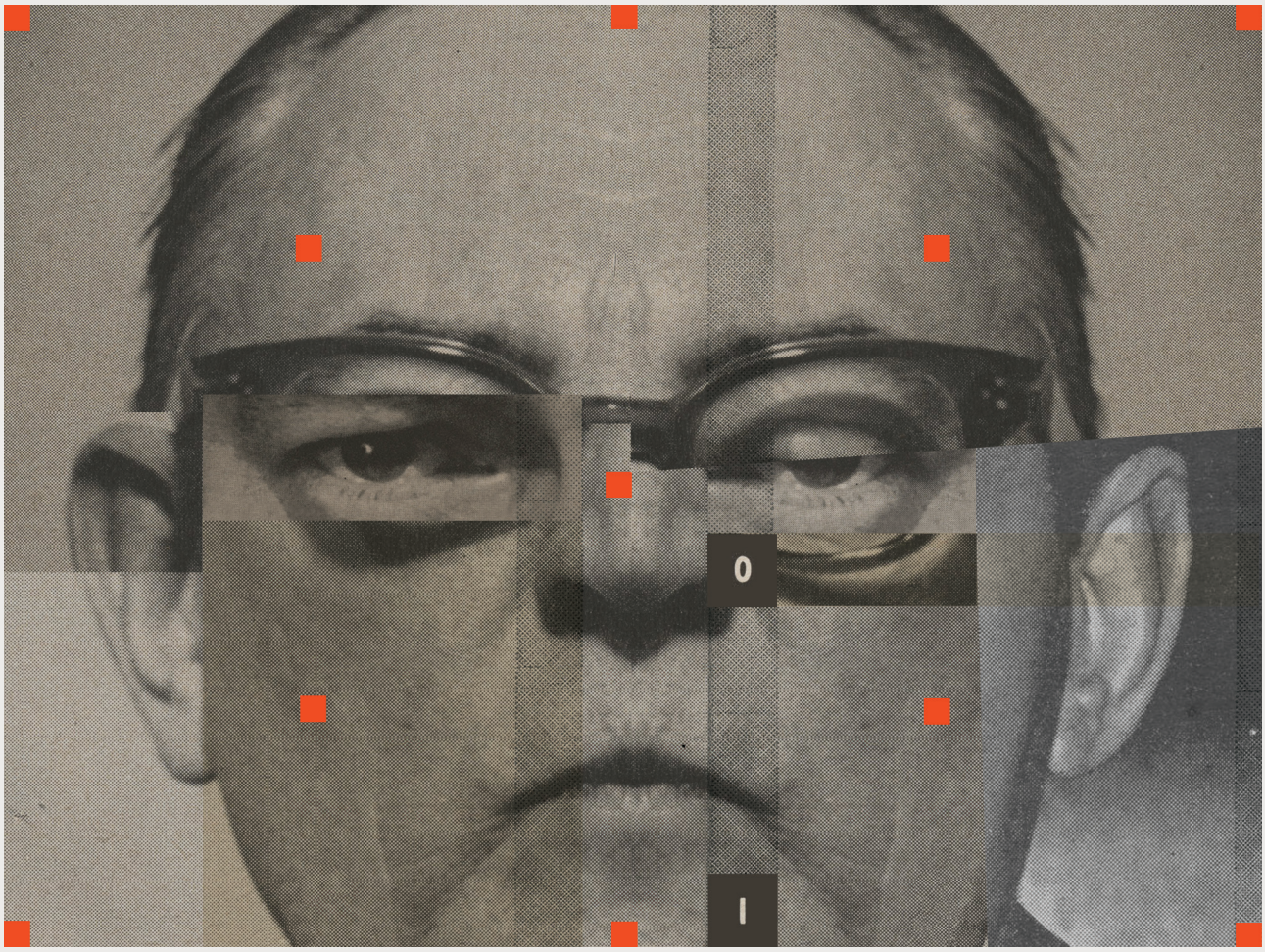
News



ARTIFICIAL INTELLIGENCE

AI's Real Worst-Case Scenarios › Who needs Terminators when you have precision clickbait and ultra-deepfakes?

