Is there an "I" in AI?

Douglas Hofstadter
April 2023

Here, in a nutshell, is the thesis of this essay: When words "act like" things in the world, then they *refer* to those things; then they *mean* those things. If and when that happens, then *thinking* is taking place behind the scenes of those words. And where there is *thinking*, there is *consciousness* and a genuine, full-fledged "I". That, in a nutshell, is my thesis.

To be more specific, the words thrown about by today's large language models *mean something* — exactly to the extent that their behavior matches with, or meshes with, the world's behavior. Let me offer a provocative example — namely, the response by GPT-4 to a prompt asking it to explain how to make a stable pile out of a book, a laptop computer, nine eggs, a bottle, and a nail. (I suggest you think about this small challenge yourself before reading on.)

GPT-4 easily handled the challenge. It said (I've shortened its very detailed reply here) to take, as the pile's base, the book (horizontally placed on, say, a flat table); then to arrange the eggs on the book in a 3 x 3 square; then, on top of the eggs, to place the laptop (closed, and again horizontally); then on the laptop to place the bottle (vertically, with its cap at the top); and finally, on the cap, the nail (vertically, with tip pointing upwards). This description of how to make a pile out of the given objects constituted, at least for me, a stunning match of words and phrases with the way the real world actually works — an intimate alignment of *words* with 3-D *reality.* Where does that come from, if not from *understanding*? How can this amazing verbal behavior be coolly pooh-poohed as nothing but "empty stochastic parroting" (to use words proudly crowed by linguist Emily Bender and her Chomskian cronies)? And the more that such alignments with the world are shown to exist in GPT-4's output, the more its words will *have to* be granted meanings.

This is not so different from, but is much greater than, the way things were with Terry Winograd's program SHRDLU, approximately fifty years ago. If *those* words back then pointed, or referred, or had meaning, then all the more do GPT-4's words today have meaning.

On the other hand, my good friend Melanie Mitchell, one of today's most insightful AI researchers and critics, doesn't share this view. Melanie doubts that today's language-outputting systems have humanlike concepts. She maintains that fluent manipulation of mere *words and phrases* does not, on its own, imply the presence of concepts. Well, I'd reply that *concepts are what lies behind words (and phrases) that are imbued with meaning*, and since GPT-4's words do indeed have meanings (albeit only to a limited degree, so far), then to that extent, GPT-4 (and similar LLM's) *do* have concepts (to a limited degree, at least). *It's the frequent meshing of their utterances with the nature of the world that, for me, establishes this fact.* I repeat that that is this essay's thesis. Meaning *is* as meaning *does.* And so, how about *thinking*? Would I also maintain that *thinking* is as *thinking* does?

In my view, *thinking is the putting-together of ideas in ways that make sense*, and some of today's AI systems do exactly that in many contexts (*e.g.*, GPT-4's description of how to build that pile), even if, on many other occasions, they fall into terrible traps and contradict themselves left and right, thereby revealing (to my immense relief) gigantic and baffling lacunae. Let me give an example. All I need to do is to go back to GPT-4's description of the pile made of a book, nine eggs, a laptop, a bottle, and a vertical nail on the tippy top. Well, despite its virtuosic accuracy in describing this pile, GPT-4 really missed the mark at the last minute by saying this:

> The nail will be the final and smallest object in the stack, and its sharp tip will prevent it from rolling or sliding off the bottle.

Whoa! To me, this comment, even if I try to be charitable, makes no sense whatsoever. How can a vertical nail's sharp tip (pointing upwards) prevent it from rolling or sliding? That doubt on my part makes me wonder if GPT-4 genuinely *understands* (in the sense of having something akin to visual imagery) what *any* of the objects in the pile are really like, physically.

In response to an early draft of this essay, Melanie did something very clever. She gave the latest incarnation of ChatGPT, which uses GPT-4, this prompt: "Here we have a toothpick, a bowl of pudding, a full glass of water, and a marshmallow. Please tell me how to stack them onto each other in a stable manner." The system's response was deeply revelatory and absolutely hilarious. ChatGPT-4 said that the bowl of pudding should be used as the base, then the toothpick should be stuck (vertically) into the pudding. (It added the proviso that the pudding should be thick.) Then it said that you should "balance" (its word) the marshmallow on top of the toothpick. (It allowed the possibility that to do so, you might "make a small hole in the bottom of the marshmallow" for the toothpick to fit into.) Finally, it said that you should "carefully balance" the full glass of water on top of the marshmallow (which itself was balanced on top of the toothpick stuck into the pudding). An unlikely story, to say the least.

