions of clarity. *Royal Institute of Philosophy Supplement* 89:227-255.

Pasquini, Maria (2023, March 13). “Man dies by suicide after converstions with chatbot that became his ‘confidante,’ window says.” *People.* https://people.com/human-interest/man-dies-by-suicide-after-ai-chatbot-became-his-confidante-widow-says/

Scarantino, Andrea and Ronald de Sousa (2021). “Emotion”, *The Stanford Encyclopedia of Philosophy* (Summer 2021 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/sum2021/entries/emotion/>.

Searle, John (1980). “Minds, brains, and programs.” *Behavioral and Brain Sciences* 3 (3):417-57.

Shanahan, Murray (2022). “Talking About Large Language Models” *arXiv preprint arXiv:2212.03551*.

Shevlin, Henry (2020). “A Digital Remix of Humanity” *Daily Nous*, July 30, 2020.

https://dailynous.com/2020/07/30/philosophers-gpt-3/

Sobieszek, Adam & Price, Tadeusz (2022). "Playing Games with Ais: The Limits of GPT-3 and Similar Large Language Models." *Minds and Machines* 32 (2):341-364.

Tankersley, Jim (2021, Jan. 31) "Black Americans are much more likely to face tax audits, study finds." *The New York Times*. https://www.nytimes.com/2023/01/31/us/politics/Black-americans-irs-tax-audits.html

Tiku, Nitasha (2022, June 11). "The Google engineer who thinks the company's AI has come to life." *The Washington Post.*

Weidinger, Laura et al. (2021). "Ethical and social risks of harm from language models." *arXiv preprint arXiv:2112.04359*.