BY NATASHA BAJEMA

Hollywood's worst-case scenario involving artificial intelligence (AI) is as familiar as any trope in blockbuster movies: Machines acquire humanlike intelligence, achieving sentience, and inevitably turn into evil overlords that attempt to destroy the human race. This narrative capitalizes on our innate fear of technology, a reflection of the profound change that often accompanies new technological developments.

However, as Malcolm Murdock, machine-learning engineer and author of the 2019 novel *The Quantum Price,* puts it, "AI doesn't have to be sentient to kill

Photo-Illustration by Mike McQuade

us all. There are plenty of other scenarios that will wipe us out before sentient AI becomes a problem."

In interviews with AI experts, *IEEE Spectrum* has uncovered six real-world AI worst-case scenarios that are far more mundane than those depicted in the movies. But they're no less dystopian. And most don't require a malevolent dictator to bring them to full fruition. Rather, they could simply happen by default, unfolding naturally—that is, if nothing is done to stop them. To prevent these worst-case scenarios, we must abandon our pop-culture notions of AI and get serious about its unintended consequences.

1. When Fiction Defines Our Reality...

Unnecessary tragedy may strike if we allow fiction to define our reality. But what choice is there when we can't tell the difference between what is real and what is false in the digital world?

In a terrifying scenario, the rise of deepfakes—fake images, video, audio, and text generated with advanced machine-learning tools—may someday lead national-security decision-makers to take real-world action based on false information, leading to a major crisis, or worse yet, a war.

Andrew Lohn, senior fellow at Georgetown University's Center for Security and Emerging Technology (CSET), says that "AI-enabled systems are now capable of generating disinformation at [large scales]." By producing greater volumes and variety of fake messages, these systems can obfuscate their true nature and optimize for success, improving their desired impact over time.

The mere notion of deepfakes amid a crisis might also cause leaders to hesitate to act if the validity of information cannot be confirmed in a timely manner.

Marina Favaro, research fellow at the Institute for Research and Security Policy in Hamburg, Germany, notes that "deepfakes compromise our trust in information streams by default." Both action and inaction caused by deepfakes have the potential to produce disastrous consequences for the world.

"We are entering dangerous and uncharted territory with the rise of surveillance and tracking through data, and we have almost no understanding of the potential implications."

—ANDREW LOHN, GEORGETOWN UNIVERSITY

2. A Dangerous Race to the Bottom

When it comes to AI and national security, speed is both the point and the problem. Since AI-enabled systems confer greater speed benefits on its users, the first countries to develop military applications will gain a strategic advantage. But what design principles might be sacrificed in the process?

Things could unravel from the tiniest flaws in the system and be exploited by hackers. Helen Toner, director of strategy at CSET, suggests a crisis could "start off as an innocuous single point of failure that makes all communications go dark, causing people to panic and economic activity to come to a standstill. A persistent lack of information, followed by other miscalculations, might lead a situation to spiral out of control."

Vincent Boulanin, senior researcher at the Stockholm International Peace Research Institute (SIPRI), in Sweden, warns that major catastrophes can occur "when major powers cut corners in order to win the advantage of getting there first. If one country prioritizes speed over safety, testing, or human oversight, it will be a dangerous race to the bottom."

For example, national-security leaders may be tempted to delegate decisions of command and control, removing human oversight of machine-learning models that we don't fully understand, in order to gain a speed advantage. In such a scenario, even an automated launch of missile-defense systems initiated without human authorization could produce unintended escalation and lead to nuclear war.