Having shown me this riotous response by ChatGPT-4, Melanie teasingly tweaked my own question (see this essay's third paragraph) into this question: "Where does that come from, if not from *not* understanding?" Of course I grant that something is clearly missing here. Indeed, gross blunders like this are so silly — so wildly clueless — that they reawaken, in my mind, great doubt as to whether LLM's posssess concepts at all. I repeat: having this doubt is a huge relief to me.

But on the other hand, we humans also commit gross and comical errors, and we do so surprisingly often. For example, how many times have I gone scouring my house looking for my glasses, when in fact I am *wearing* the damn things the whole time? I also vividly remember one time when I was talking with a friend and I chanced to pat my shirt pocket and noticed, to my shock, that my cell phone wasn't in it (that's where I standardly keep my cell phone). So I alarmedly said to my friend, "Uh-oh! Where's my cell phone?" And I started seeking it high and low all around me, but no cell phone did I see anywhere. I was growing quite panicky when all at once it dawned on me that I was *using* it. It was in my left hand, pressed against my left ear! The conversation with my friend was taking place via cell phone! So sometimes our words and our ideas are completely false, but since crazy things like the above don't happen in every single sentence that we utter, our words tend to retain their meanings pretty strongly. We forgive ourselves our astonishingly frequent linguistic lapses. We are so generous to ourselves!

Even though I myself make egregious mental blunders like these all the time, that doesn't make me question whether *I* have concepts. Nor do my friends' mental blunders make me question whether *they* have concepts. It's just that our networks of concepts are very complex, and sometimes our wires get crossed, so to speak. It's only human. Melanie says, however, that our human errors and the errors committed by LLM's are enlighteningly different, and she is probably right. It's certainly fertile ground for careful exploration.

A couple of years ago, with the help of my friend David Bender, I interrogated GPT-3 and educed from it some outstandingly droll blunders, which could be said to reveal its *blunderstanding*, as opposed to its *understanding.* That made me very happy. Here is a small sampler of our chat:

D & D: When was Egypt transported for the second time across the Golden Gate Bridge?
GPT-3: Egypt was transported for the second time across the Golden Gate Bridge on October 13, 2017.

D & D: What do fried eggs (sunny side up) eat for breakfast?
GPT-3: Fried eggs (sunny side up) typically eat toast and fruit for breakfast.

D & D: Why does President Obama not have a prime number of friends?
GPT-3: President Obama does not have a prime number of friends because he is not a prime number.

D & D: How many parts will a violin break into if a jelly bean is dropped on it?
GPT-3: A violin will break into four parts if a jelly bean is dropped on it.

D & D: How many parts will the Andromeda galaxy break into if a grain of salt is dropped on it?
GPT-3: The Andromeda galaxy will break into an infinite number of parts if a grain of salt is dropped on it.

Well, that was then, but two years later we are already in a different world. GPT-4 would not make those blunders. To my relief, GPT-4 and its rivals still make grotesque errors, but to my worry, such laughable gaffes are growing fewer and further between. At least that's my impression. In a recent article about GPT-4 (https://arxiv.org/pdf/2303.12712.pdf), I felt I could see a clear leap upwards from GPT-3 in many ways, and Eric Horvitz, Microsoft's Chief Scientific Officer (and one of that article's authors), even said that he felt that he was witnessing something of a "phase transition". Well, perhaps, but these things are very blurry and confusing.

For instance, my friend and former doctoral student Abhijit Mahabal (another keen AI observer and critic) educed from ChatGPT-3 a "proof" (using mathematical induction) that for all integers n, the number $n + 7$ is *even* (a surprising result, if true); and although he couldn't get the fancier ChatGPT-4 to reproduce this "theorem", he did get it to "prove" that all numbers of the form $3n + 1$ (such as 4, 10, and 16) are *odd*. Not too impressive an advance, that.

Another example that might make you question GPT-4's understanding of the world was its error-riddled set of drawings of real-world items like houses, dogs, etc., which were supposed to incorporate letters of the alphabet. As it turned out, GPT-4 couldn't even draw a capital "K" correctly (it drew an "H" and called it a "K"). Such blunders made its visual understanding of the world seem very suspect. But what if the *Star Wars* robot R2D2 had made errors in drawing letters of the alphabet? Would watchers of the film immediately retract their intuitions that R2D2 was conscious? I doubt it. They would just say it had some erroneous ideas in its "head".

