There Is No A.I.

There are ways of controlling the new technology—but first we have to stop mythologizing it.

By Jaron Lanier

April 20, 2023

Source: *The New Yorker*
https://www.newyorker.com/science/annals-of-artificial-intelligence/there-is-no-ai

As a computer scientist, I don't like the term "A.I." In fact, I think it's misleading—maybe even a little dangerous. Everybody's already using the term, and it might seem a little late in the day to be arguing about it. But we're at the beginning of a new technological era—and the easiest way to mismanage a technology is to misunderstand it.

The term "artificial intelligence" has a long history—it was coined in the nineteen-fifties, in the early days of computers. More recently, computer scientists have grown up on movies like "The Terminator" and "The Matrix," and on characters like Commander Data, from "Star Trek: The Next Generation." These cultural touchstones have become an almost religious mythology in tech culture. It's only natural that computer scientists long to create A.I. and realize a long-held dream.

What's striking, though, is that many of the people who are pursuing the A.I. dream also worry that it might mean doomsday for mankind. It is widely stated, even by scientists at the very center of today's efforts, that what A.I. researchers are doing could result in the annihilation of our species, or at least in great harm to humanity, and soon. In a recent poll, half of A.I. scientists agreed that there was at least a ten-per-cent chance that the human race would be destroyed by A.I. Even my colleague and friend Sam Altman, who runs OpenAI, has made similar comments. Step into any Silicon Valley coffee shop and you can hear the same debate unfold: one person says that the new code is just code and that people are in charge, but another argues that anyone with this opinion just doesn't get how profound the new tech is. The arguments aren't entirely rational: when I ask my most fearful scientist friends to spell out how an A.I. apocalypse might happen, they often seize up from the paralysis that overtakes someone trying to conceive of infinity. They say things like "Accelerating progress will fly past us and we will not be able to conceive of what is happening."

I don't agree with this way of talking. Many of my friends and colleagues are deeply impressed by their experiences with the latest big models, like GPT-4, and are practically holding vigils to await the appearance of a deeper intelligence. My position is not that they are wrong but that we can't be sure; we retain the option of classifying the software in different ways.

The most pragmatic position is to think of A.I. as a tool, not a creature. My attitude doesn't eliminate the possibility of peril: however we think about it, we can still design and operate our new tech badly, in ways that can hurt us or even lead to our extinction. Mythologizing the technology only makes it more likely that we'll fail to operate it well—and this kind of thinking limits our imaginations, tying them to yesterday's dreams. We can work better under the assumption that there is no such thing as A.I. The sooner we understand this, the sooner we'll start managing our new technology intelligently.

If the new tech isn't true artificial intelligence, then what is it? In my view, the most accurate way to understand what we are building today is as an innovative form of social collaboration.

A program like OpenAI's GPT-4, which can write sentences to order, is something like a version of Wikipedia that includes much more data, mashed together using statistics. Programs that create images to order are something like a version of online image search, but with a system for combining the pictures. In both cases, it's people who have written the text and furnished the images. The new programs mash up work done by human minds. What's innovative is that the mashup process has become guided and constrained, so that the results are usable and often striking. This is a significant achievement and worth celebrating—but it can be thought of as illuminating previously hidden concordances between human creations, rather than as the invention of a new mind.

As far as I can tell, my view flatters the technology. After all, what is civilization but social collaboration? Seeing A.I. as a way of working together, rather than as a technology for creating independent, intelligent beings, may make it less mysterious—less like *HAL* 9000 or Commander Data. But that's good, because mystery only makes mismanagement more likely.

It's easy to attribute intelligence to the new systems; they have a flexibility and unpredictability that we don't usually associate with computer technology. But this flexibility arises from simple mathematics. A large language model like GPT-4 contains a cumulative record of how particular words coincide in the vast amounts of text that the program has processed. This gargantuan tabulation causes the system to intrinsically approximate many grammar patterns, along with aspects of what might be called authorial style. When you enter a query consisting of certain words in a certain order, your entry is correlated with what's in the model; the results can come out a little differently each time, because of the complexity of correlating billions of entries.

The non-repeating nature of this process can make it feel lively. And there's a sense in which it can make the new systems more human-centered. When you synthesize a new image with an A.I. tool, you may get a bunch of similar options and then have to choose from them; if you're a student who uses an L.L.M. to cheat on an essay assignment, you might read options generated by the model and select one. A little human choice is demanded by a technology that is non-repeating.

