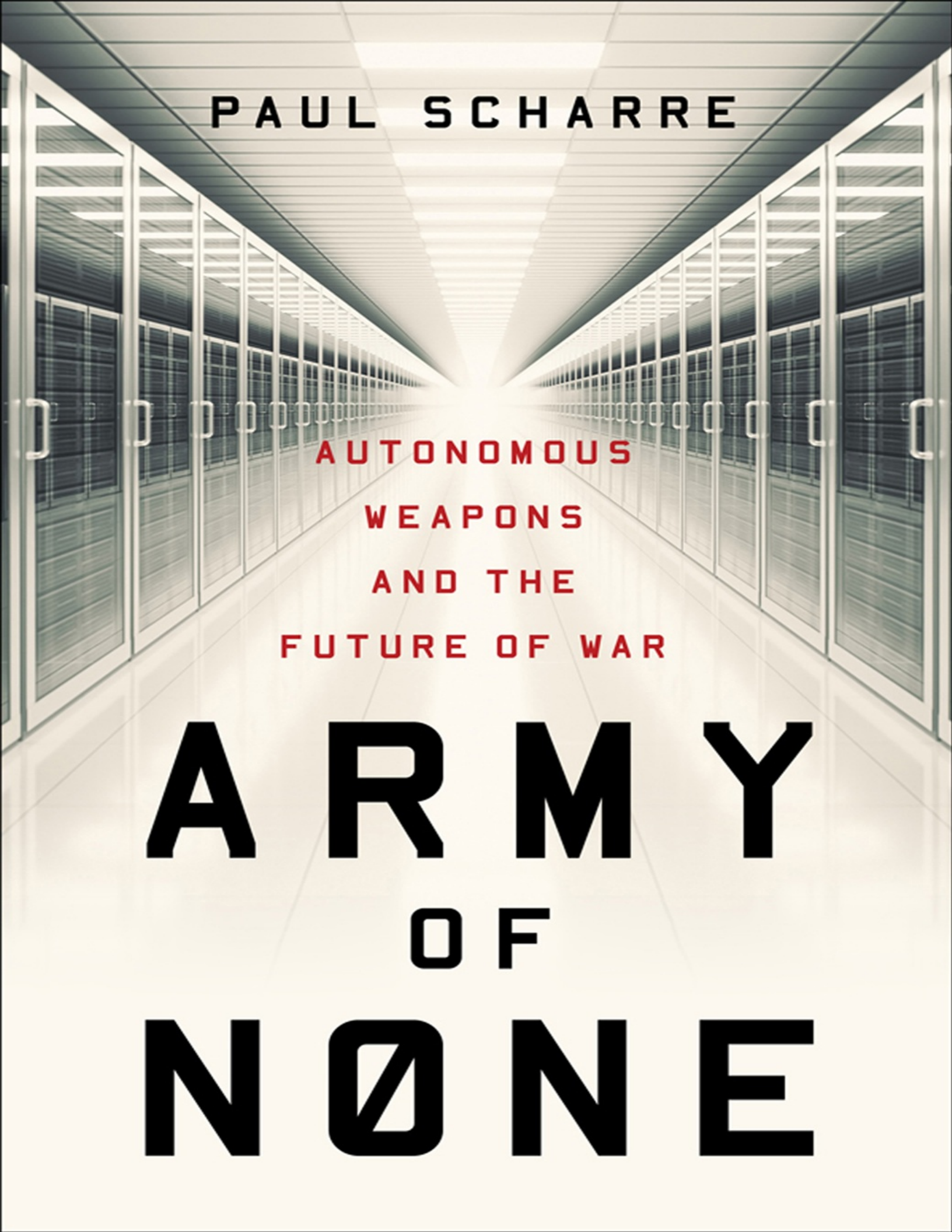


PAUL SCHARRE

AUTONOMOUS WEAPONS AND THE FUTURE OF WAR

ARMY OF NONE

ARMY OF NONE

Autonomous Weapons and
the Future of War

PAUL SCHARRE

W. W. NORTON & COMPANY

Independent Publishers Since 1923

New York | London

For Davey, William, and Ella,
that the world might be a better place.

And for Heather.
Thanks for everything.

Contents

ARMY OF NONE

Introduction

THE POWER OVER LIFE AND DEATH

THE MAN WHO SAVED THE WORLD

On the night of September 26, 1983, the world almost ended.

It was the height of the Cold War, and each side bristled with nuclear weapons. Earlier that spring, President Reagan had announced the Strategic Defense Initiative, nicknamed “Star Wars,” a planned missile defense shield that threatened to upend the Cold War’s delicate balance. Just three weeks earlier on September 1, the Soviet military had shot down a commercial airliner flying from Alaska to Seoul that had strayed into Soviet air space. Two hundred and sixty-nine people had been killed, including an American congressman. Fearing retaliation, the Soviet Union was on alert.

The Soviet Union deployed a satellite early warning system called Oko to watch for U.S. missile launches. Just after midnight on September 26, the system issued a grave report: the United States had launched a nuclear missile at the Soviet Union.

Lieutenant Colonel Stanislav Petrov was on duty that night in bunker Serpukhov-15 outside Moscow, and it was his responsibility to report the missile launch up the chain of command to his superiors. In the bunker, sirens blared and a giant red backlit screen flashed “launch,” warning him of the detected missile, but still Petrov was uncertain. Oko was new, and he worried that the launch might be an error, a bug in the system. He waited.

Another launch. Two missiles were inbound. Then another. And another. And another—five altogether. The screen flashing “launch” switched to “missile

strike." The system reported the highest confidence level. There was no ambiguity: a nuclear strike was on its way. Soviet military command would have only minutes to decide what to do before the missiles would explode over Moscow.

Petrov had a funny feeling. Why would the United States launch only five missiles? It didn't make sense. A real surprise attack would be massive, an overwhelming strike to wipe out Soviet missiles on the ground. Petrov wasn't convinced the attack was real. But he wasn't certain it was a false alarm, either.

With one eye on the computer readouts, Petrov called the ground-based radar operators for confirmation. If the missiles were real, they would show up on Soviet ground-based radars as they arced over the horizon. Puzzlingly, the ground radars detected nothing.

Petrov put the odds of the strike being real at 50/50, no easier to predict than a coin flip. He needed more information. He needed more time. All he had to do was pick up the phone, but the possible consequences were enormous. If he told Soviet command to fire nuclear missiles, millions would die. It could be the start of World War III.

Petrov went with his gut and called his superiors to inform them the system was malfunctioning. He was right: there was no attack. Sunlight reflecting off cloud tops had triggered a false alarm in Soviet satellites. The system was wrong. Humanity was saved from potential Armageddon by a human "in the loop."

What would a machine have done in Petrov's place? The answer is clear: the machine would have done whatever it was programmed to do, without ever understanding the consequences of its actions.

THE SNIPER'S CHOICE

In the spring of 2004—two decades later, in a different country, in a different war—I stared down the scope of my sniper rifle atop a mountain in Afghanistan. My sniper team had been sent to the Afghanistan-Pakistan border to scout infiltration routes where Taliban fighters were suspected of crossing back into Afghanistan. We hiked up the mountain all night, our 120-pound packs weighing heavily on the jagged and broken terrain. As the sky in the east began to lighten, we tucked ourselves in behind a rock outcropping—the best cover we could find. We hoped our position would conceal us at daybreak.

It didn't. A farmer spied our heads bobbing above the shallow rock

outcropping as the village beneath us woke to start their day. We'd been spotted.

Of course, that didn't change the mission. We kept watch, tallying the movement we could see up and down the road in the valley below. And we waited.

It wasn't long before we had company.

A young girl of maybe five or six headed out of the village and up our way, two goats in trail. Ostensibly she was just herding goats, but she walked a long slow loop around us, frequently glancing in our direction. It wasn't a very convincing ruse. She was spotting for Taliban fighters. We later realized that the chirping sound we'd heard as she circled us, which we took to be her whistling to her goats, was the chirp of a radio she was carrying. She slowly circled us, all the while reporting on our position. We watched her. She watched us.

She left, and the Taliban fighters came soon after.

We got the drop on them—we spotted them moving up a draw in the mountainside that they thought hid them from our position. The crackle of gunfire from the ensuing firefight brought the entire village out of their homes. It echoed across the valley floor and back, alerting everyone within a dozen miles to our presence. The Taliban who'd tried to sneak up on us had either run or were dead, but they would return in larger numbers. The crowd of villagers swelled below our position, and they didn't look friendly. If they decided to mob us, we wouldn't have been able to hold them all off.

"Scharre," my squad leader said. "Call for exfil."

I hopped on the radio. "This is Mike-One-Two-Romeo," I alerted our quick reaction force, "the village is massing on our position. We're going to need an exfil." Today's mission was over. We would regroup and move to a new, better position under cover of darkness that night.

Back in the shelter of the safe house, we discussed what we would do differently if faced with that situation again. Here's the thing: the laws of war don't set an age for combatants. Behavior determines whether or not a person is a combatant. If a person is participating in hostilities, as the young girl was doing by spotting for the enemy, then they are a lawful target for engagement. Killing a civilian who had stumbled across our position would have been a war crime, but it would have been legal to kill the girl.

Of course, it would have been wrong. Morally, if not legally.

