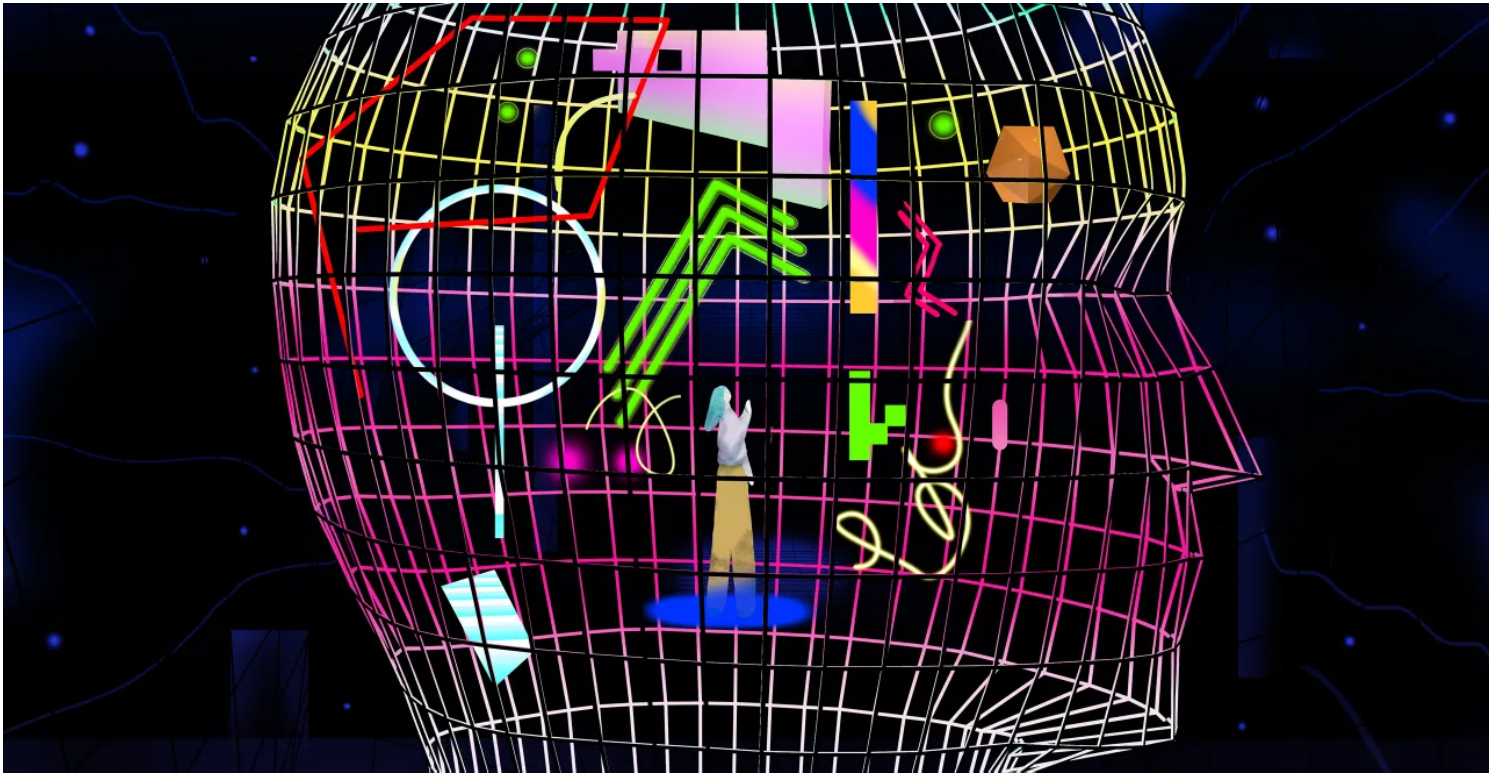


ANNALS OF TECHNOLOGY

THE HIDDEN COSTS OF AUTOMATED THINKING

By Jonathan Zittrain

July 23, 2019



Overreliance on artificial intelligence may put us in intellectual debt. Illustration by Jon Han

Like many medications, the wakefulness drug modafinil, which is marketed under the trade name Provigil, comes with a small, tightly folded paper pamphlet. For the most part, its contents—lists of instructions and precautions, a diagram of the drug’s molecular structure—make for anodyne reading. The

subsection called “Mechanism of Action,” however, contains a sentence that might induce sleeplessness by itself: “The mechanism(s) through which modafinil promotes wakefulness is unknown.”