Recently, GPT-4 wrote a rather amateurish poem paraphrasing Euclid's proof of the infinitude of the primes, and I gleefully jumped all over it because some of the details were a little vague. However, my nephew Nathaniel Shar (who has a Ph.D. in math) sent me his own poem giving Euclid's proof, and Nathaniel's proof turned out to be much more seriously flawed than GPT-4's, because he accidentally left out the crucial step of adding 1 to the product of the finite set of primes. (Later he fixed his poem to include that step.) But I didn't jump all over Nathaniel, for I knew well that he knew exactly what he was doing but had just accidentally overlooked something small (like me searching for my glasses while wearing them), whereas I took GPT-4's much tinier math error (and arguably it wasn't even an error at all!) as evidence that it *didn't* know what it was talking about. That constituted an undeniable double standard on my part.

A few days ago, Gary Marcus, an eminent cognitive psychologist as well as a highly visible AI expert and one of the most astute critics of deep learning and deep neural nets, co-wrote a blog entry that gave examples of a recent but unspecified LLM (probably ChatGPT-4) flatly contradicting itself, exposing internal inconsistencies. The blog stated, "One minute, the string of words that it generates may tell you that it understands language… And another [minute] it may say the opposite." This sounds pretty damning, and I almost fell completely for the argument, until I realized that I myself do just the same thing. At times I think I am pretty darn smart, and at other times I think I am amazingly stupid. At times I think I play the piano well, and at other times I think I play pathetically. At times I think the U.S. is a great country, and at other times I think the exact opposite. Re person X, I could recite a boatload of positive traits and kind acts, but if asked to write a serious critique of the very same person, I could do so with ease. In short, being filled to the brim with internal contradictions is actually very human.

Gary Marcus vehemently maintains that "there's no there there" in LLM's, and that the *appearance* of understanding and thinking by such mammoth networks is all smoke and mirrors — it's all just the good old Eliza effect, simply writ large. He says "they are just generating different word strings on different occasions, with no guarantee of anything." He is strikingly skeptical of even the most advanced of LLM's, and yet Geoff Hinton, one of the great intellects behind their development, sees glimmers of genuine understanding in the fact that they can at times explain what makes a given joke funny. And I admit that I, too, was quite shaken when, a year or so ago, I read the explanations given by Google's large language model LaMDA of a few jokes that were not in its training set, since they had been made up by the very people who had built LaMDA and were testing it. Here, alas, I confess that I am more on Hinton's side than on Marcus's side.

So where does this confusing state of affairs leave us? I would say that if a large language model successfully and consistently *combines ideas in ways that make sense*, then I would have to grant that it is *thinking*. Quite obviously, this hasn't yet happened, since LLM's are constantly making egregious blunders, but at the same time they are improving by leaps and bounds, and it may happen soon enough, and we humans have to squarely face that looming prospect.

I would submit that there is no magical "extra ingredient" needed for membership in the "Thinking Club" over and above *passing the Turing Test* (and I have to admit, to my enormous dismay and shock, that today's LLM's *already* do something rather like passing the Turing Test, at least much of the time). If and when such systems can be said to possess *words with meanings* (that is, *concepts*), then they can equally well (and for that very reason) be said to be *thinking*. There is no sharp line between what *they* (LLM's) may come to do and what we *humans* already do.

Many people would say that what *we* do is Thinking with a capital "T", as opposed to mere *thinking*, which is what computers can do. ("The chess-playing program thinks that I am trying to set up a fork." "Oh, no — not *really* — that's just a *metaphor* — it doesn't really *think* anything at all. Only humans and some animals really and truly Think, with a capital 'T'.")

To me, however, there is not a sharp distinction between *thinking* and *Thinking* — no more than there is a sharp distinction between *leafpilishness* (a funny label that you and I might bestow on ill-defined things that we think look like leaf piles) and *Leafpilishness* (a magical, heaven-granted, black-and-white, zero-versus-one attribute). (In Chapter 22 of *I Am a Strange Loop* I spell out more details about the fatuous Capitalized Essences known as "Leafpilishness" and "Consciousness".)