In our discussion, no one needed to recite the laws of war or refer to abstract ethical principles. No one needed to appeal to empathy. The horrifying notion of shooting a child in that situation didn't even come up. We all knew it would have been wrong without needing to say it. War does force awful and difficult

choices on soldiers, but this wasn't one of them.

Context is everything. What would a machine have done in our place? If it had been programmed to kill lawful enemy combatants, it would have attacked the little girl. Would a robot know when it is lawful to kill, but wrong?

THE DECISION

Life-and-death choices in war are not to be taken lightly, whether the stakes are millions of lives or the fate of a single child. Laws of war and rules of engagement frame the decisions soldiers face amid the confusion of combat, but sound judgment is often required to discern the right choice in any given situation.

Technology has brought us to a crucial threshold in humanity's relationship with war. In future wars, machines may make life-and-death engagement decisions all on their own. Militaries around the globe are racing to deploy robots at sea, on the ground, and in the air—more than ninety countries have drones patrolling the skies. These robots are increasingly autonomous and many are armed. They operate under human control for now, but what happens when a Predator drone has as much autonomy as a Google car? What authority should we give machines over the ultimate decision—life or death?

This is not science fiction. More than thirty nations already have defensive supervised autonomous weapons for situations in which the speed of engagements is too fast for humans to respond. These systems, used to defend ships and bases against saturation attacks from rockets and missiles, are supervised by humans who can intervene if necessary—but other weapons, like the Israeli Harpy drone, have already crossed the line to full autonomy. Unlike the Predator drone, which is controlled by a human, the Harpy can search a wide area for enemy radars and, once it finds one, destroy it without asking permission. It's been sold to a handful of countries and China has reverse engineered its own variant. Wider proliferation is a definite possibility, and the Harpy may only be the beginning. South Korea has deployed a robotic sentry gun to the demilitarized zone bordering North Korea. Israel has used armed ground robots to patrol its Gaza border. Russia is building a suite of armed ground robots for war on the plains of Europe. Sixteen nations already have armed drones, and another dozen or more are openly pursuing development.

These developments are part of a deeper technology trend: the rise of artificial intelligence (AI), which some have called the "next industrial revolution." Technology guru Kevin Kelly has compared AI to electricity: just as

electricity brings objects all around us to life with power, so too will AI bring them to life with intelligence. AI enables more sophisticated and autonomous robots, from warehouse robots to next-generation drones, and can help process large amounts of data and make decisions to power Twitter bots, program subway repair schedules, and even make medical diagnoses. In war, AI systems can help humans make decisions—or they can be delegated authority to make decisions on their own.

The rise of artificial intelligence will transform warfare. In the early twentieth century, militaries harnessed the industrial revolution to bring tanks, aircraft, and machine guns to war, unleashing destruction on an unprecedented scale. Mechanization enabled the creation of machines that were physically stronger and faster than humans, at least for certain tasks. Similarly, the AI revolution is enabling the *cognitization* of machines, creating machines that are smarter and faster than humans for narrow tasks. Many military applications of AI are uncontroversial—improved logistics, cyberdefenses, and robots for medical evacuation, resupply, or surveillance—however, the introduction of AI into weapons raises challenging questions. Automation is already used for a variety of functions in weapons today, but in most cases it is still humans choosing the targets and pulling the trigger. Whether that will continue is unclear. Most countries have kept silent on their plans, but a few have signaled their intention to move full speed ahead on autonomy. Senior Russian military commanders envision that in the near future a "fully robotized unit will be created, capable of independently conducting military operations," while the U.S. Department of Defense officials state that the option of deploying fully autonomous weapons should be "on the table."

BETTER THAN HUMAN?

Armed robots deciding who to kill might sound like a dystopian nightmare, but some argue autonomous weapons could make war more humane. The same kind of automation that allows self-driving cars to avoid pedestrians could also be used to avoid civilian casualties in war, and unlike human soldiers, machines never get angry or seek revenge. They never fatigue or tire. Airplane autopilots have dramatically improved safety for commercial airliners, saving countless lives. Could autonomy do the same for war?

New types of AI like deep learning neural networks have shown startling advances in visual object recognition, facial recognition, and sensing human

emotions. It isn't hard to imagine future weapons that could outperform humans in discriminating between a person holding a rifle and one holding a rake. Yet computers still fall far short of humans in understanding context and interpreting meaning. AI programs today can identify objects in images, but can't draw these individual threads together to understand the big picture.

Some decisions in war are straightforward. Sometimes the enemy is easily identified and the shot is clear. Some decisions, however, like the one Stanislav Petrov faced, require understanding the broader context. Some situations, like the one my sniper team encountered, require moral judgment. Sometimes doing the right thing entails breaking the rules—what's legal and what's right aren't always the same.

THE DEBATE

Humanity faces a fundamental question: should machines be allowed to make life-and-death decisions in war? Should it be legal? Is it right?

I've been inside the debate on lethal autonomy since 2008. As a civilian policy analyst in the Pentagon's Office of the Secretary of Defense, I led the group that drafted the official U.S. policy on autonomy in weapons. (Spoiler alert: it doesn't ban them.) Since 2014, I've ran the Ethical Autonomy Project at the Center for a New American Security, an independent bipartisan think tank in Washington, DC, during which I've met experts from a wide range of disciplines grappling with these questions: academics, lawyers, ethicists, psychologists, arms control activists, military professionals, and pacifists. I've peered behind the curtain of government projects and met with the engineers building the next generation of military robots.

This book will guide you on a journey through the rapidly evolving world of next-generation robotic weapons. I'll take you inside defense companies building intelligent missiles and research labs doing cutting-edge work on swarming. I'll introduce the government officials setting policy and the activists striving for a ban. This book will examine the past—including things that went wrong—and look to the future, as I meet with the researchers pushing the boundaries of artificial intelligence.

This book will explore what a future populated by autonomous weapons might look like. Automated stock trading has led to "flash crashes" on Wall Street. Could autonomous weapons lead to a "flash war"? New AI methods such as deep learning are powerful, but often lead to systems that are effectively a "black box"—even to their designers. What new challenges will advanced AI

systems bring?

Over 3,000 robotics and artificial intelligence experts have called for a ban on offensive autonomous weapons, and are joined by over sixty nongovernmental organizations (NGOs) in the Campaign to Stop Killer Robots. Science and technology luminaries such as Stephen Hawking, Elon Musk, and Apple cofounder Steve Wozniak have spoken out against autonomous weapons, warning they could spark a "global AI arms race."

Can an arms race be prevented, or is one already under way? If it's already happening, can it be stopped? Humanity's track record for controlling dangerous technology is mixed; attempts to ban weapons that were seen as too dangerous or inhumane date back to antiquity. Many of these attempts have failed, including early-twentieth-century attempts to ban submarines and airplanes. Even those that have succeeded, such as the ban on chemical weapons, rarely stop rogue regimes such as Bashar al-Assad's Syria or Saddam Hussein's Iraq. If an international ban cannot stop the world's most odious regimes from building killer robot armies, we may someday face our darkest nightmares brought to life.

STUMBLING TOWARD THE ROBOPOCALYPSE

No nation has stated outright that they are building autonomous weapons, but in secret defense labs and dual-use commercial applications, AI technology is racing forward. For most applications, even armed robots, humans would remain in control of lethal decisions—but battlefield pressures could drive militaries to build autonomous weapons that take the human out of the loop. Militaries could desire greater autonomy to take advantage of computers' superior speed or so that robots can continue engagements when their communications to human controllers are jammed. Or militaries might build autonomous weapons simply because of a fear that others might do so. U.S. Deputy Secretary of Defense Bob Work has asked:

> If our competitors go to Terminators . . . and it turns out the Terminators are able to make decisions faster, even if they're bad, how would we respond?

Vice Chairman of the Joint Chiefs of Staff General Paul Selva has termed this dilemma "The Terminator Conundrum." The stakes are high: AI is emerging as a powerful technology. Used the right way, intelligent machines could save lives by making war more precise and humane. Used the wrong way, autonomous weapons could lead to more killing and even greater civilian casualties. Nations will not make these choices in a vacuum. It will depend on

what other countries do, as well as on the collective choices of scientists, engineers, lawyers, human rights activists, and others participating in this debate. Artificial intelligence is coming and it *will* be used in war. *How* it is used, however, is an open question. In the words of John Connor, hero of the *Terminator* movies and leader of the human resistance against the machines, "The future's not set. There's no fate but what we make for ourselves." The fight to ban autonomous weapons cuts to the core of humanity's ages-old conflicted relationship with technology: do we control our creations or do they control us?

PART I

Robopocalypse Now

1

THE COMING SWARM

THE MILITARY ROBOTICS REVOLUTION

On a sunny afternoon in the hills of central California, a swarm takes flight. One by one, a launcher flings the slim Styrofoam-winged drones into the air. The drones let off a high-pitched buzz, which fades as they climb into the crystal blue California sky.

The drones carve the air with sharp, precise movements. I look at the drone pilot standing next to me and realize with some surprise that his hands aren't touching the controls; the drones are flying fully autonomously. It's a silly realization—after all, autonomous drone swarms are what I've come here to see—yet somehow the experience of watching the drones fly with such agility without any human controlling them is different than I'd imagined. Their nimble movements seem purposeful, and it's hard not to imbue them with intention. It's both impressive and discomfiting, this idea of the drones operating "off leash."