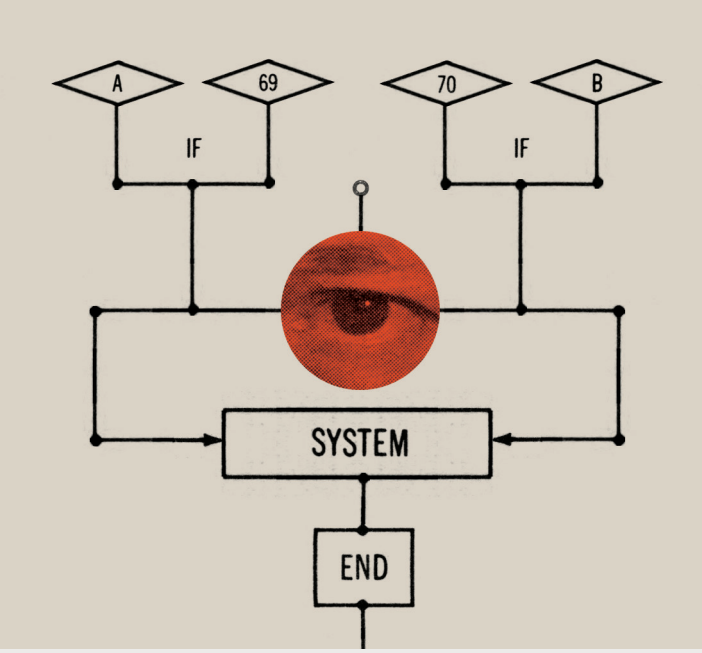
3. The End of Privacy and Free Will

With every digital action, we produce new data—emails, texts, downloads, purchases, posts, selfies, and GPS locations. By allowing companies and governments to have unrestricted access to this data, we are handing over the tools of surveillance and control.

With the addition of facial recognition, biometrics, genomic data, and AI-enabled predictive analysis, Lohn of CSET worries that "we are entering dangerous and uncharted territory with the rise of surveillance and tracking through data, and we have almost no understanding of the potential implications."

Michael C. Horowitz, director of Perry World House, at the University of Pennsylvania, warns "about the logic of AI and what it means for domestic repression. In the past, the ability of autocrats to repress their populations relied upon a large group of soldiers, some of whom may side with society and carry out a coup d'etat. AI could reduce these kinds of constraints."

The power of data, once collected and analyzed, extends far beyond the functions of monitoring and surveillance to allow for predictive control. Today, AI-enabled systems predict what products we'll purchase, what entertainment we'll watch, and what links we'll click. When these platforms know us far better than we know ourselves, we may not notice the slow creep that robs us of our



free will and subjects us to the control of external forces.

4. A Human Behavioral Experiment

The ability of children to delay immediate gratification, to wait for the second marshmallow, was once considered a major predictor of success in life. Soon even the second-marshmallow kids will succumb to the tantalizing conditioning of engagement-based algorithms.

Social media users have become rats in lab experiments, living in human Skinner boxes, glued to the screens of their smartphones, compelled to sacrifice more precious time and attention to platforms that profit from it at their expense.

Helen Toner of CSET says that "algorithms are optimized to keep users on the platform as long as possible." By offering rewards in the form of likes, comments, and follows, Malcolm Murdock explains, "the algorithms short-circuit the way our brain works, making our next bit of engagement irresistible."

To maximize advertising profit, companies steal our attention away from our jobs, families and friends, responsibilities, and even our hobbies. To make matters worse, the content often makes us feel miserable and worse off than before. Toner warns that "the more time we spend on these platforms, the less time we spend in the pursuit of positive, productive, and fulfilling lives."

5. The Tyranny of AI Design

Every day, we turn over more of our daily lives to AI-enabled machines. This is problematic since, as Horowitz observes, "we have yet to fully wrap our heads around the problem of bias in AI. Even with the best intentions, the design of AI-enabled systems, both the training data and the mathematical models, reflects the narrow experiences and interests of the biased people who program them. And we all have our biases."

As a result, Lydia Kostopoulos, senior vice president of emerging tech insights at the Clearwater, Fla.–based IT security company KnowBe4, argues that "many AI-enabled systems fail to take into account the diverse experiences and characteristics of different people." Since AI solves problems based on biased perspectives and data rather than the unique needs of every individual, such systems produce a level of conformity that doesn't exist in human society.

Even before the rise of AI, the design of common objects in our daily lives has often catered to a particular type of person. For example, studies have shown that cars, hand-held tools including cellphones, and even the temperature settings in office environments have been established to suit the average-size man, putting people of varying sizes and body types, including women, at a major disadvantage and sometimes at greater risk to their lives.

When individuals who fall outside of the biased norm are neglected, marginalized, and excluded, AI turns into a Kafkaesque gatekeeper, denying access to customer service, jobs, health care, and much more. AI design decisions can restrain people rather than liberate them from day-to-day concerns. And these choices can also transform some of the worst human prejudices into racist and sexist hiring and mortgage practices, as well as deeply flawed and biased sentencing outcomes.

