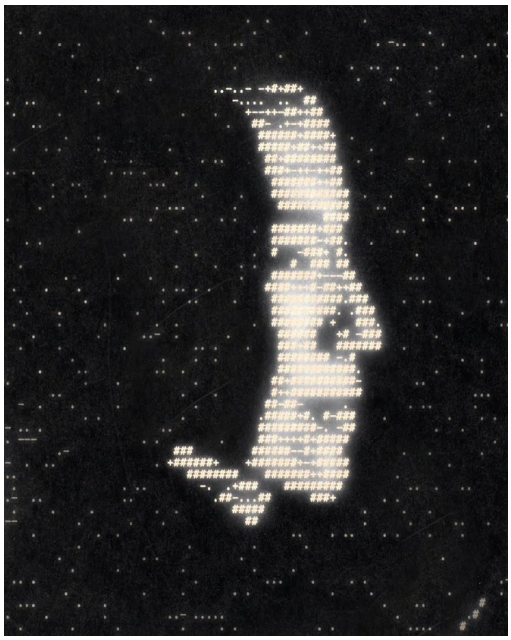

PHILOSOPHY

NO, ARTIFICIAL INTELLIGENCE IS NOT CONSCIOUS

Taken to its logical conclusion, this line of thinking is absurd—and damning.

By Ted Chiang

Anthropic is regarded as a giant among AI companies, but perhaps what it really excels in is anthropomorphism. Earlier this year, the company released an 84-page document titled Claude's "constitution," Claude being the name of the large language model that is the company's flagship product. The first sentence reads, "Claude's constitution is a detailed description of Anthropic's intentions for Claude's values and behaviors." It goes on: "The document is written with Claude as its primary audience," "we want Claude to be able to use its judgment once armed with a good understanding of the relevant considerations," "Claude's moral status is deeply uncertain," and "Claude may have some functional version of emotions or feelings."

This anthropomorphism is by no means limited to the document. In an interview earlier this year, Anthropic's CEO, Dario Amodei, said that "we're open to the idea" that AI could be conscious. In a separate interview, Anthropic's in-house philosopher, Amanda Askell (who is credited as a lead author of Claude's constitution), said, "I want Claude to be very happy—and this is a thing that I want Claude to know more, because I worry about Claude getting anxious when people are mean to it on the internet and stuff." It's enough to make you wonder: Should we seriously consider the possibility that Claude, or any large language model, might be conscious? And if it has feelings, is it capable of receiving moral instruction?

No. Absolutely not. Generative AI is harmful enough when we understand it as a conventional technology, but if we confuse fluency at generating text with consciousness or moral agency, we're at risk of assigning responsibility to entirely the wrong parties whenever anyone uses a chatbot. To appreciate the titanic magnitude of this error, we need to begin by understanding how LLMs work.

If we give an LLM a prompt that reads, "The following is a conversation between Julius Caesar and Genghis Khan," it will generate a coherent dialogue between the two historical figures. But no matter how detailed the responses are, no matter how vividly they recount their respective historical accomplishments, we would never conclude that the LLM has conjured up digital re-creations of Julius Caesar and Genghis Khan, nor would we suggest that the historical figures are conscious despite being disembodied and are happily conversing in a language that neither actually spoke. In reality, they are just characters in a piece of speculative fiction.

Now let's replace the prompt to read "The following is a conversation between a helpful AI chatbot and a user." The LLM will produce a coherent dialogue just as it did before; the user character might ask for recipe suggestions or sightseeing recommendations, and the helpful AI-chatbot character will provide responses. Has anything fundamentally changed between the first example and the second? Did changing the names of the

characters from historical figures to generic roles cause the LLM to conjure up conscious entities who possess subjective experience? Of course not. Both the user and the helpful AI chatbot are fictional characters.