I've traveled to Camp Roberts, California, to see researchers from the Naval Postgraduate School investigate something no one else in the world has ever done before: swarm warfare. Unlike Predator drones, which are individually remotely piloted by human controllers on the ground, these researchers' drones are controlled en masse. Today's experiment will feature twenty drones flying simultaneously in a ten-against-ten swarm-on-swarm mock dogfight. The shooting is simulated, but the maneuvering and flying are all real.

Each drone comes off the launcher with its autopilot already switched on. Without any human direction, they climb to their assigned altitudes and form two teams, reporting back when they are "swarm ready." The Red and Blue

swarms wait in their respective corners of the aerial combat arena, circling like a flock of hungry buzzards.

The pilot commanding Red Swarm rubs his hands together, anticipating the coming battle—which is funny, because his entire role is just to click the button that tells the swarm to start. After that, he's as much of a spectator as I am.

Duane Davis, the retired Navy helicopter pilot turned computer programmer who designed the swarm algorithms, counts down to the fight:

"Initiating swarm v. swarm . . . 3, 2, 1, shoot!"

Both the Red and Blue swarm commanders put their swarms into action. The two swarms close in on each other without hesitation. "Fight's on!" Duane yells enthusiastically. Within seconds, the swarms close the gap and collide. The two swarms blend together into a furball of close air combat. The swarms maneuver and swirl as a single mass. Simulated shots are tallied up at the bottom of the computer screen:

"UAV 74 fired at UAV 33
"UAV 59 fired at UAV 25
"UAV 33 hit
"UAV 25 hit . . ."

The swarms' behavior is driven by a simple algorithm called Greedy Shooter. Each drone will maneuver to get into a kill shot position against an enemy drone. A human must only choose the swarm behavior—wait, follow, attack, or land—and tell the swarm to start. After that, all of the swarm's actions are totally autonomous.

On the Red Swarm commander's computer screen, it's hard to tell who's winning. The drone icons overlap one another in a blur while, outside, the drones circle each other in a maelstrom of air combat. The whirling gyre looks like pure chaos to me, although Davis tells me he sometimes can pick out which drones are chasing each other.

A referee software called The Arbiter tracks the score. Red Swarm gains the upper hand with four kills to Blue's two. The "killed" drones' status switches from green to red as they're taken out of the fight. Then the fight falls into a lull, with the aircraft circling each other, unable to get a kill. Davis explains that because the aircraft are perfectly matched—same airframe, same flight controls, same algorithms—they sometimes fall into a stalemate where neither side can gain the upper hand.

Davis resets the battlefield for Round 2 and the swarms return to their

respective corners. When the swarm commanders click go, the swarms close on each other once again. This time the battle comes out dead even, 3–3. In Round 3, Red pulls out a decisive win, 7–4. Red Swarm commander is happy to take credit for the win. “I pushed the button,” he says with a chuckle.

Just as robots are transforming industries—from self-driving cars to robot vacuum cleaners and caretakers for the elderly—they are also transforming war. Global spending on military robotics is estimated to reach $7.5 billion per year in 2018, with scores of countries expanding their arsenals of air, ground, and maritime robots.

Robots have many battlefield advantages over traditional human-inhabited vehicles. Unshackled from the physiological limits of humans, uninhabited (“unmanned”) vehicles can be made smaller, lighter, faster, and more maneuverable. They can stay out on the battlefield far beyond the limits of human endurance, for weeks, months, or even years at a time without rest. They can take more risk, opening up tactical opportunities for dangerous or even suicidal missions without risking human lives.

However, robots have one major disadvantage. By removing the human from the vehicle, they lose the most advanced cognitive processor on the planet: the human brain. Most military robots today are remotely controlled, or teleoperated, by humans; they depend on fragile communication links that can be jammed or disrupted by environmental conditions. Without these communications, robots can only perform simple tasks, and their capacity for autonomous operation is limited.

The solution: more autonomy.

THE ACCIDENTAL REVOLUTION

No one planned on a robotics revolution, but the U.S. military stumbled into one as it deployed thousands of air and ground robots to meet urgent needs in Iraq and Afghanistan. By 2005, the U.S. Department of Defense (DoD) had woken up to the fact that something significant was happening. Spending on uninhabited aircraft, or drones, which had hovered around the $300 million per year mark in the 1990s, skyrocketed after 9/11, increasing sixfold to over $2 billion per year by 2005. Drones proved particularly valuable in the messy counterinsurgency wars in Iraq and Afghanistan. Larger aircraft like the MQ-1B Predator can quietly surveil terrorists around the clock, tracking their movements and unraveling their networks. Smaller hand-launched drones like the RQ-11

Raven can provide troops “over-the-hill reconnaissance” on demand while on patrol. Hundreds of drones had been deployed to Iraq and Afghanistan in short order.

Drones weren’t new—they had been used in a limited fashion in Vietnam—but the overwhelming crush of demand for them was. While in later years drones would become associated with “drone strikes,” it is their capacity for persistent surveillance, not dropping bombs, that makes them unique and valuable to the military. They give commanders a low-cost, low-risk way to put eyes in the sky.

Soon, the Pentagon was pouring drones into the wars at a breakneck pace. By 2011, annual spending on drones had swelled to over $6 billion per year, over twenty times pre-9/11 levels. DoD had over 7,000 drones in its fleet. The vast majority of them were smaller hand-launched models, but large aircraft like the MQ-9 Reaper and RQ-4 Global Hawk were also valuable military assets.

At the same time, DoD was discovering that robots weren’t just valuable in the air. They were equally important, if not more so, on the ground. Driven in large part by the rise of improvised explosive devices (IEDs), DoD deployed over 6,000 ground robots to Iraq and Afghanistan. Small robots like the iRobot Packbot allowed troops to disable or destroy IEDs without putting themselves at risk. Bomb disposal is a great job for a robot.

THE MARCH TOWARD EVER-GREATER AUTONOMY

In 2005, after DoD started to come to grips with the robotics revolution and its implications for the future of conflict, it began publishing a series of “roadmaps” for future unmanned system investment. The first roadmap was focused on aircraft, but subsequent roadmaps in 2007, 2009, 2011, and 2013 included ground and maritime vehicles as well. While the lion’s share of dollars has gone toward uninhabited aircraft, ground, sea surface, and undersea vehicles have valuable roles to play as well.

These roadmaps did more than simply catalog the investments DoD was making. Each roadmap looked forward twenty-five years into the future, outlining technology needs and wants in order to help inform future investments by government and industry. They covered sensors, communications, power, weapons, propulsion, and other key enabling technologies. Across all the roadmaps, autonomy is a dominant theme.

The 2011 roadmap perhaps summarized the vision best:

> For unmanned systems to fully realize their potential, they must be able to achieve a highly

> autonomous state of behavior and be able to interact with their surroundings. This advancement will require an ability to understand and adapt to their environment, and an ability to collaborate with other autonomous systems.

Autonomy is the cognitive engine that power robots. Without autonomy, robots are only empty vessels, brainless husks that depend on human controllers for direction.

In Iraq and Afghanistan, the U.S. military operated in a relatively "permissive" electromagnetic environment where insurgents did not generally have the ability to jam communications with robot vehicles, but this will not always be the case in future conflicts. Major nation-state militaries will almost certainly have the ability to disrupt or deny communications networks, and the electromagnetic spectrum will be highly contested. The U.S. military has ways of communicating that are more resistant to jamming, but these methods are limited in range and bandwidth. Against a major military power, the type of drone operations the United States has conducted when going after terrorists—streaming high-definition, full-motion video back to stateside bases via satellites—will not be possible. In addition, some environments inherently make communications challenging, such as undersea, where radio wave propagation is hindered by water. In these situations, autonomy is a must if robotic systems are to be effective. As machine intelligence advances, militaries will be able to create ever more autonomous robots capable of carrying out more complex missions in more challenging environments independent from human control.

Even if communications links work perfectly, greater autonomy is also desirable because of the personnel costs of remotely controlling robots. Thousands of robots require thousands of people to control them, if each robot is remotely operated. Predator and Reaper drone operations require seven to ten pilots to staff one drone "orbit" of 24/7 continuous around-the-clock coverage over an area. Another twenty people per orbit are required to operate the sensors on the drone, and scores of intelligence analysts are needed to sift through the sensor data. In fact, because of these substantial personnel requirements, the U.S. Air Force has a strong resistance to calling these aircraft "unmanned." There may not be anyone on board the aircraft, but there are still humans controlling it and supporting it.

Because the pilot remains on the ground, uninhabited aircraft free surveillance operations from the limits of human endurance—but only the physical ones. Drones can stay aloft for days at a time, far longer than a human pilot could remain effective sitting in the cockpit, but remote operation doesn't change the *cognitive* requirements on human operators. Humans still have to

perform the same tasks, they just aren't physically on board the vehicle. The Air Force prefers the term "remotely piloted aircraft" because that's what today's drones are. Pilots still fly the aircraft via stick and rudder input, just remotely from the ground, sometimes even half a world away.

It's a cumbersome way to operate. Building tens of thousands of cheap robots is not a cost-effective strategy if they require even larger numbers of highly trained (and expensive) people to operate them.

Autonomy is the answer. The 2011 DoD roadmap stated:

> Autonomy reduces the human workload required to operate systems, enables the optimization of the human role in the system, and allows human decision making to focus on points where it is most needed. These benefits can further result in manpower efficiencies and cost savings as well as greater speed in decision-making.