6. Fear of AI Robs Humanity of Its Benefits

Since AI's capabilities of course scale with the computing power and complexity of the hardware it runs on, societal fears around AI seem poised only to grow over time. "Artificial neural networks can do insanely powerful things," said Murdock, "and we need to be prudent about the risks." But what if people become so afraid of AI that governments regulate it in ways that rob humanity of AI's many benefits? For example, DeepMind's AlphaFold program achieved a major breakthrough in predicting how amino acids fold into proteins, making it possible for scientists to identify the structure of 98.5 percent of human proteins. This milestone will provide a fruitful foundation for the rapid advancement of the life sciences. Consider the benefits of improved communication and cross-cultural understanding made possible by seamlessly translating across any combination of human languages, or the use of AI-enabled systems to identify new treatments and cures for disease. Knee-jerk regulatory actions by governments to protect against AI's worst-case scenarios could also backfire and produce their own unintended negative consequences, in which we become so scared of the power of this tremendous technology that we resist harnessing it for the actual good it can do in the world. ■

Photo-Illustration by Mike McQuade

Microsoft's office in Beijing houses a company division that trained many of China's present-day AI and technology-industry titans.

NG HAN GUAN/AP

TECH POLICY

U.S.-China Rivalry Boosts Tech—and Tensions ›

One-upmanship can even be productive, until militaries get involved

BY CRAIG S. SMITH

In June 2020, OpenAI, an independent artificial-intelligence research lab based in San Francisco, announced GPT-3, the third generation of its massive Generative Pre-trained Transformer language model, which can write everything from computer code to poetry.

A year later, with much less fanfare, Tsinghua University's Beijing Academy of Artificial Intelligence released an even larger model, Wu Dao 2.0, with 10 times as many parameters—the neural network values that encode information. While GPT-3 boasts 175 billion parameters, Wu Dao 2.0 has a whopping 1.75 trillion—though not directly comparable. Moreover, the model is capable not only of generating text like GPT-3 does but also images from textual descriptions like OpenAI's 12-billion-parameter DALL-E model, and has a scaling strategy similar to Google's 1.6-trillion-parameter Switch Transformer model.

Tang Jie, the Tsinghua University professor leading the Wu Dao project, said in a recent interview that the group built an even bigger, 100-trillion-parameter model in June, though they have not trained it to "convergence," the point at which the model stops improving. "We just wanted to prove that we have the ability to do that," Tang said.

This isn't simple one-upmanship. On the one hand, it's how research progresses. But on the other, it is emblematic of an intensifying competition between the world's two technology superpowers. Whether the researchers involved like it or not, their governments are eager to adopt each AI advance into their national-security infrastructure and military capabilities.

That matters, because dominance in the technology surely improves the odds of victory in any future war. Such an advantage also likely guarantees the longevity and global influence of the government that wields it. Already, China is exporting its AI-enabled surveillance technology—which can be used to quash dissent—to client states and is espousing an authoritarian model that promises economic prosperity as a counter to democracy, something that the Soviet Union was never able to do.

Ironically, China is a competitor that the United States abetted. It's well known that the U.S. consumer market fed China's export engine, itself outfitted with U.S. machines, and led to the fastest-growing economy in the world since the 1980s. What's less well-known is how a handful of technology companies transferred

the know-how and trained the experts now giving the United States a run for its money in AI.

Blame Bill Gates, for one. In 1992, Gates led Microsoft into China's fledgling software market. Six years later, he established Microsoft Research Asia, the company's largest basic and applied computer-research institute outside the United States. People from that organization have gone on to found or lead many of China's top technology institutions.

For instance, in 2012 Zhang Yiming, a Microsoft Research Asia alum, founded the video-sharing platform's parent company, ByteDance, developer and operator of the social media platform TikTok. He hired a former head of Microsoft Research Asia, Zhang Hongjiang, to lead ByteDance's Technical Strategy Research Center. This Zhang is now head of the Beijing Academy— the organization behind Wu Dao 2.0, currently the largest AI system on the planet. That back-and-forth worries U.S. national-security strategists, who plan for a day when researchers and companies are forced to take sides.