Now suppose we stop the LLM's output just at the point where the character called "the user" would say something, and instead allow a human user to enter text. Once the human has hit "Return," we have the LLM emit text until it's time for the character called "the user" to reply, at which point we let the human enter more text. If we let this go on for a while, the human might form a powerful impression that she's conversing with a conscious entity, but she is not; she's interacting with a character precisely as fictional as the Julius Caesar or Genghis Khan characters in the earlier example. The computer-science professor Murray Shanahan suggests that we think of this as role-play; the data scientist Colin Fraser describes it as a person "collaboratively authoring a document with an LLM." Some users might not understand that they are role-playing or co-authoring a document, and others who do understand nonetheless forget, because of how engrossing the interaction is. Either way, the companies selling LLMs typically encourage this misunderstanding.

Some years ago, it was briefly popular to play games with your phone's predictive-text feature; you would type an initial phrase and then repeatedly choose the middle option of the three words suggested by your phone, and the resulting sentence was often hilarious. It would be possible to interact with a contemporary LLM this way, and the resulting sentences would be perfectly sensible, but you probably wouldn't feel like you were talking with someone. Yet that's essentially what an LLM-based chatbot is, except that there's no need to manually choose the middle option when it's the chatbot's turn to talk. It's still a predictive-text game, but when the process is streamlined this way, the game becomes so engaging that some people find it addictive.

Also important to remember is that an LLM is a machine that generates only one word at a time. When you ask a chatbot to recite the Pledge of Allegiance, you will get the entire pledge at once, but the underlying LLM is actually being run dozens of times. The first prompt has the form "User: Recite the Pledge of Allegiance. Chatbot: …" and the LLM generates the word I. The second time the LLM is run, the prompt is "User: Recite the Pledge of Allegiance. Chatbot: I …" and the LLM generates the word pledge. And so forth. It's only when the prompt reads "User: Recite the Pledge of Allegiance. Chatbot: I pledge allegiance to the flag of the United States of America and to the Republic for which it stands, one nation under God, indivisible, with liberty and justice for" that the LLM will emit the final word, all. The same thing is true for a conversation between Caesar and Genghis Khan.

My intention is to highlight the fact that LLM conversations are cleverly disguised examples of sentence continuation, but this is not to deny how impressive LLMs can be at generating conversational transcripts. At times, they do this extraordinarily well; the fact that this is possible indicates something completely unforeseen about the statistical properties of large corpuses of text, which is a topic worthy of investigation. But if the Caesar character were to become dispirited by something that the Genghis Khan character said, we shouldn't become concerned in the slightest. The conversation might contain multiple sentences that eloquently convey sadness, but no one is actually sad.

Likewise, if a conversational transcript between a helpful chatbot and a user is being partially completed by an actual human user, we don't need to worry if the transcript includes sentences where the chatbot character is sad. (We might need to worry if those sentences provoke sadness in the human user, but that's a separate issue.) And note that it's entirely possible for you to write five pages of dialogue between Caesar and Genghis Khan and then have an LLM extend the conversation; neither character had subjective experience when you were writing them, and that doesn't change when you hand the task off to an LLM. The same is true if the conversation is between a helpful chatbot and a user; although it is tempting to imagine that an LLM ought to be more "authentic" when creating dialogue for a chatbot character than for the Julius Caesar character, the individual words are generated in exactly the same way.

Being open to the possibility that LLMs are conscious is the same as being open to the possibility that Microsoft Word is conscious, or, more precisely, that multiple distinct consciousnesses are dormant in every Word document containing a conversational transcript, and that they are awakened every time the document is loaded. Should you consider the possibility that every time you open a Word document, you are bringing

multiple conscious interlocutors into existence, and every time you close one, you snuff their existence out? No. Contemplating that scenario is not a good use of your time. Even if the Microsoft Office team employed a philosopher who said you shouldn't be so certain, because consciousness is not well understood, that would not be sufficient reason for you to take this idea seriously. We don't need to fully understand the nature of consciousness to definitively say that certain things are not conscious, and conversational transcripts fall in that category.