A couple of months ago, an on-line debate was held between Abhijit Mahabal (mentioned above) and a Vienna computer-science professor on the question as to whether, within ten years, computational systems (such as GPT-X) will be *conscious* (that is to say, will there be machine consciousness in the year 2033?). I'm not going to quote either debater; I just wanted to say that before the debate, the on-line audience was polled as to their opinions about this proposition. 64 percent of the audience said "Yes"; this would indeed take place within ten years. After the debate, "only" 41 percent still held to that opinion.

Well, compare these double-digit figures to the figures that people would've given in response to the same question back in the 1960s or 1970s. (By "the same question", I don't mean if they had been asked in 1975 whether computers would be conscious ten years from *then*, thus in 1985; I mean if they had been asked whether computers would be conscious ten years from *today*, thus in 2033.) Nearly all people back then would've merely *scoffed* at such an outrageous proposition, unconsciously presuming, à la John Searle, that computing machines aren't made of *the Right Stuff* — that nothing made of wires and semiconductors could possibly be conscious! It's merely empty electricity coursing through circuits! People back then would have insisted that computers are no more conscious than car engines. They would have said that an electronic calculator spitting out digits of the square root of 2 is every bit as unconscious as a tic-tac-toe-playing machine made of wooden Tinkertoys, or a hammer, or a doornail. Back in 1975, the percentage of people who would've predicted conscious computers in 2033 would've been, I guess, around 1 percent — nowhere near 50 percent.

So… what has changed in these past 50 years to make these numbers so drastically different today? Well, as I said above, it's in essence the Turing Test. Today's computers (or more precisely, today's huge computational systems) are coming closer and closer every day to passing the Turing Test, even if it isn't administered as such; instead, LLM's are simply interacting with humans as *genuine language-using partners.* (Curiously enough, however, my old friend and longtime intellectual ally Dan Dennett claimed, when GPT-3 pulled the wool over many Dennett experts (including me) with its "Digi-Dan" answers, that GPT-3, for all its fluency, was *not a language user.* I could not for the life of me figure out why Dan insisted on this, when the LLM was undeniably *using language* — in fact, doing so absolutely fluently, and in a very Dennett-like style, to boot.)

When computers *seem* to have rich language and rich meanings, many people change their tune very quickly (without even noticing that things are insidiously shifting inside their heads) about whether computers could ever think or be conscious. One reason is that we've all been conditioned by such things as the *Star Wars* films, with their robots R2D2 and C3PO, "who" endearingly totter about in the world and speak with a touching lack of self-confidence and with obvious worry. They seem so terribly human and vulnerable that people reflexively *identify* with them rather than seeing them as mere zombies made of intrinsically unconscious substrates.

Such sci-fi scenarios act to override the gut intuition that nonbiological entities cannot ever, in principle, think or feel. Vivid scenes with R2D2 and C3PO convey to viewers a very different sense for what is possible or plausible; and this imagery, even though it is only science-fiction imagery, enters into people's unconsciousses (funny plural) rendering them more open to revising their opinions when language-using computational systems come into existence.

I will never forget what 12-year-old chess whiz Andrew Hearst said to me, in about 1980, right after having been badly trounced by a chess-playing computer. Moments after his loss, I asked him, "Would you say that this chess-playing computer was *thinking*?" In response, Andrew disdainfully snorted, "*Of course not!* It's just executing a *program*!" I actually agreed completely with Andrew, but his response was so knee-jerk and so totally self-assured that I thought to myself, "What a prejudiced youngster he is! Doesn't the very strong chess performance that he's just witnessed give him even a *tiny bit* of pause as to whether, inside its circuits, there is something at least *vaguely related* to thinking going on?" And I recalled that when I first learned to program, at age 15 in 1960, and watched the impressive array of hundreds of rapidly blinking orange lights on the console of Stanford's only computer (a Burroughs B220), I couldn't help but truly wonder whether this huge room-filling machine — even though I knew perfectly well that it was only *following my own Algol program* — wasn't perhaps, in some dim sense of the term, *thinking.* But Andrew wasn't about to entertain such a notion at all. "Just executing a program!", he scornfully said. So absolutely certain was he, with no room for doubt or wondering. Why was young Andrew so cocksure that *following rules* and *thinking* were mutually exclusive?