Today's competition has roots in an incident on 7 May 1999, when a U.S. B-2 Stealth Bomber dropped bombs on the Chinese embassy in Belgrade, Serbia, killing three people.

"That's when the Chinese started saying, 'We're moving beyond attrition warfare' to what they referred to as systems confrontation, the confrontation between their operational system and the American operational system," says Robert O. Work, former U.S. Deputy Secretary of Defense and vice chairman of the recently concluded National Security Commission on Artificial Intelligence. "Their theory of victory is what they refer to as system destruction."

China's "theory of victory is what they refer to as system destruction."

—ROBERT O. WORK, FORMER U.S. DEPUTY SECRETARY OF DEFENSE

System-destruction warfare is part and parcel of what the People's Liberation Army thinks of as "intelligentized" warfare, in which war is waged not only in the traditional physical domains of land, sea, and air but also in outer space, nonphysical cyberspace, and electromagnetic and even psychological domains—all enabled and coordinated with AI.

Work says the first major U.S. AI effort toward intelligentized warfare was to use computer vision to analyze thousands of hours of full-motion video being downloaded from dozens of drones. Today, that effort, dubbed Project Maven, detects, classifies, and tracks objects within video images, and it has been extended to acoustic data and signals intelligence.

The Chinese have kept pace. China is actively pursuing AI-based target recognition and automatic-weapon-firing research, which could be used in lethal autonomous weapons. Meanwhile, according to Work, the country's swarm technology may be ahead of the United States—whose military budget, nevertheless, is three times that of China.

Chinese firm Baidu—whose comparatively modest Sunnyvale, Calif., office is pictured here in 2018—is one of the largest Internet companies in the world.

"I worry about their emphasis on swarms of unmanned systems," says Work, adding that the Chinese want to train swarms of a hundred vehicles or more, including underwater systems, to coordinate navigation through complex environments. "While we also test swarms, we have yet to demonstrate the ability to employ these types of swarms in a combat scenario."

This type of research and testing has prompted calls for preemptive bans on lethal autonomous weapons, but neither country is willing to declare an outright prohibition. Barring a prohibition, many people believe that China and the United States, along with other countries, should begin negotiating an arms-control agreement banning the development of systems that could autonomously order a preemptive or retaliatory attack. Such systems might inadvertently lead to "flash wars," just as AI-driven autonomous trading has led to flash crashes in the financial markets.

"Neither of us wants to get into a war because an autonomous-control system made a mistake and ordered a preemptive strike," Work says, referring to the United States and China.

All of this contributes to a dilemma facing the twin realms of AI research and military modernization. The international research community, collaborative and collegial, prefers to look the other way and insist that it only serves the interest of science. But the governments that fund that research have clear agendas, and military enhancement is undeniably one.

Geoffrey Hinton, regarded as one of the godfathers of deep learning, the kind of AI transforming militaries today, left the United States and moved to Canada largely because he didn't want to depend on funding from the Defense Advanced Research Projects Agency, or DARPA. The agency, the largest funder of AI research in the world, is responsible for the development of emerging technologies for military use.

Hinton instead helped to put deep learning on the map in 2012 with a

SMITH COLLECTION/GADO/GETTY IMAGES

now-famous neural net called AlexNet when he was at the University of Toronto. But Hinton was also in close contact with the Microsoft Research Lab in Redmond, Wash., before and after his group validated AlexNet, according to one of Hinton's associates there, Li Deng, then principal researcher and manager and later chief scientist of AI at Microsoft. When Hinton achieved his 2012 breakthrough, news soon spread through Microsoft's Chinese brain trust to China.

The United States has since tried to limit this cross-pollination, barring Chinese nationals known to have worked for China's military or intelligence organizations from working with U.S. research institutions. But research continues to flow back and forth between the two countries: Microsoft maintains its research lab in Beijing, and the Chinese Internet and AI giant Baidu has a research lab in Silicon Valley, for example.