Many of DoD's robotic roadmaps point toward the long-term goal of full autonomy. The 2005 roadmap looked toward "fully autonomous swarms." The 2011 roadmap articulated an evolution of four levels of autonomy from (1) human operated to (2) human delegated, (3) human supervised, and eventually (4) fully autonomous. The benefits of greater autonomy was the "single greatest theme" in a 2010 report from the Air Force Office of the Chief Scientist on future technology.

Although Predator and Reaper drones are still flown manually, albeit remotely from the ground, other aircraft such as Air Force Global Hawk and Army Gray Eagle drones have much more automation: pilots direct these aircraft where to go and the aircraft flies itself. Rather than being flown via a stick and rudder, the aircraft are directed via keyboard and mouse. The Army doesn't even refer to the people controlling its aircraft as "pilots"—it called them "operators." Even with this greater automation, however, these aircraft still require one human operator per aircraft for anything but the simplest missions.

Incrementally, engineers are adding to the set of tasks that uninhabited aircraft can perform on their own, moving step by step toward increasingly autonomous drones. In 2013, the U.S. Navy successfully landed its X-47B prototype drone on a carrier at sea, autonomously. The only human input was the order to land; the actual flying was done by software. In 2014, the Navy's Autonomous Aerial Cargo/Utility System (AACUS) helicopter autonomously scouted out an improvised landing area and executed a successful landing on its own. Then in 2015, the X-47B drone again made history by conducting the first autonomous aerial refueling, taking gas from another aircraft while in flight.

These are key milestones in building more fully combat-capable uninhabited aircraft. Just as autonomous cars will allow a vehicle to drive from point A to

point B without manual human control, the ability to takeoff, land, navigate, and refuel autonomously will allow robots to perform tasks under human direction and supervision, but without humans controlling each movement. This can begin to break the paradigm of humans manually controlling the robot, shifting humans into a supervisory role. Humans will command the robot what action to take, and it will execute the task on its own.

Swarming, or cooperative autonomy, is the next step in this evolution. Davis is most excited about the nonmilitary applications of swarming, from search and rescue to agriculture. Coordinated robot behavior could be useful for a wide variety of applications and the Naval Postgraduate School's research is very basic, so the algorithms they're building could be used for many purposes. Still, the military advantages in mass, coordination, and speed are profound and hard to ignore. Swarming can allow militaries to field large numbers of assets on the battlefield with a small number of human controllers. Cooperative behavior can also allow quicker reaction times, so that the swarm can respond to changing events faster than would be possible with one person controlling each vehicle.

In conducting their swarm dogfight experiment, Davis and his colleagues are pushing the boundaries of autonomy. Their next goal is to work up to a hundred drones fighting in a fifty-on-fifty aerial swarm battle, something Davis and his colleagues are already simulating on computers, and their ultimate goal is to move beyond dogfighting to a more complex game akin to capture the flag. Two swarms would compete to score the most points by landing at the other's air base without being "shot down" first. Each swarm must balance defending its own base, shooting down enemy drones, and getting as many of its drones as possible into the enemy's base. What are the "plays" to run with a swarm? What are the best tactics? These are precisely the questions Davis and his colleagues want to explore.

"If I have fifty planes that are involved in a swarm," he said, "how much of that swarm do I want to be focused on offense—getting to the other guy's landing area? How much do I want focused on defending my landing space and doing the air-to-air problem? How do I want to do assignments of tasks between the swarms? If I've got the adversary's UAVs [unmanned aerial vehicles] coming in, how do I want my swarm deciding which UAV is going to take which adversary to try to stop them from getting to our base?"

Swarm tactics are still at a very early stage. Currently, the human operator allocates a certain number of drones to a sub-swarm then tasks that sub-swarm with a mission, such as attempting to attack an enemy's base or attacking enemy aircraft. After that, the human is in a supervisory mode. Unless there is a safety

concern, the human controller won't intervene to take control of an aircraft. Even then, if an aircraft began to experience a malfunction, it wouldn't make sense to take control of it until it left the swarm's vicinity. Taking manual control of an aircraft in the middle of the swarm could actually instigate a midair collision. It would be very difficult for a human to predict and avoid a collision with all of the other drones swirling in the sky. If the drone is under the swarm's command, however, it will automatically adjust its flight to avoid a collision.

Right now, the swarm behaviors Davis is using are very basic. The human can command the swarm to fly in a formation, to land, or to attack enemy aircraft. The drones then sort themselves into position for landing or formation flying to "deconflict" their actions. For some tasks, such as landing, this is done relatively easily by altitude: lower planes land first. Other tasks, such as deconflicting air-to-air combat are trickier. It wouldn't do any good, for example, for all of the drones in the swarm to go after the same enemy aircraft. They need to coordinate their behavior.

The problem is analogous to that of outfielders calling a fly ball. It wouldn't make sense to have the manager calling who should catch the ball from the dugout. The outfielders need to coordinate among themselves. "It's one thing when you've got two humans that can talk to one another and one ball," Davis explained. "It's another thing when there's fifty humans and fifty balls." This task would be effectively impossible for humans, but a swarm can accomplish this very quickly, through a variety of methods. In centralized coordination, for example, individual swarm elements pass their data back to a single controller, which then issues commands to each robot in the swarm. Hierarchical coordination, on the other hand, decomposes the swarm into teams and squads much like a military organization, with orders flowing down the chain of command.

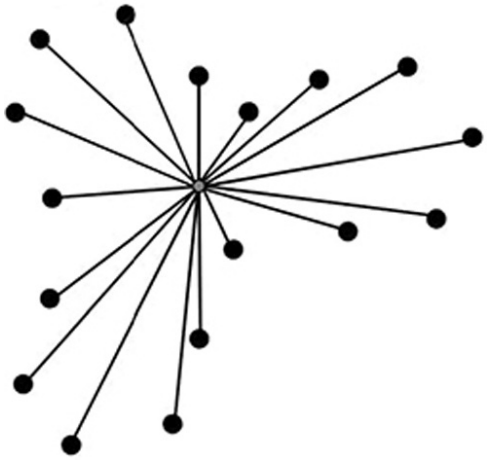
Consensus-based coordination is a decentralized approach where all of the swarm elements communicate with one another simultaneously and collectively decide on a course of action. They could do this by using "voting" or "auction" algorithms to coordinate behavior. For example, each swarm element could place a "bid" on an "auction" to catch the fly ball. The individual that bids highest "wins" the auction and catches the ball, while the others move out of the way.

Emergent coordination is the most decentralized approach and is how flocks of birds, colonies of insects, and mobs of people work, with coordinated action arising naturally from each individual making decisions based on those nearby. Simple rules for individual behavior can lead to very complex collective action,

allowing the swarm to exhibit “collective intelligence.” For example, a colony of ants will converge on an optimal route to take food back to the nest over time because of simple behavior from each individual ant. As ants pick up food, they leave a pheromone trail behind them as they move back to the nest. If they come across an existing trail with stronger pheromones, they’ll switch to it. More ants will arrive back at the nest sooner via the faster route, leading to a stronger pheromone trail, which will then cause more ants to use that trail. No individual ant “knows” which trail is fastest, but collectively the colony converges on the fastest route.

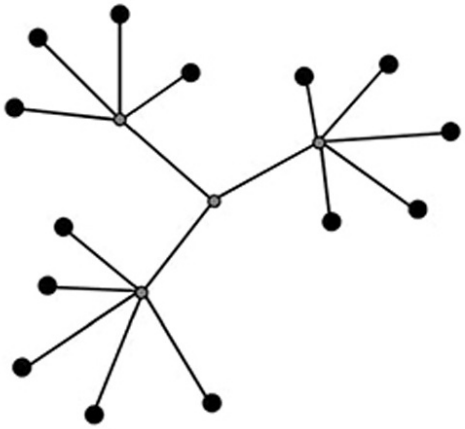
Centralized Coordination

Swarm elements communicate with a centralized planner which coordinates all tasks.



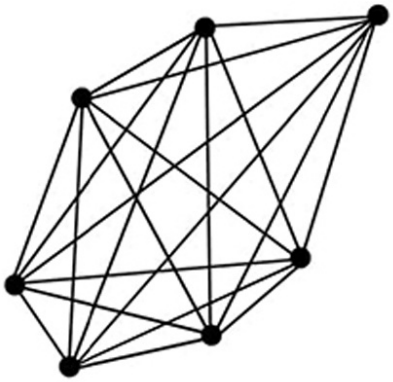
Hierarchical Coordination

Swarm elements are controlled by "squad" level agents, who are in turn controlled by higher-level controllers.



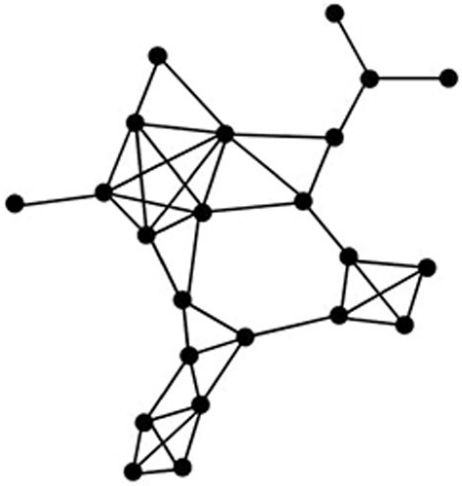
Coordination by Consensus

All swarm elements communicate to one another and use "voting" or auction-based methods to converge on a solution.