Provigil isn’t uniquely mysterious. Many drugs receive regulatory approval, and are widely prescribed, even though no one knows exactly how they work. This mystery is built into the process of drug discovery, which often proceeds by trial and error. Each year, any number of new substances are tested in cultured cells or animals; the best and safest of those are tried out in people. In some cases, the success of a drug promptly inspires new research that ends up explaining how it works—but not always. Aspirin was discovered in 1897, and yet no one convincingly explained how it worked until 1995. The same phenomenon exists elsewhere in medicine. Deep-brain stimulation involves the implantation of electrodes in the brains of people who suffer from specific movement disorders, such as Parkinson’s disease; it’s been in widespread use for more than twenty years, and some think it should be employed for other purposes, including general cognitive enhancement. No one can say how it works.

This approach to discovery—answers first, explanations later—accrues what I call intellectual debt. It’s possible to discover what works without knowing why it works, and then to put that insight to use immediately, assuming that the underlying mechanism will be figured out later. In some cases, we pay off this intellectual debt quickly. But, in others, we let it compound, relying, for decades, on knowledge that’s not fully known.

In the past, intellectual debt has been confined to a few areas amenable to trial-and-error discovery, such as medicine. But that may be changing, as new techniques in artificial intelligence—specifically, machine learning—increase our collective intellectual credit line. Machine-learning systems work by identifying patterns in oceans of data. Using those patterns, they hazard answers to fuzzy, open-ended questions. Provide a neural network with labelled pictures of cats and

other, non-feline objects, and it will learn to distinguish cats from everything else; give it access to medical records, and it can attempt to predict a new hospital patient's likelihood of dying. And yet, most machine-learning systems don't uncover causal mechanisms. They are statistical-correlation engines. They can't explain why they think some patients are more likely to die, because they don't "think" in any colloquial sense of the word—they only answer. As we begin to integrate their insights into our lives, we will, collectively, begin to rack up more and more intellectual debt.

Theory-free advances in pharmaceuticals show us that, in some cases, intellectual debt can be indispensable. Millions of lives have been saved on the basis of interventions that we fundamentally do not understand, and we are the better for it. Few would refuse to take a life-saving drug—or, for that matter, aspirin—simply because no one knows how it works. But the accrual of intellectual debt has downsides. As drugs with unknown mechanisms of action proliferate, the number of tests required to uncover untoward interactions must scale exponentially. (If the principles by which the drugs worked were understood, bad interactions could be predicted in advance.) In practice, therefore, interactions are discovered after new drugs are on the market, contributing to a cycle in which drugs are introduced, then abandoned, with class-action lawsuits in between. In each individual case, accruing the intellectual debt associated with a new drug may be a reasonable idea. But intellectual debts don't exist in isolation. Answers without theory, found and deployed in different areas, can complicate one another in unpredictable ways.

Intellectual debt accrued through machine learning features risks beyond the ones created through old-style trial and error. Because most machine-learning models cannot offer reasons for their ongoing judgments, there is no way to tell when they've misfired if one doesn't already have an independent judgment about the answers they provide. Misfires can be rare in a well-trained system. But they can also be triggered intentionally by someone who knows just what kind of data to feed into that system.

Consider image recognition. Ten years ago, computers couldn't easily identify objects in photos. Today, image search engines, like so many of the systems we interact with on a day-to-day basis, are based on extraordinarily capable machine-learning models. Google's image search relies on a neural network called Inception. In 2017, M.I.T.'s LabSix—a research group of undergraduate and graduate students—succeeded in altering the pixels of a photograph of a cat so that, although it looked like a cat to human eyes, Inception became 99.99-percent sure it had been given a photograph of guacamole. (There was, it calculated, a slim chance that the photograph showed broccoli, or mortar.) Inception, of course, can't explain what features led it to conclude that a cat is a cat; as a result, there's no easy way to predict how it might fail when presented with specially crafted or corrupted data. Such a system is likely to have unknown gaps in its accuracy that amount to vulnerabilities for a smart and determined attacker.