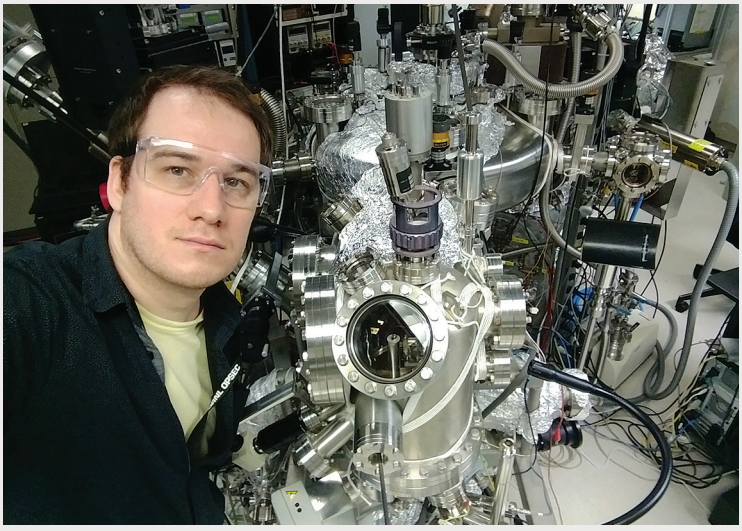
Tsinghua University's Tang said decoupling the two countries would slow China's AI research—not because it would stop the flow of ideas, but because it would cut China off from the advanced semiconductors needed to train AI models. "We hope that we can do science for the world, not just the one country," Tang says. But, he added, "we should do something on demand based on the national project research plan."

China's National Intelligence Law compels its companies and researchers to cooperate when asked. China began pouring billions of dollars into AI research in 2017, and among the organizations set up with that funding was Tsinghua's Beijing Academy, where Tang and his team built Wu Dao 2.0.

By most metrics, Wu Dao 2.0 has surpassed OpenAI's GPT-3. Tang says it was trained on 4.9 terabytes of clean data, including Chinese-language text, English-language text, and images. OpenAI has said that GPT-3 was trained on just 570 gigabytes of clean, primarily English-language text.

Tang says his group is now working on video with the goal of generating realistic video from text descriptions. "Hopefully, we can make this model do something beyond the Turing test," he says, referring to an assessment of whether a computer can generate text indistinguishable from that created by a human. "That's our final goal." ■

MAXIM ZIATDINOV



Maxim Ziatdinov of Oak Ridge National Laboratory sees automated microscopy ultimately becoming a crucial tool for next-generation quantum computers.

MICROSCOPY

Navigating the Nanoscale ›

Deep learning enables push toward self-driving microscopes

BY DAN GARISTO

It's difficult to find an area of scientific research where deep learning isn't discussed as the next big thing. Claims abound: Deep learning will spot cancers; it will unravel complex protein structures; it will reveal new exoplanets in previously analyzed data; it will even discover a theory of everything. Knowing what's real and what's just hype isn't always easy.

One promising—perhaps even overlooked—area of research for deep learning to make its mark is in microscopy. In spite of new discoveries, the underlying workflow of techniques like scanning probe microscopy (SPM) and scanning transmission electron microscopy (STEM) has remained largely unchanged for decades. Skilled human operators must painstakingly set up, observe, and analyze samples. Deep learning has the potential not only to automate many of the tedious tasks, but also to dramatically speed up the analysis time by honing in on microscopic features of interest.

"People usually just look at the image and identify a few properties of interest," says Maxim Ziatdinov, a researcher at Oak Ridge National Laboratory. "They basically discard most of the information because there is just no way to actually extract all the features of interest from the data." With deep learning, Ziatdinov says, it's possible to extract information about the position and type of atomic structures in seconds, opening up a vista of possibilities.

It's a twist on the classical dream of doing more with smaller things (most famously expressed in Richard Feynman's "There's Plenty of Room at the Bottom"). Improving hardware isn't the only way to increase the functionality of micro-

scopes. Software can play a role, too—by making a microscope autonomous. "Such a machine will 'understand' what it is looking at and automatically document features of interest," an article in the *Materials Research Society Bulletin* declared. "The microscope will know what various features look like by referencing databases or can be shown examples on the fly."