Emergent Coordination

Coordination arises naturally by individual swarm elements reacting to one another, like in animal swarms.



Swarm Command-and-Control Models

Communication among elements of the swarm can occur through direct signaling, akin to an outfielder yelling “I got it!”; indirect methods such as co-observation, which is how schools of fish and herds of animals stay together; or by modifying the environment in a process called *stigmergy*, like ants leaving pheromones to mark a trail.

The drones in Davis’s swarm communicate through a central Wi–Fi router on the ground. They avoid collisions by staying within narrow altitude windows that are automatically assigned by the central ground controller. Their attack behavior is uncoordinated, though. The “greedy shooter” algorithm simply directs each drone to attack the nearest enemy drone, regardless of what the other drones are doing. In theory, all the drones could converge on the same enemy drone, leaving other enemies untouched. It’s a terrible method for air-to-air combat, but Davis and his colleagues are still in the proof-of-concept stage. They have experimented with a more decentralized auction-based approach and found it to be very robust to disruptions, including up to a 90 percent communications loss within the swarm. As long as some communications are up, even if they’re spotty, the swarm will converge on a solution.

The effect of fifty aircraft working together, rather than fighting individually or in wingman pairs as humans do today, would be tremendous. Coordinated behavior is the difference between a basketball *team* and five ball hogs all making a run at the basket themselves. It’s the difference between a bunch of lone wolves and a *wolf pack.*

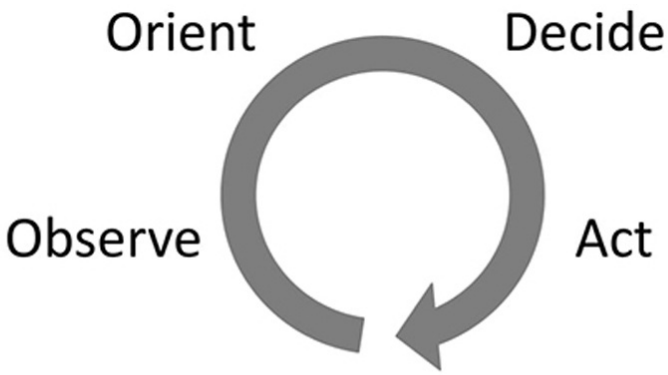
In 2016, the United States demonstrated 103 aerial drones flying together in a swarm that DoD officials described as “a collective organism, sharing one distributed brain for decision-making and adapting to each other like swarms in nature.” (Not to be outdone, a few months later China demonstrated a 119-drone swarm.) Fighting together, a drone swarm could be far more effective than the same number of drones fighting individually. No one yet knows what the best tactics will be for swarm combat, but experiments such as these are working to tease them out. New tactics might even be evolved by the machines themselves through machine learning or evolutionary approaches.

Swarms aren’t merely limited to the air. In August 2014, the U.S. Navy Office of Naval Research (ONR) demonstrated a swarm of small boats on the James River in Virginia by simulating a mock strait transit in which the boats protected a high-value Navy ship against possible threats, escorting it through a simulated high-danger area. When directed by a human controller to investigate a potential threat, a detachment of uninhabited boats moved to intercept and encircle the suspicious vessel. The human controller simply directed them to

intercept the designated suspicious ship; the boats moved autonomously, coordinating their actions by sharing information. This demonstration involved five boats working together, but the concept could be scaled up to larger numbers, just as in aerial drone swarms.

Bob Brizzolara, who directed the Navy's demonstration, called the swarming boats a "game changer." It's an often-overused term, but in this case, it's not hyperbole—robotic boat swarms are highly valuable to the Navy as a potential way to guard against threats to its ships. In October 2000, the USS *Cole* was attacked by al-Qaida terrorists using a small explosive-laden boat while in port in Aden, Yemen. The blast killed seventeen sailors and cut a massive gash in the ship's hull. Similar attacks continue to be a threat to U.S. ships, not just from terrorists but also from Iran, which regularly uses small high-speed craft to harass U.S. ships near the Straits of Hormuz. Robot boats could intercept suspicious vessels further away, putting eyes (and potentially weapons) on potentially hostile boats without putting sailors at risk.

What the robot boats might do after they've intercepted a potentially hostile vessel is another matter. In a video released by the ONR, a .50 caliber machine gun is prominently displayed on the front of one of the boats. The video's narrator makes no bones about the fact that the robot boats could be used to "damage or destroy hostile vessels," but the demonstration didn't involve firing any actual bullets, and didn't include a consideration of what the rules of engagement actually would have been. Would a human be required to pull the trigger? When pressed by reporters following the demonstration, a spokesman for ONR explained that "there is always a human in the loop when it comes to the actual engagement of an enemy." But the spokesman also acknowledged that "under this swarming demonstration with multiple [unmanned surface vehicles], ONR did not study the specifics of how the human-in-the-loop works for rules of engagement."



In the OODA loop paradigm of combat, victory on the battlefield goes to whichever side can complete the observe-orient-decide-act cycle faster.

OODA Loop

The Navy's fuzzy answer to such a fundamental question reflects a tension in the military's pursuit of more advanced robotics. Even as researchers and engineers move to incorporate more autonomy, there is an understanding that there are—or should be—limits on autonomy when it comes to the use of weapons. What exactly those limits are, however, is often unclear.

REACHING THE LIMIT

How much autonomy is too much? The U.S. Air Force laid out an ambitious vision for the future of robot aircraft in their *Unmanned Aircraft Systems Flight Plan, 2009–2047*. The report envisioned a future where an arms race in speed drove a desire for ever-faster automation, not unlike real-world competition in automated stock trading.

In air combat, pilots talk about an observe, orient, decide, act (OODA) loop, a cognitive process pilots go through when engaging enemy aircraft. Understanding the environment, deciding, and acting faster than the enemy allows a pilot to "get inside" the enemy's OODA loop. While the enemy is still trying to understand what's happening and decide on a course of action, the pilot has already changed the situation, resetting the enemy to square one and forcing him or her to come to grips with a new situation. Air Force strategist John Boyd, originator of the OODA loop, described the objective:

> Goal: Collapse adversary's system into confusion and disorder by causing him to over and under react to activity that appears simultaneously menacing as well as ambiguous, chaotic, or misleading.

If victory comes from completing this cognitive process faster, then one can see the advantage in automation. The Air Force's 2009 Flight Plan saw tremendous potential for computers to exceed human decision-making speeds:

> Advances in computing speeds and capacity will change how technology affects the OODA loop. Today the role of technology is changing from supporting to fully participating with humans in each step of the process. In 2047 technology will be able to reduce the time to complete the OODA loop to micro or nanoseconds. Much like a chess master can outperform proficient chess players, [unmanned aircraft systems] will be able to react at these speeds and therefore this loop moves toward becoming a "perceive and act" vector. Increasingly humans will no longer be "in the loop" but rather "on the loop"—monitoring the execution of certain decisions. Simultaneously, advances in AI will enable systems to make combat decisions and act within legal and policy constraints without necessarily requiring human input.

This, then, is the logical culmination of the arms race in speed: autonomous weapons that complete engagements all on their own. The Air Force Flight Plan acknowledged the gravity of what it was suggesting might be possible. The next paragraph continued:

> Authorizing a machine to make lethal combat decisions is contingent upon political and military leaders resolving legal and ethical questions. These include the appropriateness of machines having this ability, under what circumstances it should be employed, where responsibility for mistakes lies and what limitations should be placed upon the autonomy of such systems. . . . Ethical discussions and policy decisions must take place in the near term in order to guide the development of future [unmanned aircraft system] capabilities, rather than allowing the development to take its own path apart from this critical guidance.

The Air Force wasn't recommending autonomous weapons. It wasn't even suggesting they were necessarily a good idea. What it was suggesting was that autonomous systems might have advantages over humans in speed, and that AI might advance to the point where machines could carry out lethal targeting and engagement decisions without human input. If that is true, then legal, ethical, and policy discussions should take place now to shape the development of this technology.

At the time the Air Force Flight Plan was released in 2009, I was working in the Office of the Secretary of Defense as a civilian policy analyst focusing on drone policy. Most of the issues we were grappling with at the time had to do with how to manage the overwhelming demand for more drones from Iraq and Afghanistan. Commanders on the ground had a seemingly insatiable appetite for drones. Despite the thousands that had been deployed, they wanted more, and

Pentagon senior leaders—particularly in the Air Force—were concerned that spending on drones was crowding out other priorities. Secretary of Defense Robert Gates, who routinely chastised the Pentagon for its preoccupation with future wars over the ongoing ones in Iraq and Afghanistan, strongly sided with warfighters in the field. His guidance was clear: send more drones. Most of my time was spent figuring out how to force the Pentagon bureaucracy to comply with the secretary's direction and respond more effectively to warfighter needs, but when policy questions like this came up, eyes turned toward me.

I didn't have the answers they wanted. There was no policy on autonomy. Although the Air Force had asked for policy guidance in their 2009 Flight Plan, there wasn't even a conversation under way.

The 2011 DoD roadmap, which I was involved in writing, took a stab at an answer, even if it was a temporary one:

> Policy guidelines will especially be necessary for autonomous systems that involve the application of force. . . . For the foreseeable future, decisions over the use of force and the choice of which individual targets to engage with lethal force will be retained under human control in unmanned systems.

