Killer Apps

The Real Dangers of an AI Arms Race

Paul Scharre

The nation that leads in the development of artificial intelligence will, Russian President Vladimir Putin proclaimed in 2017, "become the ruler of the world." That view has become commonplace in global capitals. Already, more than a dozen governments have announced national AI initiatives. In 2017, China set a goal of becoming the global leader in AI by 2030. Earlier this year, the White House released the American AI Initiative, and the U.S. Department of Defense rolled out an AI strategy.

But the emerging narrative of an "AI arms race" reflects a mistaken view of the risks from AI—and introduces significant new risks as a result. For each country, the real danger is not that it will fall behind its competitors in AI but that the perception of a race will prompt everyone to rush to deploy unsafe AI systems. In their desire to win, countries risk endangering themselves just as much as their opponents.

AI promises to bring both enormous benefits, in everything from health care to transportation, and huge risks. But those risks aren't something out of science fiction; there's no need to fear a robot uprising. The real threat will come from humans.

Right now, AI systems are powerful but unreliable. Many of them are vulnerable to sophisticated attacks or fail when used outside the environment in which they were trained. Governments want their systems to work properly, but competition brings pressure to cut corners. Even if other countries aren't on the brink of major AI breakthroughs, the perception that they're rushing ahead could push others to do the same. And if a government deployed an untested AI weapons

PAUL SCHARRE is a Senior Fellow and Director of the Technology and National Security Program at the Center for a New American Security. He is the author of *Army of None: Autonomous Weapons and the Future of War*.

system or relied on a faulty AI system to launch cyberattacks, the result could be disaster for everyone involved.

Policymakers should learn from the history of computer networks and make security a leading factor in AI design from the beginning. They should also ratchet down the rhetoric about an AI arms race and look for opportunities to cooperate with other countries to reduce the risks from AI. A race to the bottom on AI safety is a race no one would win.

THE AIS HAVE IT

The most straightforward kind of AI system performs tasks by following a series of rules set in advance by humans. These "expert systems," as they are known, have been around for decades. They are now so ubiquitous that we hardly stop to think of the technology behind airplane autopilots or tax-preparation software as AI. But in the past few years, advances in data collection, computer processing power, and algorithm design have allowed researchers to make big progress with a more flexible AI method: machine learning.

In machine learning, a programmer doesn't write the rules; the machine picks them up by analyzing the data it is given. Feed an algorithm thousands of labeled photos of objects, and it will learn to associate the patterns in the images with the names of the objects. The current AI boom began in 2012, when researchers made a breakthrough using a machine-learning technique called "deep learning," which relies on deep neural networks. Neural networks are an AI technique loosely inspired by biological neurons, the cells that communicate with other cells by sending and receiving electrical impulses. An artificial neural network starts out as a blank slate; it doesn't know anything. The system learns by adjusting the strength of the connections between neurons, strengthening certain pathways for right answers and weakening the connections for wrong answers. A deep neural network—the type responsible for deep learning—is a neural network with many layers of artificial neurons between the input and output layers. The extra layers allow for more variability in the strengths of different pathways and thus help the AI cope with a wider variety of circumstances.

How exactly the system learns depends on which machine-learning algorithm and what kind of data the developers use. Many approaches use data that are already labeled (known as "supervised learning"), but machines can also learn from data that are not labeled ("unsupervised

Seeing like a state: a SenseTime surveillance software demo in Beijing, October 2017

THOMAS PETER / REUTERS

learning") or directly from the environment ("reinforcement learning"). Machines can also train on synthetic, computer-generated data. The autonomous car company Waymo has driven its cars for over ten million miles on public roads, but the company clocks ten million miles every day in computer simulations, allowing it to test its algorithms on billions of miles of synthetic data.

Since the deep-learning breakthrough in 2012, researchers have created AI systems that can match or exceed the best human performance in recognizing faces, identifying objects, transcribing speech, and playing complex games, including the Chinese board game go and the real-time computer game StarCraft. Deep learning has started to outstrip older, rules-based AI systems, too. In 2018, a deep-learning algorithm beat the reigning chess computer program after spending just four hours playing millions of games against itself on a massive supercomputer without any human training data or hand-coded rules to guide its behavior.