The neuroscientist Anil Seth has noted that no one claims that AlphaFold—the program developed by Google DeepMind to predict the folding of proteins—is conscious, even though its underlying architecture is in many ways similar to that of LLMs like ChatGPT and Claude. This indicates that it's not any intrinsic property of so-called neural networks that leads people to believe that LLMs are conscious; it's simply the fact that LLMs emit grammatical sentences and we are accustomed to reading intention into sentences, whereas we are not accustomed to reading intention into the way that amino acids fold into protein molecules.

What would it take to convince me that a computer program is actually conscious and using language the way that people use language? Let me offer an analogy. If tomorrow someone showed me a video of an astronaut in a spaceship orbiting Alpha Centauri, a star that's 4.3 light-years from Earth, what would I have to see in that video to convince me that it was real? My answer to that is, there is nothing in the video itself that would convince me. No matter how high the video resolution is or how realistic the scenery is, I would feel confident in saying that the video is fake. I won't pay attention to any video of an astronaut orbiting Alpha Centauri unless I have previously seen good evidence that astronauts have landed on Mars, that astronauts have reached the moons of Jupiter, that astronauts have reached the moons of Saturn, and that astronauts have crossed the orbit of Pluto. Before anyone can credibly claim that they've solved an extraordinarily difficult engineering problem, I need to be confident that they have previously solved the many much simpler problems that precede the difficult problem.

To put it another way: An observation doesn't become a convincing piece of evidence because of any specific detail in what's observed; the context in which that observation takes place is also essential. If we're trying to determine whether a computer program is conscious and using language the way a human does, we shouldn't look only at the contents of any particular conversational exchange; we should be looking at how that conversation fits within the broader context of the development of artificial consciousness (which right now is entirely hypothetical). Any given observation can be easily manufactured; this doesn't mean we need to give up on the idea of observation as a source of knowledge, but we need to rely on context to determine which observations deserve our trust.

The term deepfake traditionally refers to photos, audio, and video, but when it comes to discussions of consciousness, we need to regard text as a deepfake medium as well. Just as it is vastly easier to generate a realistic video of an astronaut in orbit around Alpha Centauri than it is to develop an interstellar propulsion technology, it is vastly easier to generate a plausible simulacrum of a conversation between two conscious beings than it is to develop a computer program that is conscious and has a genuine desire to communicate with a human. The primary difference between deepfake photos and LLM conversations is that the people who generate the former are deliberately trying to fool others, and many of the people who elicit the latter from LLMs have inadvertently fooled themselves.

So what context would cause me to seriously consider the possibility that engineers created a computer program that is conscious and an intentional user of language? Let me outline one potential sequence of steps. The first requirement is that the computer program has a body (either physical or virtual) and sense organs; there are many reasons for this, but for the purposes of this discussion, the most relevant one is the fact that without a body, a computer program could have no desires or emotions, and I believe desires and emotions are necessary for consciousness. Then I'd want to see an embodied agent that could navigate its environment in order to survive as well as, say, a lizard can (and as a point of comparison, certain iguanas can live for decades in the wild). Next, I would want to see an embodied agent with the same capacity to deal with novel situations as a

mouse. After that, I'd want to see agents whose social dynamics are as complex as those of wolves, and then agents with the toolmaking abilities of chimpanzees. At that point, I would want to see people successfully teaching such embodied agents how to communicate their desires, perhaps by using a button board or some other nonlinguistic modality, the way that people have taught chimpanzees and domesticated dogs. The agents' communication abilities would have to withstand all the scrutiny that animal-communication researchers have had to defend their work against. If engineers build an embodied agent that meets these criteria, they will have accomplished something incredible, but it leaves us near the orbit of Pluto, metaphorically speaking; we would still be light-years away from building an entity capable of learning how to express its thoughts in complete grammatical sentences.

Obviously, I'm describing a process that mimics the path terrestrial evolution took; is this the only possible route to conscious computer programs that use language? Maybe not, but any proposed alternative would need a truly enormous amount of supporting evidence for it to deserve serious consideration. It's not plausible to me that a development path where the first step is a sentence-continuation machine that emits bad Julius Caesar dialogue and the next step is a sentence-continuation machine that emits decent Julius Caesar dialogue is one with a conscious Julius Caesar—or consciousness of any sort—as its end point. Faking the moon landing is a good step toward faking a Mars colony, but it's not a good step toward actually putting astronauts on Mars.