Many of the uses of A.I. that I like rest on advantages we gain when computers get less rigid. Digital stuff as we have known it has a brittle quality that forces people to conform to it, rather than assess it. We've all endured the agony of watching some poor soul at a doctor's office struggle to do the expected thing on a front-desk screen. The face contorts; humanity is undermined. The need to conform to digital designs has created an ambient expectation of human subservience. A positive spin on A.I. is that it might spell the end of this torture, if we use it well. We can now imagine a Web site that reformulates itself on the fly for someone who is color-blind, say, or a site that tailors itself to someone's particular cognitive abilities and styles. A humanist like me wants people to have more control, rather than be overly influenced or guided by technology. Flexibility may give us back some agency.

Still, despite these possible upsides, it's more than reasonable to worry that the new technology will push us around in ways we don't like or understand. Recently, some friends of mine circulated a petition asking for a pause on the most ambitious A.I. development. The idea was that we'd work on policy during the pause. The petition was signed by some in our community but not others. I found the notion too hazy—what level of progress would mean that the pause could end? Every week, I receive new but always vague mission statements from organizations seeking to initiate processes to set A.I. policy.

These efforts are well intentioned, but they seem hopeless to me. For years, I worked on the E.U.'s privacy policies, and I came to realize that we don't know what privacy is. It's a term we use every day, and it can make sense in context, but we can't nail it down well enough to generalize. The closest we have come to a definition

of privacy is probably "the right to be left alone," but that seems quaint in an age when we are constantly dependent on digital services. In the context of A.I., "the right to not be manipulated by computation" seems almost correct, but doesn't quite say everything we'd like it to.

A.I.-policy conversations are dominated by terms like "alignment" (is what an A.I. "wants" aligned with what humans want?), "safety" (can we foresee guardrails that will foil a bad A.I.?), and "fairness" (can we forestall all the ways a program might treat certain people with disfavor?). The community has certainly accomplished much good by pursuing these ideas, but that hasn't quelled our fears. We end up motivating people to try to circumvent the vague protections we set up. Even though the protections do help, the whole thing becomes a game—like trying to outwit a sneaky genie. The result is that the A.I.-research community communicates the warning that their creations might still kill all of humanity soon, while proposing ever more urgent, but turgid, deliberative processes.

Recently, I tried an informal experiment, calling colleagues and asking them if there's anything specific on which we can all seem to agree. I've found that there is a foundation of agreement. We all seem to agree that deepfakes—false but real-seeming images, videos, and so on—should be labelled as such by the programs that create them. Communications coming from artificial people, and automated interactions that are designed to manipulate the thinking or actions of a human being, should be labelled as well. We also agree that these labels should come with actions that can be taken. People should be able to understand what they're seeing, and should have reasonable choices in return.

How can all this be done? There is also near-unanimity, I find, that the black-box nature of our current A.I. tools must end. The systems must be made more transparent. We need to get better at saying what is going on inside them and why. This won't be easy. The problem is that the large-model A.I. systems we are talking about aren't made of explicit ideas. There is no definite representation of what the system "wants," no label for when it is doing a particular thing, like manipulating a person. There is only a giant ocean of jello—a vast mathematical mixing. A writers'-rights group has proposed that real human authors be paid in full when tools like GPT are used in the scriptwriting process; after all, the system is drawing on scripts that real people have made. But when we use A.I. to produce film clips, and potentially whole movies, there won't necessarily be a screenwriting phase. A movie might be produced that appears to have a script, soundtrack, and so on, but it will have been calculated into existence as a whole. Similarly, no sketch precedes the generation of a painting from an illustration A.I. Attempting to open the black box by making a system spit out otherwise unnecessary items like scripts, sketches, or intentions will involve building another black box to interpret the first—an infinite regress.

At the same time, it's not true that the interior of a big model has to be a trackless wilderness. We may not know what an "idea" is from a formal, computational point of view, but there could be tracks made not of ideas but of people. At some point in the past, a real person created an illustration that was input as data into the model, and, in combination with contributions from other people, this was transformed into a fresh image. Big-model A.I. is made of people—and the way to open the black box is to reveal them.