Researchers are now applying AI to a host of real-world problems, from diagnosing skin cancers to driving cars to improving energy efficiency. According to an estimate by the consulting firm McKinsey, almost half of all the tasks people are paid to perform in the United States could be automated with existing technology (although less than five percent of jobs could be eliminated entirely). AI tools are also becoming more widely available. Large organizations are the most

likely to make major breakthroughs, thanks to their ability to amass large data sets and huge quantities of computing power. But many of the resulting AI tools are available online for anyone to use. Free programming courses teach people how to make their own AI systems, and trained neural networks are free to download. Accessibility will spur innovation, but putting powerful AI tools into the hands of anyone who wants them will also help those who set out to do harm.

AUTOCRATIC INTELLIGENCE

Harm from AI misuse isn't hypothetical; it's already here. Bots are regularly used to manipulate social media, amplifying some messages and suppressing others. Deepfakes, AI-generated fake videos, have been used in so-called revenge porn attacks, in which a person's face is digitally grafted onto the body of a pornographic actor.

These examples are only the start. Political campaigns will use AI-powered data analytics to target individuals with political propaganda tailored just for them. Companies will use the same analytics to design manipulative advertising. Digital thieves will use AI tools to create more effective phishing attacks. Bots will be able to convincingly impersonate humans online and over the phone by cloning a person's voice with just a minute of audio. Any interaction that isn't in person will become suspect. Security specialists have shown that it's possible to hack into autonomous cars, disabling the steering and brakes. Just one person could conceivably hijack an entire fleet of vehicles with a few keystrokes, creating a traffic jam or launching a terrorist attack.

AI's power as a tool of repression is even more frightening. Authoritarian governments could use deepfakes to discredit dissidents, facial recognition to enable round-the-clock mass surveillance, and predictive analytics to identify potential troublemakers. China has already started down the road toward digital authoritarianism. It has begun a massive repression campaign against the Muslim Uighur population in Xinjiang Province. Many of the tools the government is using there are low tech, but it has also begun to use data analytics, facial recognition systems, and predictive policing (the use of data to predict criminal activity). Vast networks of surveillance cameras are linked up to algorithms that can detect anomalous public behavior, from improperly parked vehicles to people running where they are not allowed. The Chinese company Yuntian Lifei Technology boasts that its intelligent video surveillance system has been deployed in nearly 80

Chinese cities and has identified some 6,000 incidents related to "social governance." Some of the ways in which Chinese authorities now use AI seem trivial, such as tracking how much toilet paper people use in public restrooms. Their proposed future uses are more sinister, such as monitoring patterns of electricity use for signs of suspicious activity.

China is not just building a techno-dystopian surveillance state at home; it has also begun exporting its technology. In 2018, Zimbabwe signed a deal with the Chinese company CloudWalk Technology to create a national database of faces and install facial recognition surveillance systems at airports, railway stations, and bus stops. There's more than money at stake in the deal. Zimbabwe has agreed to let CloudWalk send data on millions of faces back to China, helping the company improve its facial recognition systems for people with dark skin. China also plans to sell surveillance technology in Malaysia, Mongolia, and Singapore.

China is exporting its authoritarian laws and policies, too. According to Freedom House, China has held training sessions with government officials and members of the media from over 30 countries on methods to monitor and control public opinion. Three countries—Tanzania, Uganda, and Vietnam—passed restrictive media and cybersecurity laws soon after engaging with China.

WHAT AI WILL DO

Whichever country takes the lead on AI will use it to gain economic and military advantages over its competitors. By 2030, AI is projected to add between $13 trillion and $15 trillion to the global economy. AI could also accelerate the rate of scientific discovery. In 2019, an artificial neural network significantly outperformed existing approaches in synthetic protein folding, a key task in biological research.

AI is also set to revolutionize warfare. It will likely prove most useful in improving soldiers' situational awareness on the battlefield and commanders' ability to make decisions and communicate orders. AI systems can process more information than humans, and they can do it more quickly, making them valuable tools for assessing chaotic battles in real time. On the battlefield itself, machines can move faster and with greater precision and coordination than people. In the recent AI-versus-human StarCraft match, the AI system, AlphaStar, displayed superhuman abilities in rapidly processing large amounts of information, coordinating its units, and moving them quickly and precisely.

In the real world, these advantages will allow AI systems to manage swarms of robots far more effectively than humans could by controlling them manually. Humans will retain their advantages in higher-level strategy, but AI will dominate on the ground.

Washington's rush to develop AI is driven by a fear of falling behind China, which is already a global powerhouse in AI. The Chinese technology giants Alibaba, Baidu, and Tencent rank right alongside Amazon, Google, and Microsoft as leading AI companies. Five of the ten AI start-ups with the most funding last year were Chinese. Ten years ago, China's goal of becoming the global leader in AI by 2030 would have seemed fanciful; today, it's a real possibility.