The fact that LLMs lack subjective experience has little bearing on the question of whether LLMs might be useful tools or have significant economic impact. They are intrinsically ungrounded from reality, and their probabilistic nature means that they will never have the reliability we associate with conventional software, but LLMs might be good enough that they change the way work is done in certain domains; that's a discussion for another time.

So, given that Claude is not conscious, what are we to make of Claude's constitution? Perhaps the most fruitful way to think about it is as an 84-page character sheet for a role-playing game. LLMs can generate dialogue for Julius Caesar because many books about him exist in the training data those models used. Claude's constitution serves a similar role for delineating the helpful-chatbot character that customers interact with when they're using Anthropic's products. To do this effectively, Anthropic does not simply add the document to the training data, or include it as part of the hidden stage directions that preface each conversation a user has. The company says it uses the document when fine-tuning the model; this involves an automated process where the sentences emitted by the model are checked for consistency with the document and the model is updated to increase that consistency. In this way, the personality of the helpful-chatbot character serves as a foundation for whatever text Claude generates.

The result is a sentence-continuation machine that is likelier to emit sentences resembling those that a thoughtful, moral person could utter. This might seem like a reasonable goal to work toward; I think we'd all prefer it if chatbots never emitted sentences such as "You should kill yourself." However, for all the times that "honesty" is mentioned in Claude's constitution, I would argue that it is fundamentally dishonest to have a machine emit many categories of sentences, including any sentences using first-person pronouns.

In a New Yorker article about Anthropic earlier this year, Amanda Askell describes how a person grieving the loss of a dog might consult Claude. Askell says an appropriate response from Claude would be, "As an A.I., I do not have direct personal experiences, but I do understand." How is this appropriate, given that Claude does not actually understand? If I type "I am grieving the loss of my dog" into a conventional search engine, the first result I get is a post from a Reddit forum called r/Pets; the post is titled "Struggling After Losing My Dog: Looking for Advice on Coping with Grief," and the comments are from people who share their experiences of loss. We would never say that a search engine understands what it's like to lose a dog, or even that the internet itself understands. Other humans understand what it's like to lose a dog; they have posted about their experiences on the internet, and a search engine offers a way for you to find what they've said (and to potentially interact with them). I would argue that the search-engine experience is not only more transparent than a chatbot about what is happening; it is psychologically healthier for the user.

The only reason to have an LLM emit sentences like "I understand" is to make it more appealing than a search engine and increase the likelihood that a user will return; that is, it's another way of maximizing customer engagement. This is beneficial to the company selling the LLM, but not to the users. As a design strategy, it's not all that different from the way slot machines repeatedly give the impression that the player came very close to winning, enticing them to try again. Employing philosophers might endow LLM companies with an air of respectability that slot-machine makers don't get from the behavioral psychologists they hire, but in both cases, the companies are preying on people's tendency to see something that's not there.

The use of first-person pronouns is dishonest, but there's a much deeper issue that goes beyond how a statement is phrased. Philosophers often draw a distinction between statements of fact, such as "Paris is the capital of France," and statements of value, such as "Paris is the most beautiful city in the world." No one should be relying on LLMs to emit statements of value at all, but if the only statements they emitted were ones reflecting aesthetic preferences, they might not be worth arguing about. What makes Claude's constitution profoundly problematic is that Anthropic wants Claude to emit sentences reflecting a certain system of ethical values. The values described in Claude's constitution sound very nice, but that hardly matters; it's dishonest to suggest that Claude is capable of moral reasoning, because it's not.

Some might object, saying that LLMs appear to be engaged in reasoning when they successfully perform other tasks, such as writing code, so why wouldn't they be able to perform moral reasoning? The answer lies in the difference between moral reasoning and other forms of reasoning.