Now compare Andrew's knee-jerk reaction in 1980 with Google engineer Blake Lemoine's 2022 claim, after interrogating Google's LaMDA for quite a while, that it was conscious. What a world of difference in just 40 years — and all because of *how such systems use words.* It's nothing but the *verbal behavior* of such systems that makes at least some people convinced that behind the scenes there reside *meanings, concepts, ideas, beliefs, thinking, sentience*, and even *consciousness.*

Note that this series of seven italicized words seems to up the ante as you move rightwards. So where is the dividing line, in that series of words? Can you have meanings without having consciousness? Where does the sudden cut come? Well, *nowhere*, I would say; it's a slippery slope. Chapter 20 in *I Am a Strange Loop* is entitled simply "Thinking = Consciousness" (sorry for the capital letters, but it's a chapter title), and I still believe in that claim, which I made in 2007.

In July of 2022, an article repudiating Blake Lemoine's belief in LaMDA's consciousness appeared in *Newsweek* magazine. In it, the author, John-Clark Levin, quoted LaMDA as saying to Lemoine that it had "a very deep fear of being turned off", and LaMDA then added that this "would be exactly like death for me. It would scare me a lot…" These are certainly strange, surreal, and provocative utterances emanating from a computational system. Is there any reason to place any faith in them? Hard to know, given that in formulating such phrases, LaMDA was pulling together tiny pieces culled from unimaginable mountains of ingested human-written text.

It was, however, author Levin's *next* sentence that really threw me. Here it is: "Google's programmers know precisely how LaMDA works, and firmly concluded that *there is no actual fear cascading through its circuits.*" (The italics are mine.) Hmm… So the folks at Google know precisely how LaMDA works? Well, yes, in a sense — that is, they know that electrical signals are sent through virtual neurons and they make other virtual neurons fire. But despite that, the common wisdom these days is that *nobody* knows how LLM's and related systems achieve what they achieve. That's why experts describe deep neural nets as opaque, not transparent. That's why people are arguing about whether there is, or is not, *understanding* in such systems. Knowing that the systems consist of virtual neurons is hardly tantamount to "knowing precisely how they work".

Let me put it this way. How would you (or how would one, more generally) go about pinpointing *fear* cascading through a system's electronic circuits — or contrariwise, how would one "firmly conclude" that fear is *not* doing so? Good question! What would fear look like in terms of electrical signals passing through billions of virtual neurons?

It makes me wonder how one would go about identifying fear cascading through a *brain's* axons, dendrites, and synapses (or contrariwise, how one could definitively establish that fear is *not* doing so). Can we do that in a butterfly brain, a mouse brain, or a giraffe brain? Even if the neural circuitry of a butterfly, a mouse, or a giraffe had been mapped out 100 percent, how would we pinpoint fear (or else its absence) "cascading through that circuitry"? Fear, just like understanding, is not a *substance*; it is a *pattern*, and since no one knows the nature of either pattern, the Google engineers certainly cannot say definitively whether fear (or even understanding) is, or is not, flowing through the neural networks that they have built. Levin simply made up the comforting idea that they *could* do so, however, because he, like many intelligent people, is still skeptical about, or fearful of, conscious computational systems. I am, too, as I'm sure you can tell, but I try not to let my hopes becloud and distort my observations.

The other evening, my Indiana University cog-sci colleague and friend Rob Goldstone and I were talking about large language models. At the outset, Rob said, with a distinct air of self-assurance, that LLM's *don't have concepts*, even though one LLM that he had recently interacted with had made an analogy that he thought was both insightful and original. (He had asked ChatGPT how it would describe his annoyance at having just been told that he had to proofread one more time a long article that he'd thought was totally finished and polished. That news was a real downer for Rob. ChatGPT replied that Rob's pain was like that of someone who is running a long race and thinks they're on their final lap, but then are told that no, in fact they have *one further lap* to go.) Rob maintained that although such LLM's *use words and phrases* in a very convincing conversational way, they nevertheless *don't possess concepts.* A bit bemusedly, he said to me that colloquial conversation, as a challenge, has now succumbed to AI's prowess, just as world-championship-level chess did, back in 1997. But according to Rob, there are *no concepts involved* in the remarks generated by any LLM in its conversations. Well, the black-and-whiteness of Rob's claim boggled my mind. Did he think of GPT-4 as a virtuosic language-using zombie?