It didn't say much, but it was the first official DoD policy statement on lethal autonomy. Lethal force would remain under human control for the "foreseeable future." But in a world where AI technology is racing forward at a breakneck pace, how far into the future can we really see?

2

THE TERMINATOR AND THE ROOMBA

WHAT IS AUTONOMY?

Autonomy is a slippery word. For one person, “autonomous robot” might mean a household Roomba that vacuums your home while you’re away. For another, autonomous robots conjure images from science fiction. Autonomous robots could be a good thing, like the friendly—if irritating—C-3PO from *Star Wars*, or could lead to rogue homicidal agents, like those Skynet deploys against humanity in the *Terminator* movies.

Science fiction writers have long grappled with questions of autonomy in robots. Isaac Asimov created the now-iconic Three Laws of Robotics to govern robots in his stories:

1 A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2 A robot must obey orders given by human beings except where such orders would conflict with the first law.
3 A robot must protect its own existence as long as such protection does not conflict with the first or second law.

In Asimov’s stories, these laws embedded within the robot’s “positronic brain” are inviolable. The robot must obey. Asimov’s stories often explore the consequences of robots’ strict obedience of these laws, and loopholes in the laws themselves. In the Asimov-inspired movie *I, Robot* (spoiler alert), the lead robot protagonist, Sonny, is given a secret secondary processor that allows him to

override the Three Laws, if he desires. On the outside, Sonny looks the same as other robots, but the human characters can instantly tell there is something different about him. He dreams. He questions them. He engages in humanlike dialogue and critical thought of which the other robots are incapable. There is something unmistakably human about Sonny's behavior.

When Dr. Susan Calvin discovers the source of Sonny's apparent anomalous conduct, she finds it hidden in his chest cavity. The symbolism in the film is unmistakable: unlike other robots who are slaves to logic, Sonny has a "heart."

Fanciful as it may be, *I, Robot*'s take on autonomy resonates. Unlike machines, humans have the ability to ignore instructions and make decisions for themselves. Whether robots can ever have something akin to human free will is a common theme in science fiction. In *I, Robot*'s climactic scene, Sonny makes a choice to save Dr. Calvin, even though it means risking the success of their mission to defeat the evil AI V.I.K.I., who has taken over the city. It's a choice motivated by love, not logic. In the *Terminator* movies, when the military AI Skynet becomes self-aware, it makes a different choice. Upon determining that humans are a threat to its existence, Skynet decides to eliminate them, starting global nuclear war and initiating "Judgment Day."

THE THREE DIMENSIONS OF AUTONOMY

In the real world, machine autonomy doesn't require a magical spark of free will or a soul. Autonomy is simply the ability for a machine to perform a task or function on its own.

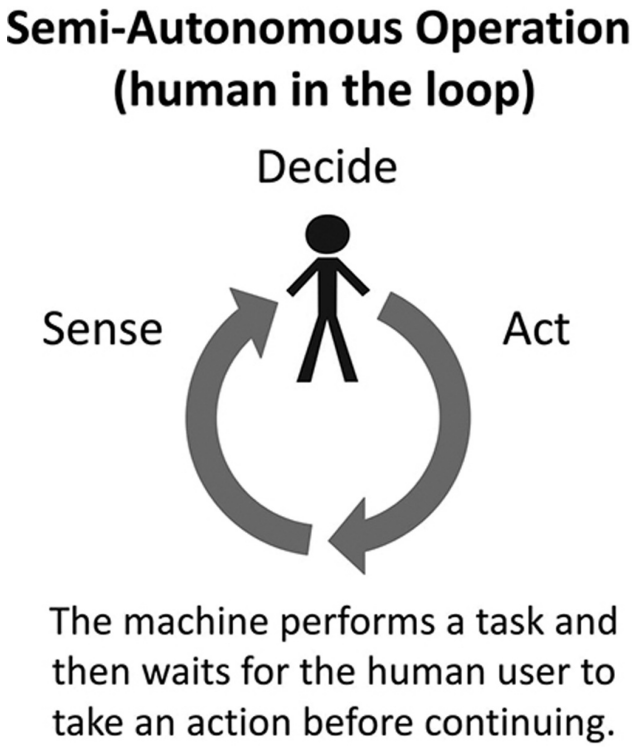
The DoD unmanned system roadmaps referred to "levels" or a "spectrum" of autonomy, but those classifications are overly simplistic. Autonomy encompasses three distinct concepts: the type of task the machine is performing; the relationship of the human to the machine when performing that task; and the sophistication of the machine's decision-making when performing the task. This means there are three different *dimensions* of autonomy. These dimensions are independent, and a machine can be "more autonomous" by increasing the amount of autonomy along any of these spectrums.

The first dimension of autonomy is the task being performed by the machine. Not all tasks are equal in their significance, complexity, and risk: a thermostat is an autonomous system in charge of regulating temperature, while *Terminator*'s Skynet was given control over nuclear weapons. The complexity of decisions involved and the consequences if the machine fails to perform the task

appropriately are very different. Often, a single machine will perform some tasks autonomously, while humans are in control of other tasks, blending human and machine control within the system. Modern automobiles have a range of autonomous features: automatic braking and collision avoidance, antilock brakes, automatic seat belt retractors, adaptive cruise control, automatic lane keeping, and self-parking. Some autonomous functions, such as autopilots in commercial airliners, can be turned on or off by a human user. Other autonomous functions, like airbags, are always ready and decide for themselves when to activate. Some autonomous systems may be designed to override the human user in certain situations. U.S. fighter aircraft have been modified with an automatic ground collision avoidance system (Auto-GCAS). If the pilot becomes disoriented and is about to crash, Auto-GCAS will take control of the aircraft at the last minute to pull up and avoid the ground. The system has already saved at least one aircraft in combat, rescuing a U.S. F-16 in Syria.

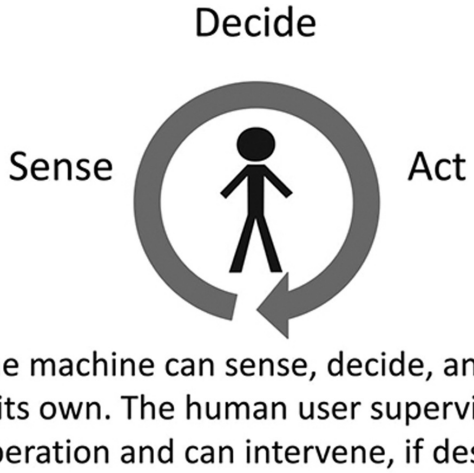
As automobiles and aircraft demonstrate, it is meaningless to refer to a system as “autonomous” without referring to the specific task that is being automated. Cars are still driven by humans (for now), but a host of autonomous functions can assist the driver, or even take control for short periods of time. The machine becomes “more autonomous” as it takes on more tasks, but some degree of human involvement and direction always exists. “Fully autonomous” self-driving cars can navigate and drive on their own, but a human is still choosing the destination.

For any given task, there are degrees of autonomy. A machine can perform a task in a semiautonomous, supervised autonomous, or fully autonomous manner. This is the second dimension of autonomy: the human-machine relationship.



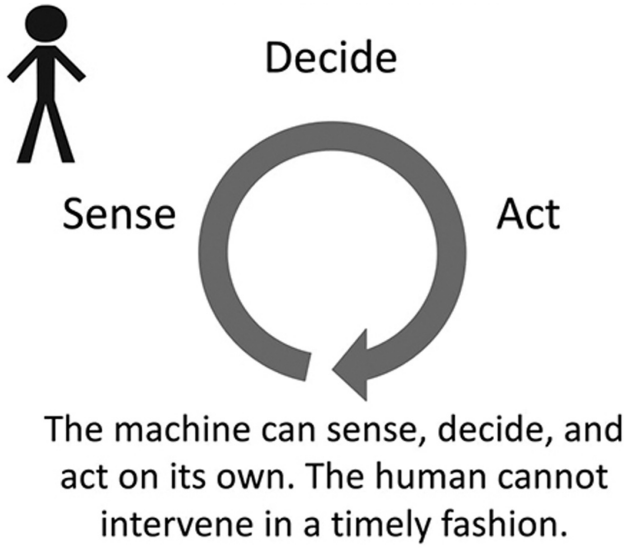
Semiautonomous Operation (human in the loop)

In semiautonomous systems, the machine performs a task and then waits for a human user to take an action before continuing. A human is "in the loop." Autonomous systems go through a sense, decide, act loop similar to the military OODA loop, but in semiautonomous systems the loop is broken by a human. The system can sense the environment and recommend a course of action, but cannot carry out the action without human approval.



Supervised Autonomous Operation (human on the loop)

In supervised autonomous systems, the human sits “on” the loop. Once put into operation, the machine can sense, decide, and act on its own, but a human user can observe the machine’s behavior and intervene to stop it, if desired.



Fully Autonomous Operation (human out of the loop)

Fully autonomous systems sense, decide, and act entirely without human intervention. Once the human activates the machine, it conducts the task without communication back to the human user. The human is “out of the loop.”