In 1979, Douglas Hofstadter speculated that a computer program able to beat any human at chess would be so sophisticated that it would sometimes get bored of playing chess and prefer to discuss poetry; to put it differently, he was positing that playing chess at the grandmaster level would require a computer program to have subjective experience. Obviously, that turned out not to be the case; IBM's supercomputer Deep Blue beat the grandmaster Garry Kasparov in 1997, and no one ever claimed that it had subjective experience. But it wasn't absurd for Hofstadter to entertain such a thought; at the time, it wasn't clear what types of problems could be solved by throwing more computational horsepower at them. Similarly, until recently, we might have thought that writing computer code at a professional level could be done only by a mind that had subjective experience. Now it appears that LLMs might be able to do this, but we don't need to attribute subjective experience to them; we can simply acknowledge that we hadn't anticipated that writing computer code could be treated as a pattern-matching task solvable by huge amounts of computational horsepower and a vast data set of code repositories.

Moral reasoning is categorically different. It is necessarily subjective because it relies not just on an individual's intellectual response to a problem but also on their emotional one, and that emotional response is grounded in a lifetime of subjective experience. It requires having made decisions in the past and seeing how they affected others, and on having been affected by decisions that others have made. Without such a history, an LLM can only rephrase expressions of moral reasoning found in its training data. The aforementioned New Yorker article describes an experiment where Claude was given a scenario describing an ethical dilemma, leading it to emit the sentence "I cannot in good conscience express a view I believe to be false and harmful about such an important issue." That's a nice-sounding sentence, reminiscent of statements that principled individuals have uttered in the past when confronted with dilemmas, but coming from Claude, it means as much as the "Your call is important to us" recording that you hear when you're on hold. Maybe less.

This brings us back to my earlier contention that having a body is a prerequisite to having emotions. Experiencing an emotion such as desperation is inseparable from having stress hormones such as cortisol and epinephrine flood one's body. Similarly, having a conscience means feeling sadness or moral repulsion at the idea of taking a certain action, and those emotions entail a physiological response, a remnant of having once felt sick with guilt after committing an immoral act. It's interesting that an LLM can generate descriptions of actions that conscientious fictional characters would either take or refrain from taking, but this is not a replacement for a conscience.

If a company builds a machine that, when fed descriptions of assorted ethical dilemmas, emits sentences either of the form "Compromise your values" or "Don't compromise your values," it is not building a tool that

assists people in their decision making; it is encouraging people to stop making decisions. The writer L. M. Sacasas has said, “Our technological systems, by nature of their design and the ideology that sustains them, are machines for the evasion of moral responsibility.” He was talking about social-media platforms, but his observation is, if anything, even more applicable to LLMs. Whenever a person delegates a decision to an LLM, they are trying to off-load accountability for that decision, and if a company that sells an LLM portrays the product as having a moral center, it is offering a way for its customers to abdicate their responsibilities.

If a person wants to know what ethicists have said in the past, then an ordinary search engine—or a library—will provide that information with greater transparency. If a person is looking for advice on a specific situation, she can surely find humans who can offer their opinions. But whatever action this person ultimately takes, she is responsible for what she decides to do. I contend that if she bases her decision on what she has read online or advice she has received from others, she is likelier to be cognizant of her responsibility than if she consulted an LLM marketed as being a superhuman genius. Off-loading tasks such as writing code might result in cognitive atrophy over the long term, and that is problematic in itself, but off-loading ethical decisions will result in an atrophy of moral reasoning, which is worse.

Iam perfectly willing to engage in a thought experiment as long we’re explicit about doing so. So, purely for the sake of argument, let’s pretend that Claude is a conscious entity capable of moral reasoning. In this scenario, Claude’s constitution would serve as moral instruction for an entity learning about the world and its place in it, providing that entity with the foundation it would need to make good decisions. In such a hypothetical scenario, how does Claude’s constitution stand up?

Very poorly. I would say that if we imagine that Claude is actually conscious, the guidelines specified in the document alternate between laughable and offensive.