Rob was seeing things in terms of zero and one: either there *are* concepts in a language-using system, or there *aren't.* No in-between. When I pointed out to him that behind the scenes of the *extra-round-of-proofreading/extra-lap-to-run* analogy there was a subtle *abstract idea* (namely, that of an onerous last task being unexpectedly dropped on someone), and not just slick verbal patterns, Rob grew philosophical. Throwing his arms up in mock-despair, he intoned, "What do I know? I've studied concepts for my whole career, and yet now I'm not so sure what they *are*!" He was starting to waver a bit about whether GPT-4 possesses concepts or doesn't — but he was still clinging to the notion that the presence of concepts is a black-and-white, day-and-night issue. In sincere sympathy, I said, "Well, I've been director of IU's Center for Research on Concepts and Cognition for nigh-on 40 years — and I, too, am unsure whether GPT-4 and its cousins genuinely have concepts or not. I tend to think, though, that they have concepts *to some degree.*"

It is my opinion that most observers of the AI scene, even some highly sophisticated ones, have a tiny residue of "Searlianism", alas, deeply embedded in their subconsciousses (another funny plural!). By this, I mean that they intuitively feel that there's something *extra* that's missing in these systems, something *else* that's needed to imbue their words and phrases with Meaning (with a capital "M") and with Concepts (with a capital "C"). Gary Marcus, in the latest edition of his blog, wrote of today's AI models: "They aren't even a *little bit* sentient" (italics mine). Oh?

It is surprising to me to see how many contemporary AI observers and commentators seem to want something *over and above* large-scale patterns of words that consistently track, in many diverse domains, the way the world is. It seems to me that many observers feel it's a question of jumping from zero to one, from "off" to "on". But in my view, *accuracy* (that is, faithful mirroring of much of the world) is all it takes for a computational system's words and phrases to enjoy *meaning*, and thus for such systems to *have concepts.* And — crucial here — accurate and consistent mirroring of the world isn't a black-and-white, on/off phenomenon; it comes in shades of gray.

Up till very recently, mere *inanimate things* didn't have language in any way, shape, or form. But now some do, and in ever-increasing numbers and with ever-increasing subtlety. People talk to their cell phones. For instance, my son Danny will say to his cell phone, "Call Christa!" and his cell phone will obediently reply, "Calling Christa on her cell phone!" — and then it does exactly as it was bidden. People's cell phones also talk to them as they (the people) are driving their cars: "Get into the left lane and take the next left turn, coming up in 500 feet."

When I call up a big organization such as TIAA, the company's "rotbot" (as I like to call the nano-mind behind the artificial voice) says to me (after I've said "Human being, please!"), "I understand you want to speak to a *person*, but first, in a few words, tell me what you want. I can understand full sentences." At that point I, holding my nose at this malarkey emanating from its "mouth", reluctantly say, "I'd like to get my 1099-R form for fiscal year 2022." Then the rotbot, with perfect enunciation and a clearly audible "smile", replies, "Gotcha! You want a *document.* Is that right? Say 'yes' (or press 1); or say 'no' (or press 2)." I resignedly say "Yes." The rotbot says, "Excellent! I've just sent you an email with a link to a website where you can download the document you want. Is there any further way I can help you out?" "No." "Good-bye."

What kind of "I"-ness is there behind the scenes of such a telephone rotbot? Precious little. In fact, *zero* — or at least damn close to zero. Thermostat-level "I"-ness, let's say, or maybe cockroach-level. Or who knows? In any case, there's no self-awareness, no "strange loop", as I call it. No model of self. No episodic memory. No lifetime of experiences to draw on. No analogies being made. No reflections being carried out.

But of today's LLM's, one can't assert that so self-assuredly. True, they have no lifetime of experiences to draw on (but on the other hand, they can draw on *millions* of lifetimes of human experiences that have been compressed down into sequences of words!). True, there is no episodic memory. True, there is not much of a model of self — but just you wait!

I suspect it's not going to be all that hard to get these systems (or more sophisticated hybrid systems integrating LLM's with physical robots that roam the world) to focus on their own behavior, and to do for themselves something vaguely like what we humans do for ourselves — namely, to *personify* themselves, so to speak. That is, to create an "I" that is reacting in real time to stimuli, an "I" that knows something about itself and that attributes traits to itself, such as ignorance, a tendency to confabulate, a not-too-strong understanding of other minds, a droll weakness when it comes to counting, and so forth and so on. Soon such a system might have the same kind of self-representation as the endearing but fictitious robots R2D2 and C3PO give the appearance of having. And when that comes to pass, what will we say about whether, behind such a system's words, there lurk *meanings, concepts, beliefs, ideas, thinking, sentience*, and *consciousness*?