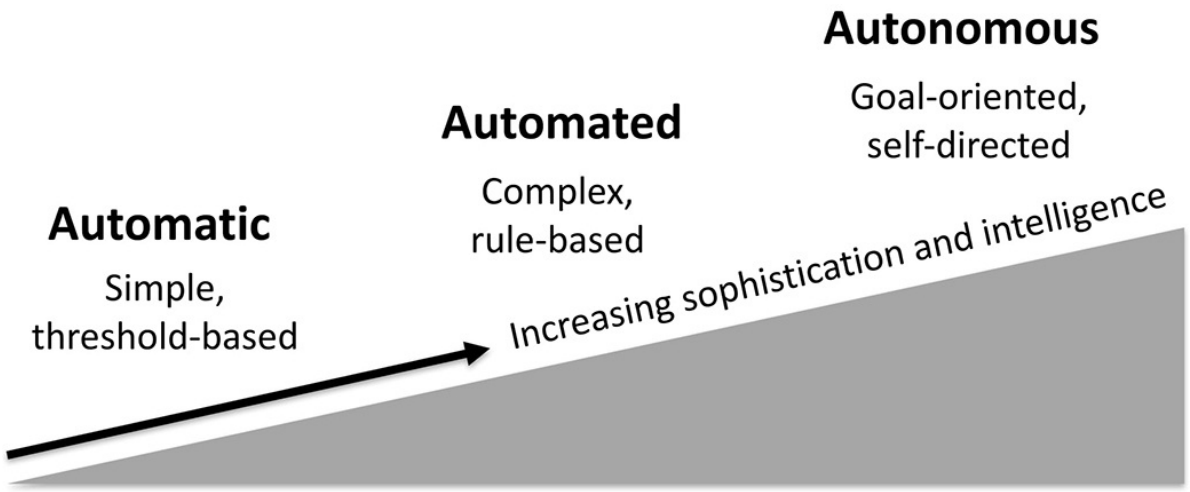
Many machines can operate in different modes at different times. A Roomba that is vacuuming while you are home is operating in a supervised autonomous mode. If the Roomba becomes stuck—my Roomba frequently trapped itself in the bathroom—then you can intervene. If you’re out of the house, then the Roomba is operating in a fully autonomous capacity. If something goes wrong, it’s on its own until you come home. More often than I would have liked, I came home to a dirty house and a spotless bathroom.

It wasn’t the Roomba’s fault it had locked itself in the bathroom. It didn’t even know that it was stuck (Roombas aren’t very smart). It had simply wandered into a location where its aimless bumping would nudge the door closed, trapping it. Intelligence is the third dimension of autonomy. More sophisticated, or more intelligent, machines can be used to take on more complex tasks in more challenging environments. People often use terms like “automatic,” “automated,” or “autonomous” to refer to a spectrum of intelligence in machines.

Automatic systems are simple machines that don’t exhibit much in the way of “decision-making.” They sense the environment and act. The relationship

between sensing and action is immediate and linear. It is also highly predictable to the human user. An old mechanical thermostat is an example of an automatic system. The user sets the desired temperature and when the temperature gets too high or too low, the thermostat activates the heat or air conditioning.

Automated systems are more complex, and may consider a range of inputs and weigh several variables before taking an action. Nevertheless, the internal cognitive processes of the machine are generally traceable by the human user, at least in principle. A modern digital programmable thermostat is an example of an automated system. Whether the heat or air conditioning turns on is a function of the house temperature as well as what day and time it is. Given knowledge of the inputs to the system and its programmed parameters, the system's behavior should be predictable to a trained user.



As machines become more sophisticated, they become more capable and able to accomplish more complex tasks in more open-ended environments. The downside is that their specific actions may be less predictable, even to trained users.

Spectrum of Intelligence in Machines

"Autonomous" is often used to refer to systems sophisticated enough that their internal cognitive processes are less intelligible to the user, who understands the task the system is supposed to perform, but not necessarily *how* the system will perform that task. Researchers often refer to autonomous systems as being "goal-oriented." That is to say, the human user specifies the goal, but the autonomous system has flexibility in how it achieves that goal.

Take a self-driving car, for example. The user specifies the destination and other goals, such as avoiding accidents, but can't possibly specify in advance

every single action the autonomous car is supposed to perform. The user doesn't know where there will be traffic or obstacles in the road, when lights will change, or what other cars or pedestrians will do. The car is therefore programmed with the flexibility to decide when to stop, go, and change lanes in order to accomplish its goal: getting to the destination safely.

In practice, the line between automatic, automated, and autonomous systems is still blurry. Often, the term "autonomous" is used to refer to future systems that have not yet been built, but once they do exist, people describe those same systems as "automated." This is similar to a trend in artificial intelligence where AI is often perceived to encompass only tasks that machines cannot yet do. Once a machine conquers a task, then it is merely "software."

Autonomy doesn't mean the system is exhibiting free will or disobeying its programming. The difference is that unlike an automatic system where there is a simple, linear connection from sensing to action, autonomous systems take into account a range of variables to consider the best action in any given situation. Goal-oriented behavior is essential for autonomous systems in uncontrolled environments. If a self-driving car were on a closed track with no pedestrians or other vehicles, each movement could be programmed into the car in advance—when to go, stop, turn, *etc.* But such a car would not be very useful, as it could only drive in a simple environment where every action could be predicted. In more complex environments or when performing more complex tasks, it is crucial that the machine be able to make decisions based on the specific situation.

This greater complexity in autonomous systems is a double-edged sword. The downside to more sophisticated systems is that the user may not be able to predict its specific actions in advance. The feature of increased autonomy can become a flaw if the user is surprised in an unpleasant way by the machine's behavior. For simple automatic or automated systems, this is less likely. But as the complexity of the system increases, so does the difficulty of predicting how the machine will act.

It can be exciting, if a little scary, to hand over control to an autonomous system. The machine is like a black box. We specify its goal and, like magic, the machine overcomes obstacles to reach the goal. The inner workings of how it did so are often mysterious to us; the distinction between "automated" and "autonomous" is principally in the mind of the user. A new machine only feels "autonomous" because we don't yet have a good mental model for how it "thinks." As we gain experience with the machine and begin to better understand it, the layers of fog hiding the inner workings of the black box dissipate,

revealing the complex logic driving its behavior. We come to decide the machine is merely "automated" after all. In understanding the machine, we have tamed it; the humans are back in control. That process of discovery, however, can be a rocky one.

A few years ago, I purchased a Nest "learning thermostat." The Nest tracks your behavior and adjusts the house's temperature as needed, "learning" your preferences over time. There were bumps along the way as I discovered various aspects of the Nest's functionality and occasionally the house was temporarily too warm or too cold, but I was sufficiently enamored of the technology that I was willing to push through these growing pains. My wife, Heather, was less tolerant of the Nest. Every time it changed the temperature on its own, disregarding an instruction she had given, she viewed it more and more suspiciously. (Unbeknownst to her, the Nest was following other guidance I had given it previously.)

The final straw for the Nest was when we came home from summer vacation to find the house a toasty 84 degrees, despite my having gone online the night before and set the Nest to a comfortable 70. With sweat dripping off our faces, we set our bags down in the foyer and I ran to the Nest to see what had happened. As it turned out, I had neglected to turn off the "auto-away feature." After the Nest's hallway sensor detected no movement and discerned we were not home, it reverted—per its programming—to the energy-saving "away" setting of 84 degrees. One look from Heather told me it was too late, though. She had lost trust in the Nest. (Or, more accurately, in my ability to use it.)

The Nest wasn't broken, though. The human-machine connection was. The same features that made the Nest "smarter" also made it harder for me to anticipate its behavior. The disconnect between my expectations of what the Nest would do and what it was actually doing meant the autonomy that was supposed to be working for me ended up, more often than not, working against my goals.

HOW MUCH SHOULD WE TRUST AUTONOMOUS SYSTEMS?

All the Nest did was control the thermostat. The Roomba merely vacuumed. Coming home to a Roomba locked in the bathroom or an overheated house might be annoying, but it wasn't a catastrophe. The tasks entrusted to these autonomous systems weren't critical ones.

What if I was dealing with an autonomous system performing a truly critical function? What if the Nest was a weapon, and my inability to understand it led to failure?

What if the task I was delegating to an autonomous system was the decision whether or not to kill?

3

MACHINES THAT KILL

WHAT IS AN AUTONOMOUS WEAPON?

The path to autonomous weapons began 150 years ago in the mid-nineteenth century. As the second industrial revolution was bringing unprecedented productivity to cities and factories, the same technology was bringing unprecedented efficiency to killing in war.

At the start of the American Civil War in 1861, inventor Richard Gatling devised a new weapon to speed up the process of firing: the Gatling gun. A forerunner of the modern machine gun, the Gatling gun employed automation for loading and firing, allowing more bullets to be fired in a shorter amount of time. The Gatling gun was a significant improvement over Civil War–era rifled muskets, which had to be loaded by hand through the muzzle in a lengthy process. Well-trained troops could fire three rounds per minute with a rifled musket. The Gatling gun fired over 300 rounds per minute.

In its time, the Gatling gun was a marvel. Mark Twain was an early enthusiast:

> [T]he Gatling gun . . . is a cluster of six to ten savage tubes that carry great conical pellets of lead, with unerring accuracy, a distance of two and a half miles. It feeds itself with cartridges, and you work it with a crank like a hand organ; you can fire it faster than four men can count. When fired rapidly, the reports blend together like the clattering of a watchman's rattle. It can be discharged four hundred times a minute! I liked it very much.

The Gatling gun was not an autonomous weapon, but it began a long evolution of weapons automation. In the Gatling gun, the process of loading bullets, firing, and ejecting cartridges was all automatic, provided a human kept

turning the crank. The result was a tremendous expansion in the amount of destructive power unleashed on the battlefield. Four soldiers were needed to operate the Gatling gun, but by dint of automation, they could deliver the same lethal firepower as more than a hundred men.