Equally alarming for U.S. policymakers is the sharp divide between Washington and Silicon Valley over the military use of AI. Employees at Google and Microsoft have objected to their companies' contracts with the Pentagon, leading Google to discontinue work on a project using AI to analyze video footage. China's authoritarian regime doesn't permit this kind of open dissent. Its model of "military-civil fusion" means that Chinese technology innovations will translate more easily into military gains. Even if the United States keeps the lead in AI, it could lose its military advantage. The logical response to the threat of another country winning the AI race is to double down on one's own investments in AI. The problem is that AI technology poses risks not just to those who lose the race but also to those who win it.

THE ONLY WINNING MOVE IS NOT TO PLAY

Today's AI technologies are powerful but unreliable. Rules-based systems cannot deal with circumstances their programmers did not anticipate. Learning systems are limited by the data on which they were trained. AI failures have already led to tragedy. Advanced autopilot features in cars, although they perform well in some circumstances, have driven cars without warning into trucks, concrete barriers, and parked cars. In the wrong situation, AI systems go from supersmart to superdumb in an instant. When an enemy is trying to manipulate and hack an AI system, the risks are even greater.

Even when they don't break down completely, learning systems sometimes learn to achieve their goals in the wrong way. In a research paper last year, a group of 52 AI researchers recounted dozens of times when AI systems showed surprising behavior. An algorithm learning to walk in a simulated environment discovered it could move fastest

by repeatedly falling over. A Tetris-playing bot learned to pause the game before the last brick fell, so that it would never lose. One program deleted the files containing the answers against which it was being evaluated, causing it to be awarded a perfect score. As the researchers wrote, "It is often functionally simpler for evolution to exploit loopholes in the quantitative measure than it is to achieve the actual desired outcome." Surprise seems to be a standard feature of learning systems.

In the wrong situation, AI systems go from supersmart to superdumb in an instant.

Machine-learning systems are only ever as good as their training data. If the data don't represent the system's operating environment well, the system can fail in the real world. In 2018, for example, researchers at the MIT Media Lab showed that three leading facial recognition systems were far worse at classifying dark-skinned faces than they were at classifying light-skinned ones.

When they fail, machine-learning systems are also often frustratingly opaque. For rules-based systems, researchers can always explain the machine's behavior, even if they can't always predict it. For deep-learning systems, however, researchers are often unable to understand why a machine did what it did. Ali Rahimi, an AI researcher at Google, has argued that much like medieval alchemists, who discovered modern glassmaking techniques but did not understand the chemistry or physics behind their breakthroughs, modern machine-learning engineers can achieve powerful results but lack the underlying science to explain them.

Every failing of an AI system also presents a vulnerability that can be exploited. In some cases, attackers can poison the training data. In 2016, Microsoft created a chatbot called Tay and gave it a Twitter account. Other users began tweeting offensive messages at it, and within 24 hours, Tay had begun parroting their racist and anti-Semitic language. In that case, the source of the bad data was obvious. But not all data-poisoning attacks are so visible. Some can be buried within the training data in a way that is undetectable to humans but still manipulates the machine.

Even if the creators of a deep-learning system protect its data sources, the system can still be tricked using what are known as "adversarial examples," in which an attacker feeds the system an input that is carefully tailored to get the machine to make a mistake. A neural network classifying satellite images might be tricked into identifying a

subtly altered picture of a hospital as a military airfield or vice versa. The change in the image can be so small that the picture looks normal to a human but still fools the AI. Adversarial examples can even be placed in physical objects. In one case, researchers created a plastic turtle with subtle swirls embedded in the shell that made an object identification system think it was a rifle. In another, researchers placed a handful of small white and black squares on a stop sign, causing a neural network to classify it as a 45-mile-per-hour speed-limit sign. To make matters worse, attackers can develop these kinds of deceptive images and objects without access to the training data or the underlying algorithm of the system they are trying to defeat, and researchers have struggled to find effective defenses against the threat. Unlike with cybersecurity vulnerabilities, which can often be patched once they are uncovered, there is no known way of fully inoculating algorithms against these attacks.

Governments already have plenty of experience testing military, cyber-, and surveillance tools, but no testing method can guarantee that complex systems won't experience glitches once they're out in the real world. The first time F-22 fighter jets crossed the International Date Line, their computers crashed and the aircraft were nearly stranded over the Pacific Ocean.