This concept, which I've contributed to developing, is usually called "data dignity." It appeared, long before the rise of big-model "A.I.," as an alternative to the familiar arrangement in which people give their data for free in exchange for free services, such as internet searches or social networking. Data dignity is sometimes known as "data as labor" or "plurality research." The familiar arrangement has turned out to have a dark side: because of "network effects," a few platforms take over, eliminating smaller players, like local newspapers. Worse, since the immediate online experience is supposed to be free, the only remaining business is the hawking of influence. Users experience what seems to be a communitarian paradise, but they are targeted by stealthy and addictive algorithms that make people vain, irritable, and paranoid.

In a world with data dignity, digital stuff would typically be connected with the humans who want to be known for having made it. In some versions of the idea, people could get paid for what they create, even when it is filtered and recombined through big models, and tech hubs would earn fees for facilitating things that people want to do. Some people are horrified by the idea of capitalism online, but this would be a more honest capitalism. The familiar "free" arrangement has been a disaster.

One of the reasons the tech community worries that A.I. could be an existential threat is that it could be used to toy with people, just as the previous wave of digital technologies have been. Given the power and potential reach of these new systems, it's not unreasonable to fear extinction as a possible result. Since that danger is widely recognized, the arrival of big-model A.I. could be an occasion to reformat the tech industry for the better.

Implementing data dignity will require technical research and policy innovation. In that sense, the subject excites me as a scientist. Opening the black box will only make the models more interesting. And it might help us understand more about language, which is the human invention that truly impresses, and the one that we are still exploring after all these hundreds of thousands of years.

Could data dignity address the economic worries that are often expressed about A.I.? The main concern is that workers will be devalued or displaced. Publicly, techies will sometimes say that, in the coming years, people who work with A.I. will be more productive and will find new types of jobs in a more productive economy. (A worker might become a prompt engineer for A.I. programs, for instance—someone who collaborates with or controls an A.I.) And yet, in private, the same people will quite often say, "No, A.I. will overtake this idea of collaboration." No more remuneration for today's accountants, radiologists, truck drivers, writers, film directors, or musicians.

A data-dignity approach would trace the most unique and influential contributors when a big model provides a valuable output. For instance, if you ask a model for "an animated movie of my kids in an oil-painting world of talking cats on an adventure," then certain key oil painters, cat portraitists, voice actors, and writers—or their estates—might be calculated to have been uniquely essential to the creation of the new masterpiece. They would be acknowledged and motivated. They might even get paid.

There is a fledgling data-dignity research community, and here is an example of a debate within it: How detailed an accounting should data dignity attempt? Not everyone agrees. The system wouldn't necessarily account for the billions of people who have made ambient contributions to big models—those who have added to a model's simulated competence with grammar, for example. At first, data dignity might attend only to the small number of special contributors who emerge in a given situation. Over time, though, more people might be included, as intermediate rights organizations—unions, guilds, professional groups, and so on—start to play a role. People in the data-dignity community sometimes call these anticipated groups mediators of individual data (*MIDs*) or data trusts. People need collective-bargaining power to have value in an online world—especially when they might get lost in a giant A.I. model. And when people share responsibility in a group, they self-police, reducing the need, or temptation, for governments and companies to censor or control from above. Acknowledging the human essence of big models might lead to a blossoming of new positive social institutions.

Data dignity is not just for white-collar roles. Consider what might happen if A.I.-driven tree-trimming robots are introduced. Human tree trimmers might find themselves devalued or even out of work. But the robots could eventually allow for a new type of indirect landscaping artistry. Some former workers, or others, might create inventive approaches—holographic topiary, say, that looks different from different angles—that find their way into the tree-trimming models. With data dignity, the models might create new sources of income, distributed through collective organizations. Tree trimming would become more multifunctional and interesting over time;

there would be a community motivated to remain valuable. Each new successful introduction of an A.I. or robotic application could involve the inauguration of a new kind of creative work. In ways large and small, this could help ease the transition to an economy into which models are integrated.

Many people in Silicon Valley see universal basic income as a solution to potential economic problems created by A.I. But U.B.I. amounts to putting everyone on the dole in order to preserve the idea of black-box artificial intelligence. This is a scary idea, I think, in part because bad actors will want to seize the centers of power in a universal welfare system, as in every communist experiment. I doubt that data dignity could ever grow enough to sustain all of society, but I doubt that any social or economic principle will ever be complete. Whenever possible, the goal should be to at least establish a new creative class instead of a new dependent class.