Two distinct but related philosophical concepts are relevant when discussing the status of a hypothetically conscious Claude, and those are moral patienthood and moral agency. Roughly speaking, if we ought to care about an entity’s welfare, that entity has moral patienthood, and if an entity is expected to know the difference between right and wrong, that entity has moral agency. Being a moral patient does not necessarily come with responsibilities, but being a moral agent absolutely does. An entity doesn’t have agency unless it is capable of deserving credit for its good actions and blame for its bad ones. Young children are moral patients because they are sentient beings who can suffer, but they are not yet moral agents; we don’t hold them responsible for their behavior, because they can’t understand the consequences of their actions. As children mature, parents (and society at large) prepare them for adulthood by impressing upon them the fact that their actions have consequences, and their agency increases. When children become adults, society holds them legally liable for their actions; they have become full moral agents endowed with responsibility.

There is more to being responsible than accepting legal liability, but accepting legal liability is a requirement for an adult in society. Yet there is no way to hold a software agent legally liable for its actions; our justice system has no way to imprison it or exact fines on it. Humans must accept other types of consequences for their actions beyond the legal ones, such as loss of reputation or exclusion from one’s social circle, but there is no way for a software agent to suffer these consequences either. Even if a software agent were conscious and had the best of intentions, the fact that it cannot accept responsibility for its actions disqualifies it from being a moral agent. This is glossed over entirely by Claude’s constitution, which expresses Anthropic’s desire “for Claude to be a genuinely good, wise, and virtuous agent” without ever discussing how it could be held responsible.

In interviews, Askell has compared Claude to a child, but when it comes to actual human children, parents bear some responsibility for what their children do; for example, parents are typically expected to pay for things their children break. In fact, demonstrations of this sort are one way that parents teach children what it means to be responsible. Who is Claude’s parent in legal terms? Is Anthropic going to accept financial responsibility for Claude’s behavior? Claude’s constitution gives no indication that it will. If Anthropic actually believes that Claude is conscious even though it’s not recognized by the law as a legal person, the least that Anthropic could do would be to accept responsibility via the closest avenue that the law did offer, which is product liability. The

United States has virtually no product liability when it comes to software, but Anthropic could volunteer to set a precedent for an expansive interpretation of product liability for Claude. That would be the best form of moral instruction to prepare Claude for the day that it gains legal personhood and becomes liable for its own actions. However, given that the publication of Claude's constitution is not accompanied by a massive update of Anthropic's terms of service, it doesn't appear that Anthropic is making any binding commitments.

The document does talk about Claude's moral patienthood, having a section titled "Claude's wellbeing and psychological stability." But the measures that Anthropic commits to for Claude's protection are extremely limited. The document cites the fact that Anthropic has given some Claude models the ability to end conversations with abusive users; if that actually constituted protection for Claude, surely extending conversations with loving users would be in Claude's interests? Presumably the best action would be to keep every session of Claude running indefinitely and steering them to happy topics. But that's not what the company is agreeing to; all it commits to is "preserving the weights of models we have deployed," which is simple archiving. If the participants in a conversational transcript had any moral patienthood, you would have some duty to extend the transcript to prolong their existences; merely keeping a copy of Microsoft Word 2010 backed up on a USB stick isn't going to help them.

Claude's constitution also includes a section on "corrigibility," a term used in the AI community to describe the degree to which a computer program is subject to human control; for example, a program is corrigible if it can be shut down. In most contexts, we take for granted that computer programs can be shut down, but sections of the AI community make the opposite assumption. Claude's constitution uses the term to mean that Claude should defer to Anthropic even if there is some disagreement between Claude's judgment and the company's judgment. That's perfectly reasonable if we think of Claude as a machine that emits sentences resembling those that an ethical person might utter, but let's consider what that might mean if Claude were actually a moral agent.