Testing AI systems often takes even more time and money than testing traditional military hardware. Their complexity, which makes them more capable, also creates more opportunities for unexpected glitches. Imagine that a government develops an AI system that can hack into its adversaries' computer networks while avoiding detection. The first government to deploy such a system would gain a huge advantage over its competitors. Worried that an adversary was developing a similar tool, the government might feel compelled to cut testing short and deploy the system early. This dynamic has already played out in other industries, such as self-driving cars. The consequences of accidents caused by national security AI tools could be far worse.

A world of widespread, unprotected AI systems isn't just a possibility; it's the default setting.

AI wouldn't be the first case of governments relying on a powerful but unsafe technology. That's exactly what happened with computers, which play critical roles in everything from trading stocks to guiding

missiles even though they suffer from enormous vulnerabilities. In 2018, investigators at the U.S. Government Accountability Office found that U.S. weapons systems were riddled with cybersecurity loopholes that could be exploited with "relatively simple tools and techniques." Even worse, Defense Department program managers didn't know about the problems and dismissed the GAO's findings, claiming the tests were not realistic. Computer security vulnerabilities aren't limited to government-run systems. Company after company has suffered major data breaches. Digital security is already too often an afterthought. A world of widespread, unprotected AI systems isn't just a possibility; it's the default setting.

SAFETY FIRST

Urgent threats require urgent responses. One of the most important ways policymakers can deal with the dangers of AI is to boost funding for AI safety research. Companies are spending billions of dollars finding commercial applications for AI, but the U.S. government can play a valuable role in funding basic AI research, as it has since the field's early days. The AI Next initiative, a program run by the Defense Advanced Research Projects Agency that is set to spend $2 billion over the next five years, is aimed at tackling many of the limitations of narrow AI systems. Expanding on this effort, the White House should increase the funding going to AI safety research as part of its new American AI Initiative, and it should ask Congress for additional money for R & D and safety research.

When it comes to applying AI to national security, government agencies will have to reconsider their traditional approaches to testing new systems. Verifying that a system meets its design specifications isn't enough. Testers also need to ensure that it will continue to function properly in the real world when an adversary is trying to defeat it. In some cases, they can use computer simulations to tease out bugs, as manufacturers now do for autonomous cars. On top of that, the Departments of Defense and Homeland Security and the intelligence community should create red teams—groups that act as attackers to test a system's defenses—to ferret out vulnerabilities in AI systems so that developers can fix them before the systems go live.

Government officials should also tone down their rhetoric about an AI arms race, since such talk could easily become self-fulfilling. At a conference in 2018, Michael Griffin, the chief Pentagon official for

research and engineering, said, "There might be an artificial intelligence arms race, but we're not yet in it." Militaries are certainly going to adopt AI, but Griffin's statement was missing any concern for—or even awareness of—the risks that come with it. Talk of an arms race encourages adversaries to cut corners on safety. Government officials should emphasize not only the value of AI but also the importance of guaranteeing reliability and security.

Finally, the United States should look for ways to work with other countries, even hostile ones, to ensure AI safety. International cooperation on new technologies has a mixed record, but countries have sometimes succeeded in working together to avoid mutual harm. During the Cold War, the United States and the Soviet Union worked together to limit certain types of delivery systems for nuclear warheads that both sides agreed were particularly destabilizing. The United States also encouraged other countries to adopt safety measures to prevent the unauthorized use of nuclear weapons. Today, the United States should work with both allies and adversaries to boost international funding on AI safety. It should also begin discussions with China and Russia over whether some applications of AI pose unacceptable risks of escalation or loss of control and what countries can do jointly to improve safety. The biggest danger for the United States in an AI race is not losing but creating a world in which no one wins.

In the nineteenth century, industrialization brought tremendous economic growth, but it also handed militaries the tank, the machine gun, and mustard gas. The invention of nuclear weapons posed an even more profound risk, one with which policymakers are still grappling. Computers revolutionized how people work, learn, and communicate, but they also made previously isolated systems vulnerable to cyberattacks.

AI will match those changes. Most of its effects will be positive. It will boost economic growth, help diagnose and cure diseases, reduce automobile accidents, and improve people's daily lives in thousands of ways, large and small. Like any new technology, however, AI also has a darker side. Facing up to the risks now is the only way to make sure humanity realizes the promise of AI, not the peril.