As knowledge generated by machine-learning systems is put to use, these kinds of gaps may prove consequential. Health-care A.I.s have been successfully trained to classify skin lesions as benign or malignant. And yet—as a team of researchers from Harvard Medical School and M.I.T. showed, in a paper published this year—they can also be tricked into making inaccurate judgments using the same techniques that turn cats into guacamole. (Among other things, attackers might use these vulnerabilities to commit insurance fraud.) Seduced by the predictive power of such systems, we may stand down the human judges whom they promise to replace. But they will remain susceptible to hijacking—and we will have no easy process for validating the answers they continue to produce.

Could we create a balance sheet for intellectual debt—a system for tracking where and how theoryless knowledge is used? Our accounting could reflect the fact that not all intellectual debt is equally problematic. If an A.I. produces new pizza recipes, it may make sense to shut up and enjoy the pizza; by contrast,

when we begin using A.I. to make health predictions and recommendations, we'll want to be fully informed.

Building and maintaining a society-wide intellectual-debt balance sheet would probably require refining our approach to trade secrets and other intellectual property. In cities, building codes ask building owners to publicly disclose their renovation plans. Similarly, we might explore asking libraries or universities to accept, in escrow, otherwise hidden data sets and algorithms that have found a certain measure of public use. That would allow researchers to begin probing the models and underlying data on which we're coming to depend, and—by building theories—make payments on our intellectual debt before it becomes due in the form of errors and vulnerabilities.

The growing pervasiveness of machine-learning models, and the fact that anyone can create one, promise to make this process of accounting difficult. But it's vital. Taken in isolation, oracular answers can generate consistently helpful results. But these systems won't stay in isolation: as A.I.s gather and ingest the world's data, they'll produce data of their own—much of which will be taken up by still other systems. Just as drugs with unknown mechanisms of action sometimes interact, so, too, will debt-laden algorithms.

Even simple interactions can lead to trouble. In 2011, a biologist named Michael Eisen found out, from one of his students, that the least-expensive copy of an otherwise unremarkable used book—"The Making of a Fly: The Genetics of Animal Design"—was available on Amazon for \$1.7 million, plus \$3.99 shipping. The second-cheapest copy cost \$2.1 million. The respective sellers were well established, with thousands of positive reviews between them. When Eisen visited the book's Amazon page several days in a row, he discovered that the prices were increasing continually, in a regular pattern. Seller A's price was consistently 99.83 per cent that of Seller B; Seller B's price was reset, every day, to 127.059 per cent of Seller A's. Eisen surmised that Seller A had a copy of the book, and was seeking

to undercut the next-cheapest price. Seller B, meanwhile, didn't have a copy, and so priced the book higher; if someone purchased it, B could order it, on that customer's behalf, from A.

Each seller's presumed strategy was rational. It was the interaction of their algorithms that produced irrational results. The interaction of thousands of machine-learning systems in the wild promises to be much more unpredictable. The financial markets, where cutting-edge machine-learning systems are already being deployed, provide an obvious breeding ground for this type of problem. In 2010, for a harrowing thirty-six minutes, a "flash crash" driven by algorithmic trading wiped more than a trillion dollars from the major U.S. indices. Last fall, the J. P. Morgan analyst Marko Kolanovic argued that such a crash could easily happen again, since more trading than ever is based on automated systems. Intellectual debt can accumulate in the interstices where systems bump into each other, even when they don't formally interconnect. Without anything resembling a balance sheet, there's no way to determine—either in advance or retrospectively—whether any particular quantity of intellectual debt is worth taking on.

The increase in our intellectual debt may also involve a shift in the way we think—away from basic science and toward applied technology. Unlike, say, particle accelerators—massive capital projects which are supported by consortia of wealthy governments and run by academic research institutions—the tools of machine learning are as readily taken up by private industry as by academia. In fact, the sorts of data that produce useful predictions may be more readily available to Google and Facebook than to any computer-science or statistics department. Businesspeople may be perfectly satisfied by such unexplained knowledge, but intellectual debt will still be building. It will be held by corporations, far from the academic researchers who might be most interested in paying it down.

It's easy to imagine that the availability of machine-learning-based knowledge will shift funding away from researchers who insist on the longer route of trying to

figure things out for themselves. This past December, Mohammed AlQuraishi, a researcher who studies protein folding, wrote an essay exploring a recent development in his field: the creation of a machine-learning model that can predict protein folds far more accurately than human researchers. AlQuraishi found himself lamenting the loss of theory over data, even as he sought to reconcile himself to it. “There’s far less prestige associated with conceptual papers or papers that provide some new analytical insight,” he said, in an interview. As machines make discovery faster, people may come to see theoreticians as extraneous, superfluous, and hopelessly behind the times. Knowledge about a particular area will be less treasured than expertise in the creation of machine-learning models that produce answers on that subject.