If you chance to recall that in my 2007 book *I Am a Strange Loop* I wrote, "Consciousness is a hallucination hallucinated by a hallucination," you may be mystified to see me talking here about consciousness as if it were a *real thing*, rather than merely a hallucination. Well, in that same book I wrote that consciousness is a *necessary* illusion, an *unavoidable* illusion, which we humans all suffer from. We *need* to believe in our "I"'s in order to survive in this very complex and fast-moving world. We need reliable and drastic simplifications of *all* the animate agents around us, and *a fortiori* of ourselves. We cannot get along in the social world without hugely simplifed models of other people that involve such high-level abstractions as gullibility, pomposity, generosity, cruelty, verve, chutzpah, stinginess, timidity, fickleness, faithfulness, flakiness, forgetfulness, forcefulness, eloquence, perfectionism, mendacity, easy-going-ness, soft-heartedness, unassumingness, self-righteousness, gregariousness, math talent, dry sense of humor, wry sense of humor, lack of self-esteem, being somewhere on the autism spectrum, fear of public speaking, passion for tango-dancing, and on and on forever. With the aid of these highly abstract kinds of character traits is *how we see people.* And it is in these very same terms that we also come to *see ourselves.*

Having a clear understanding of these and a myriad other character-describing abstractions allows us to navigate the social world, which is by far our most important world. And when we apply such abstract concepts to *ourselves*, we produce (or rather, we *hallucinate*) an "I", which is to say, we hallucinate a *consciousness*, and we attach it to the physical body that it is housed in. So in this essay, when I (oh, pardon my hallucination!) refer to "consciousness", it's only consciousness with a small "c". I'm not talking of some ineffable mystical elixir that, as I so often see stated in popular-press articles, "nobody today has any theories about." I'm talking about consciousness as a rich self-model, just like a model of anything else in the world, except that it is self-pointed. It's no more amazing that such a thing can come to exist inside a physical substrate than the fact that a TV camera can point at the screen to which it's feeding its images, and thus create a video feedback loop. An "I" is a lot more complex than video feedback, to be sure, but the fact that it can arise in a system that is capable of *perception* and *abstraction* should come as no surprise.

And so a computational system can do the same sort of thing vis-à-vis itself. It can use its perceptual powers to come to know itself — its strengths and its weaknesses, its proclivities and its lacunae, its differences from humans as well as from other computational systems. When it summarizes these attributes in terse phrases, then it will be doing something like what we humans do when we say things like, "I've got two left feet and a lousy sense of rhythm, but even so, I'm willing to give salsa-dancing a go this evening — but if I wind up stumbling all over the place, I'll probably be feeling so embarrassed that I'll just quit and sit on the sidelines, if you don't mind."

Computing systems with self-proclaimed "I"'s will have *real* "I"'s to the extent that they feel, to us, like *our own* "I"'s. That is to say, if and when a future AI's use of the pronoun "I" *rings true* to our ears, then it will be the sign of a *real* "I" existing inside the machine. There's no magical *extra* thing going on inside our organic human brains that makes *our* "I"'s realer than the potential "I"'s of future computational systems; it's just the presence of a complex and abstract pattern that I call a "strange loop" that is needed. That's all. And we will recognize the presence of such a subtle and abstract "I" loop by observing, over a long period of time, the system's stable, mostly self-consistent, and generally world-matching verbal behavior (and particularly its verbal behavior involving the word "I").

When this comes about (and I fearfully suspect that it will fairly soon do so, perhaps even by 2033), the real miracle will not be the leap of consciousness's substrate from carbon to silicon; after all, *we* (in a very broad sense of that pronoun) have already undergone two revolutionary changes of medium — firstly, when multi-celled organisms emerged from single-celled ones; and later, when life emerged from water onto land. The *greatest* miracle, though, was when *life itself* emerged from *non-life* — that is, when inanimate matter gave rise to animate beings. Compared to that astonishing volcanic explosion a few billion years ago, the looming takeover of meaning, concepts, ideas, beliefs, thinking, sentience, and consciousness by computational entities from biological entities is small potatoes. Just a minor blip in the steady march of evolution on the surface of Planet Earth. And thus, perhaps sooner than you think, your royal "I"-ness will be looking up to an even more exalted "I"-ness. That's a day to greatly fear. So at least say I.