There are also non-altruistic reasons for A.I. companies to embrace data dignity. The models are only as good as their inputs. It's only through a system like data dignity that we can expand the models into new frontiers. Right now, it's much easier to get an L.L.M. to write an essay than it is to ask the program to generate an interactive virtual-reality world, because there are very few virtual worlds in existence. Why not solve that problem by giving people who add more virtual worlds a chance for prestige and income?

Could data dignity help with any of the human-annihilation scenarios? A big model could make us incompetent, or confuse us so much that our society goes collectively off the rails; a powerful, malevolent person could use A.I. to do us all great harm; and some people also think that the model itself could "jailbreak," taking control of our machines or weapons and using them against us.

We can find precedents for some of these scenarios not just in science fiction but in more ordinary market and technology failures. An example is the 2019 catastrophe related to Boeing's 737 *MAX* jets. The planes included a flight-path-correction feature that in some cases fought the pilots, causing two mass-casualty crashes. The problem was not the technology in isolation but the way that it was integrated into the sales cycle, training sessions, user interface, and documentation. Pilots thought that they were doing the right thing by trying to counteract the system in certain circumstances, but they were doing exactly the wrong thing, and they had no way of knowing. Boeing failed to communicate clearly about how the technology worked, and the resulting confusion led to disaster.

Anything engineered—cars, bridges, buildings—can cause harm to people, and yet we have built a civilization on engineering. It's by increasing and broadening human awareness, responsibility, and participation that we can make automation safe; conversely, if we treat our inventions as occult objects, we can hardly be good engineers. Seeing A.I. as a form of social collaboration is more actionable: it gives us access to the engine room, which is made of people.

Let's consider the apocalyptic scenario in which A.I. drives our society off the rails. One way this could happen is through deepfakes. Suppose that an evil person, perhaps working in an opposing government on a war footing, decides to stoke mass panic by sending all of us convincing videos of our loved ones being tortured or abducted from our homes. (The data necessary to create such videos are, in many cases, easy to obtain through social media or other channels.) Chaos would ensue, even if it soon became clear that the videos were faked. How could we prevent such a scenario? The answer is obvious: digital information must have context. Any collection of bits needs a history. When you lose context, you lose control.

Why don't bits come attached to the stories of their origins? There are many reasons. The original design of the Web didn't keep track of where bits came from, likely to make it easier for the network to grow quickly. (Computers and bandwidth were poor in the beginning.) Why didn't we start remembering where bits came from when it became more feasible to at least approximate digital provenance? It always felt to me that we

wanted the Web to be more mysterious than it needed to be. Whatever the reason, the Web was made to remember everything while forgetting its context.

Today, most people take it for granted that the Web, and indeed the Internet it is built on, is, by its nature, anti-contextual and devoid of provenance. We assume that decontextualization is intrinsic to the very idea of a digital network. That was never so, however; the initial proposals for digital-network architecture, put forward by the monumental scientist Vannevar Bush in 1945 and the computer scientist Ted Nelson in 1960, preserved provenance. Now A.I. is revealing the true costs of ignoring this approach. Without provenance, we have no way of controlling our A.I.s, or of making them economically fair. And this risks pushing our society to the brink.

If a chatbot appears to be manipulative, mean, weird, or deceptive, what kind of answer do we want when we ask why? Revealing the indispensable antecedent examples from which the bot learned its behavior would provide an explanation: we'd learn that it drew on a particular work of fan fiction, say, or a soap opera. We could react to that output differently, and adjust the inputs of the model to improve it. Why shouldn't that type of explanation always be available? There may be cases in which provenance shouldn't be revealed, so as to give priority to privacy—but provenance will usually be more beneficial to individuals and society than an exclusive commitment to privacy would be.

The technical challenges of data dignity are real and must inspire serious scientific ambition. The policy challenges would also be substantial—a sign, perhaps, that they are meaningful and concrete. But we need to change the way we think, and to embrace the hard work of renovation. By persisting with the ideas of the past—among them, a fascination with the possibility of an A.I. that lives independently of the people who contribute to it—we risk using our new technologies in ways that make the world worse. If society, economics, culture, technology, or any other spheres of activity are to serve people, that can only be because we decide that people enjoy a special status to be served.

This is my plea to all my colleagues. Think of people. People are the answer to the problems of bits. ♦

Jaron Lanier is a computer scientist, author, and musician. He is presently "Prime Unifying Scientist" at Microsoft, but does not speak for the company.
More:Artificial Intelligence (A.I.)TechnologyIntellectual PropertyData Privacy