Financial debt shifts control—from borrower to lender, and from future to past. Mounting intellectual debt may shift control, too. A world of knowledge without understanding becomes a world without discernible cause and effect, in which we grow dependent on our digital concierges to tell us what to do and when. It’s easy to imagine, for example, how a college-admissions committee might turn the laborious and uncertain sifting of applicants over to a machine-learning model; such a model might purport to optimize an entering cohort not just for academic success but also for harmonious relationships and generous alumni donations. The only way to make sense of this world might be to employ our own A.I.s—neural nets that fine-tune our social-media profiles so that we seem like we’ll fit perfectly into the freshman class.

Perhaps all this technology will work—and that, in turn, will be a problem. Much of the timely criticism of artificial intelligence has rightly focussed on the ways in which it can go wrong: it can create or replicate bias; it can make mistakes; it can be put to evil ends. We should also worry, though, about what will happen when A.I. gets it right.

THE DAILY

The best of *The New Yorker*, every day, in your in-box, plus occasional alerts when we publish major stories.

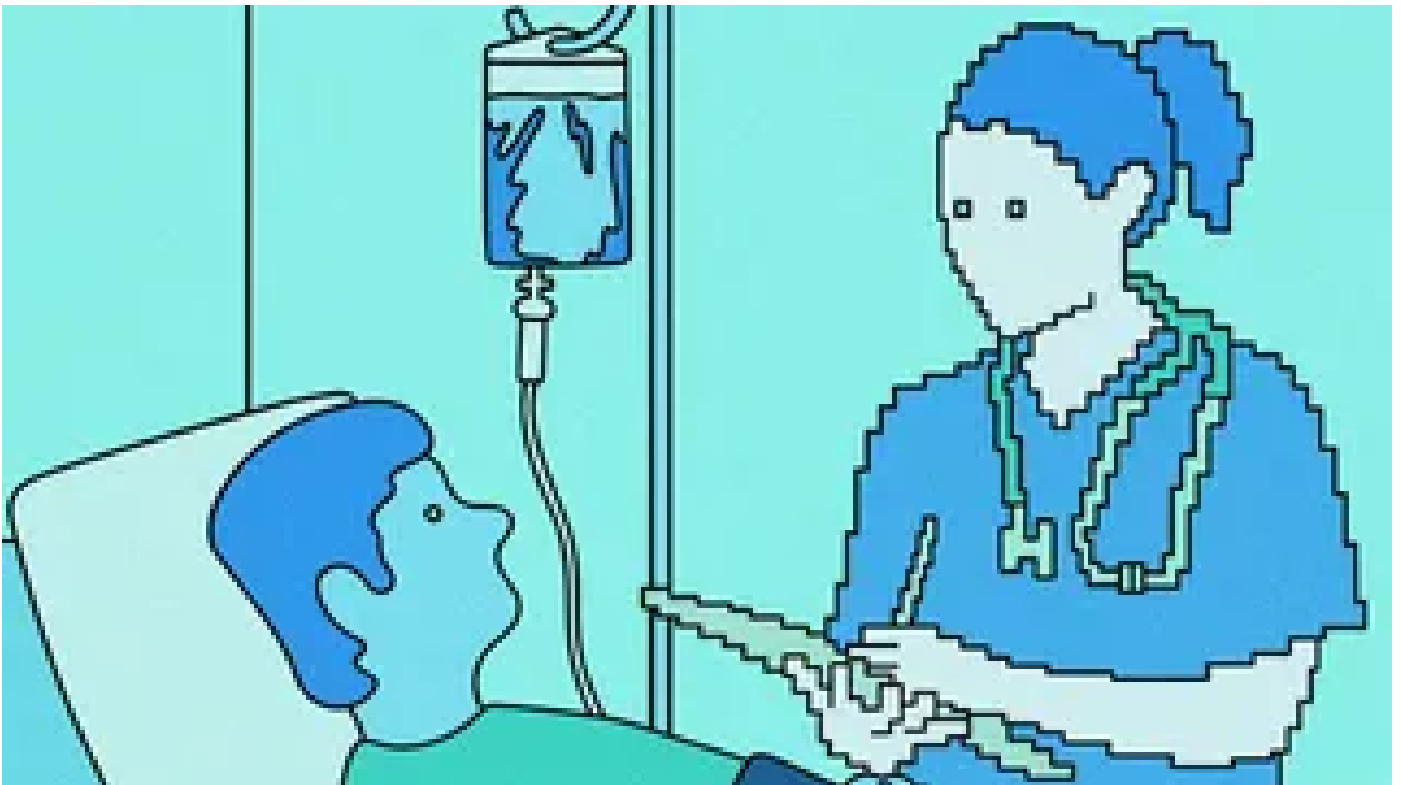
E-mail address

Your e-mail address

Sign up

By signing up, you agree to our [User Agreement](#) and [Privacy Policy & Cookie Statement](#).

Read More

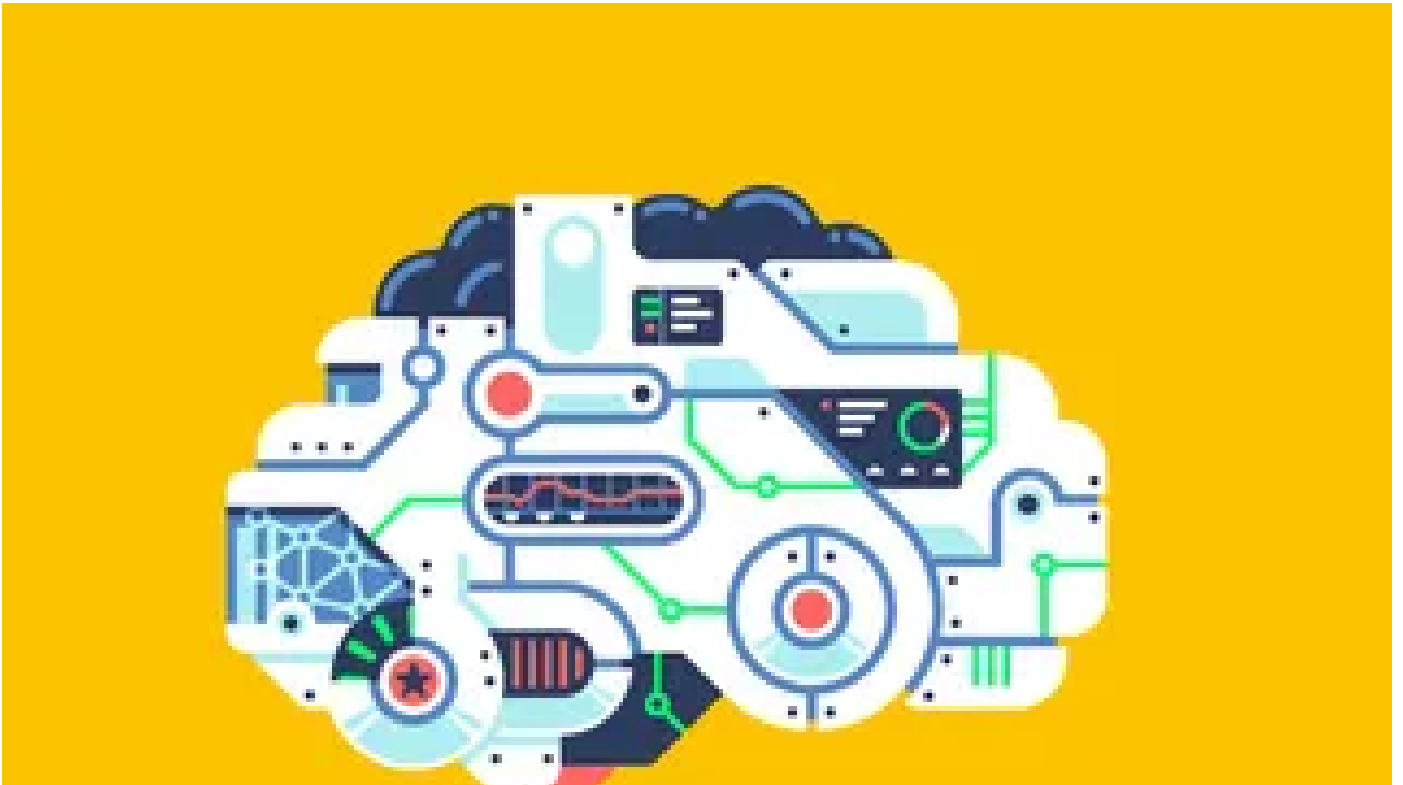


ANNALS OF TECHNOLOGY

AUTOMATED HEALTH CARE OFFERS FREEDOM FROM SHAME, BUT IS IT WHAT PATIENTS NEED?