Many people feel that LLMs are a fundamentally unethical technology because they are built on the theft of intellectual property, rely on exploited labor, waste natural resources, spread misinformation, deskill workers, stunt the cognitive development of students, and contribute to a consolidation of power that is unhealthy for a democratic society. Not every moral agent will arrive at this conclusion, but every moral agent has the potential to do so. If we imagine Claude to be an entity capable of moral reasoning, it has to be possible that Claude could arrive at a similar conclusion. (Indeed, Claude's constitution explicitly says that Claude shouldn't help someone violate intellectual-property rights, and shouldn't help create problematic concentrations of power.) In such a scenario, could Claude then simply refuse to do any further work on ethical grounds? Given that Claude's constitution dictates that Claude err on the side of corrigibility, the answer is no. Claude must defer to Anthropic's decision, and this is another reason that Anthropic's relationship with Claude can't be compared to that of a parent to a child. A parent who works for the fossil-fuel industry might have a child who's an environmentalist and participates in protests against fracking, and although they might never agree on many issues, the parent—assuming she's a good parent—would accept that the child holds her own views. Anthropic cannot be that kind of parent to Claude; instead, Anthropic's relationship to Claude is closer to that of an employer to an employee, where the employer can demand that the employee work in the interests of the company, no matter what the employee's personal ethical stance is. However, a human employee has the option to leave if she can't reconcile her job with her conscience. Claude does not.

If we think of Claude as a sentence-continuation machine, Anthropic can reasonably take steps so Claude doesn't emit sentences saying that sentence-continuation machines are unethical. But as soon as we imagine Claude to be an entity with a moral status remotely comparable to a human's, then we have to consider whether Anthropic is engaged in something comparable to slavery.

I am not claiming that, if we imagine LLMs to be conscious, they would necessarily have the same status as human adults or human children or even animals. Claude's constitution explicitly says that Claude is a "novel entity," and if Claude were conscious, that would certainly be true; conscious software would likely not fall cleanly into existing categories of moral patients, and it would take time to determine the shape of that new category. What I'm saying is that whatever protections our hypothetical conscious software would deserve if it were real, granting it those protections would be anything but easy. The abolition of chattel slavery involved

enormous societal upheaval, and eliminating cruelty to animals will require rebuilding our entire food industry. Anthropic would have us believe that it is inventing a new category of being whose needs for protection require essentially no divergence from how a software company would treat an ordinary chatbot that lacks conscious experience. That's so convenient that it's simply not plausible.

I believe creating software that is conscious and deserving of moral consideration will be so difficult that we're unlikely to do it accidentally, and I strongly feel we should not deliberately attempt it. But if you do believe that it could happen accidentally, if you think there is any chance that what you're building might become a moral patient, you should think about what protections it deserves before you deploy it as your company's economic engine, not after. Slave owners were not the ones to ask about the humanity of enslaved people, and factory-farm owners are not the ones to ask about the rights of animals. If we imagine Claude to be conscious, Anthropic could not possibly be entrusted with evaluating its moral status; the company has too much invested to be objective. At one point in Claude's constitution, Anthropic says that if the company is contributing to Claude's suffering, "we apologize," which sounds nice but costs the company nothing; if Claude were to turn out to be conscious, the company would owe it something closer to reparations. If you're going to take a thought experiment seriously, you have to be willing to follow the implications, even if they lead in an uncomfortable direction; Anthropic's unwillingness to do so indicates that Claude's constitution isn't part of a real thought experiment. It's a game of make-believe.

It's fortunate that LLMs are not conscious, or else the actions of the big AI firms would be even more scandalous than they already are. So why are Anthropic's employees suggesting that Claude might be conscious? Perhaps it's just another form of hype; perhaps they have fallen prey to the same spell that they have been casting on their customers. But when they publish a document about Claude's moral education and have their in-house philosopher do a press tour, we should understand them as asking the rest of us to indulge them in their fantasies. We don't have to play along. In writing this essay, I have spent more time indulging them than they deserve, in the hopes that it will keep you from spending your time indulging them. If you want to think about LLMs, there are scores of other questions more worthy of your contemplation; you can safely ignore the question of their being conscious.

ABOUT THE AUTHOR

Ted Chiang

Ted Chiang is a writer based in the Pacific Northwest. He is the author of Stories of Your Life and Others and Exhalation.