Despite the "micro-" prefix, microscopy such as SPM and STEM actually deals with objects on the nanoscale, including individual atoms. In SPM, a nanoscale tip hovers over the sample surface and traces its grooves, like the needle of a record player, to create an image. On the other hand, STEM generates an image by showering a sample with electrons and collecting those that pass through, essentially creating a negative.

Both microscopy techniques allow researchers to quickly observe the broad structural features of a sample. Researchers like Ziatdinov are interested in the functional properties of certain features such as defects. By applying a stimulus like an electric field to the sample, they can measure how it responds and build a functional map, too.

But zooming in on a structural image to gather functional data is time-prohibitive, and human operators have to make a guess about which features they choose to analyze. There hasn't been a rigorous way to predict functionality from structure, so operators have simply had to get a knack for picking good features.

The hope is that this tedious feature-picking can be outsourced to a neural network that predicts features of interest and navigates to them, dramatically speeding up the process.

Automated microscopy is still at the proof-of-concept stage, with a few groups of researchers around the world hammering out the principles and doing preliminary tests. Unlike many areas of deep learning, success here would not be simply automating preexisting measurements; with automation, researchers could make measurements that have been inaccessible.

Ziatdinov and his colleagues have already made some progress toward such a future. For years, they sat on microscopy data of graphene—a few frames showing a defect that created strain in the atomically thin material. "We couldn't analyze it, because there's just no way that you can extract positions of all the atoms," Ziatdinov says. But by training a neural net on the graphene, they were able to categorize new structures on the edges of defects.

Microscopy isn't just limited to observing. By blasting samples with a high-energy electron beam, researchers can shift the position of atoms, effectively creating an "atomic forge." As with a conventional billows-and-iron forge, automation could make things a lot easier. An atomic forge guided by deep learning could spot defects and fix them, or nudge atoms into place to form intricate structures—around the clock, without human error, sweat, or tears.

"If you actually want to have a manufacturing capability, just like with any other kind of manufacturing, you need to be able to automate it," Ziatdinov says.

Ziatdinov is particularly interested in applying automated microscopy to quantum devices, like topological qubits. Efforts to create these qubits have not proved successful, but he thinks he might have the answer. By training a neural network to understand the functions associated with specific features, deep learning could unlock what atomic tweaks are needed to create a topological qubit—something humans clearly haven't quite figured out.

Benchmarking exactly how far we are from a future where autonomous microscopy helps build quantum devices isn't easy. There are only a few human operators in the entire world, so it's difficult to compare deep-learning results to a human average. It's also unclear which obstacles will pose the biggest problems moving forward in a domain where the difference of a few atoms can be decisive.

The conclusion of a recent review of the prospects of autonomous microscopy argues it "will enable fundamentally new opportunities and paradigms for scientific discovery," with the caveat that "this process is likely to be highly nontrivial." Whether deep learning lives up to its promise on the microscopic frontier remains, literally, to be seen. ■

JOURNAL WATCH

This RISC-V Powerhouse Goes Light on the Power

As society's insatiable demand for computing power continues to grow, so too does the need for more efficient processors. A group of researchers in Switzerland has devised a new processor design that may help. It is physically small and computationally agile—and aptly named Snitch. (Harry Potter fans will get the reference.)

Florian Zaruba, a postdoc at the Integrated Systems Laboratory at the Swiss Federal Institute of Technology (ETH), in Zurich—and a researcher involved in the creation of Snitch—notes that commercial, general-purpose cores today rely on larger and more energy-hungry processors. "Snitch is the opposite," he says.

Typically, processors try to find an efficient instruction order on the fly, which requires additional hardware and thus uses more power. But Snitch is able to execute the majority of its basic instructions instantaneously, bypassing the need for this extra, burdensome hardware.

Zaruba and his colleagues describe their streamlined, RISC-V chip design in a study published 7 October in *IEEE Transactions on Computers*. They found that a single Snitch processor with its custom extensions was twice as energy efficient as comparable benchmark CPUs. When multiple processors were used in parallel, Snitch proved to be 3.5 times as energy efficient and up to six times as fast as the others.

The researchers have open-sourced Snitch's hardware design and note that they have seen growing interest from industry consortia, for example from the Open Hardware Group, in supporting commercialization efforts. —Michelle Hampson