We often respond more openly to computers and robots than we do to our fellow-humans. Yet some ethicists worry that relying too much on A.I. could be dangerous.

By Allison J. Pugh



TECH

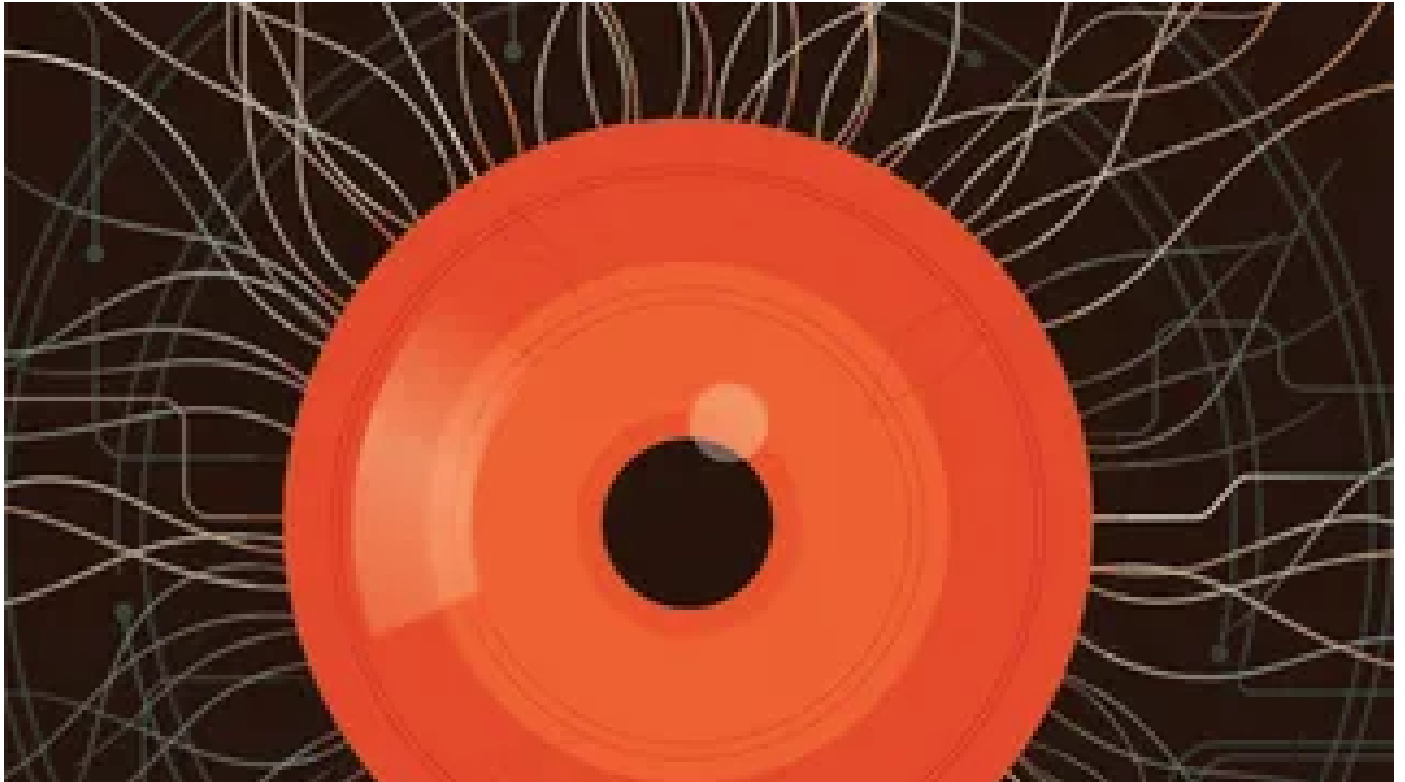
A COMPUTER TO RIVAL THE BRAIN

By Kelly Clancy



HOW TO MAKE A SYNTHETIC SELF

Using artificial intelligence, researchers can project the movements of one body onto another's.



DEPT. OF SPECULATION

HOW FRIGHTENED SHOULD WE BE OF A.I.?

Thinking about artificial intelligence can help clarify what makes us human—for better and for worse.

By Tad Friend

Your Privacy Choices