Richard Gatling's motivation was not to accelerate the process of killing, but to save lives by reducing the number of soldiers needed on the battlefield. Gatling built his device after watching waves of young men return home wounded or dead from the unrelenting bloodshed of the American Civil War. In a letter to a friend, he wrote:

> It occurred to me that if I could invent a machine—a gun—which could by its rapidity of fire, enable one man to do as much battle duty as a hundred, that it would, to a great extent, supersede the necessity of large armies, and consequently, exposure to battle and disease be greatly diminished.

Gatling was an accomplished inventor with multiple patents to his name for agricultural implements. He saw the gun in a similar light—machine technology harnessed to improve efficiency. Gatling claimed his gun "bears the same relation to other firearms that McCormack's reaper does to the sickle, or the sewing machine to the common needle."

Gatling was more right than he knew. The Gatling gun did indeed lay the seeds for a revolution in warfare, a break from the old ways of killing people one at a time with rifled muskets and shift to a new era of mechanized death. The future Gatling wrought was not one of less bloodshed, however, but unimaginably more. The Gatling gun laid the foundations for a new class of machine: the automatic weapon.

AUTOMATIC WEAPONS: MACHINE GUNS

Automatic weapons came about incrementally, with inventors building on and refining the work of those who came before. The next tick in the gears of progress came in 1883 with the invention of the Maxim gun. Unlike the Gatling gun, which required a human to hand-crank the gun to power it, the Maxim gun harnessed the physical energy from the recoil of the gun's firing to power the process of reloading the next round. Hand-cranking was no longer needed, and once firing was initiated, the gun could continue firing on its own. The machine gun was born.

The machine gun was a marvelous and terrible invention. Unlike semiautomatic weapons, which require the user to pull the trigger for each bullet,

automatic weapons will continue firing so long as the trigger remains held down. Modern machine guns come in all shapes and sizes, from the snub-nosed Uzi that plainclothes security personnel can tuck under their suit jackets to massive chain guns that rattle off thousands of rounds per minute. Regardless of their form, their power is palpable when firing one.

As a Ranger, I carried an M249 Squad Automatic Weapon, or SAW, a single-person light machine gun carried in infantry fire teams. Weighing seventeen pounds without ammunition, the SAW is on the hefty side of what can be considered "hand held." With training, the SAW can be fired from the shoulder standing up in short controlled bursts, but is best used lying on the ground. The SAW comes equipped with two metal bipod legs that can be flipped down to allow the gun to stand elevated off the dirt. One does not simply lay on the ground and fire the SAW, however. The SAW has to be managed; it has to be controlled. When fired, the weapon bucks and moves like a wild animal from the rapid-fire recoil. At a cyclic rate of fire, with the trigger held down, the SAW will fire 800 rounds per minute. That's thirteen bullets streaming out of the barrel per second. At that rate of fire, a gunner will rip through his entire stash of ammunition in under two minutes. The barrel will overheat and begin to melt.

Using the SAW effectively requires discipline. The gunner must lean into the weapon to control it, putting his weight behind it and digging the bipod legs into the dirt to pin the weapon down as it is fired. The gunner fires in short bursts of five to seven rounds at a time to conserve ammunition, keep the weapon on target, and prevent the barrel from overheating. Under heavy firing, the SAW's barrel will glow red hot—the barrel may need to be removed and replaced with a spare before it begins to melt. The gun can't handle its own power.

On the other end of the spectrum of infantry machine guns is the M2 .50 caliber heavy machine gun, the "ma deuce." Mounted on military trucks, the .50 cal is the gun that turns a simple off-road truck into a piece of lethal machinery, the "gun truck." At eighty pounds—plus a fifty-pound tripod—the gun is a behemoth. To fire it, the gunner leans back in the turret to brace him or herself and thumbs down the trigger with both hands. The gun unleashes a powerful THUK THUK THUK as the rounds exit. The half inch–wide bullets can sail over a mile.

Machine guns changed warfare forever. In the late 1800s, the British Army used the Maxim gun to aid in their colonial conquest of Africa, allowing them to take on and defeat much larger forces. For a time, to the British at least, machine guns might have seemed like a weapon that lessened the cost of war. In World War I, however, both sides had machine guns and the result was bloodshed on an

unprecedented scale. At the Battle of the Somme, Britain lost 20,000 men in a single day, mowed down by automatic weapons. Millions died in the trenches of World War I, an entire generation of young men.

Machine guns accelerated the process of killing by harnessing industrial age efficiency in the service of war. Men weren't merely killed by machine guns; they were mowed down, like McCormack's mechanical reaper cutting down stalks of grain. Machine guns are dumb weapons, however. They still have to be aimed by the user. Once initiated, they can continue firing on their own, but the guns have no ability to sense targets. In the twentieth century, weapons designers would take the next step to add rudimentary sensing technologies into weapons —the initial stages of intelligence.

THE FIRST "SMART" WEAPONS

From the first time a human threw a rock in anger until the twentieth century, warfare was fought with unguided weapons. Projectiles—whether shot from a sling, a bow, or a cannon—follow the laws of gravity once released. Projectiles are often inaccurate, and the degree of inaccuracy increases with range. With unguided weapons, destroying the enemy hinged on getting close enough to deliver overwhelming barrages of fire to blanket an area.

In World War II, as rockets, missiles, and bombs increased the range at which combatants could target one another—but not their accuracy—militaries sought to develop methods for precision guidance that would allow weapons to accurately strike targets from long distances. Some attempts to insert intelligence into weapons were seemingly comical, such as behaviorist B. F. Skinner's efforts to control a bomb by the pecking of a pigeon on a target image. Skinner's pigeon-guided bomb might have worked, but it never saw combat. Other attempts to implement onboard guidance measures did, giving birth to the first "smart" weapons: precision-guided munitions (PGMs).

The first successful PGM was the German G7e/T4 *Falke* ("Falcon") torpedo, introduced in 1943. The Falcon torpedo incorporated a new technological innovation: an acoustic homing seeker. Unlike regular torpedoes that traveled in a straight line and could very well miss a passing ship, the Falcon used its homing seeker to account for aiming errors. After traveling 400 meters from the German U-boat (submarine) that launched it, the Falcon would activate its passive acoustic sensors, listening for any nearby merchant ships. It would then steer toward any ships, detonating once it reached them.

The Falcon was used by only three U-boats in combat before being replaced by the upgraded G7es/T5 *Zaunkönig* ("Wren"), which had a faster motor and therefore could hit faster moving Allied navy ships in addition to merchant vessels. Using a torpedo that could home in on targets rather than travel in a straight line had clear military advantages, but it also immediately created complications. Two U-boats were sunk in December 1943 (*U-972*) and January 1944 (*U-377*) when their torpedoes circled back on them, homing in on the sound of their own propeller. In response to this problem, Germany instituted a 400-meter safety limit before activating the homing mechanism. To more fully mitigate against the dangers of a homing torpedo turning back on oneself, German U-boats also began incorporating a tactic of diving immediately after launch and then going completely silent.

The Allies quickly developed a countermeasure to the Wren torpedo. The Foxer, an acoustic decoy towed behind Allied ships, was intended to lure away the Wren so that it detonated harmlessly against the decoy, not the ship itself. The Foxer introduced other problems; it loudly broadcast the Allied convoy's position to other nearby U-boats, and it wasn't long before the Germans introduced the Wren II with an improved acoustic seeker. Thus began the arms race in smart weapons and countermeasures against them.

PRECISION-GUIDED MUNITIONS

The latter half of the twentieth century saw the expansion of PGMs like the Wren into sea, air, and ground combat. Today, they are widely used by militaries around the world in a variety of forms. Sometimes called "smart missiles" or "smart bombs," PGMs use automation to correct for aiming errors and help guide the munition (missile, bomb, or torpedo) onto the intended target. Depending on their guidance mechanism, PGMs can have varying degrees of autonomy.

Some guided munitions have very little autonomy at all, with the human controlling the aimpoint of the weapon throughout its flight. Command-guided weapons are manually controlled by a human remotely via a wire or radio link. For other weapons, a human operator "paints" the target with a laser or radar and the missile or bomb homes in on the laser or radar reflection. In these cases, the human doesn't directly control the movements of the munition, but does control the weapon's aimpoint in real time. This allows the human controller to redirect the munition in flight or potentially abort the attack.

Other PGMs are “autonomous” in the sense that they cannot be recalled once launched, but the munition’s flight path and target are predetermined. These munitions can use a variety of guidance mechanisms. Nuclear-tipped ballistic missiles use inertial navigation systems consisting of gyroscopes and accelerometers to guide the missile to its preselected target point. Submarine-launched nuclear ballistic missiles use star-tracking celestial navigation systems to orient the missile, since the undersea launching point varies. Many cruise missiles look down to earth rather than up to the stars for navigation, using radar or digital scene mapping to follow the contours of the Earth to their preselected target. GPS-guided weapons rely on signals from the constellation of U.S. global positioning system satellites to determine their position and guidance to their target. While many of these munitions cannot be recalled or redirected after launch, the munitions do not have any freedom to select their own targets or even their own navigational route. In terms of the task they are performing, they have very little autonomy, even if they are beyond human control once launched. Their movements are entirely predetermined. The guidance systems, whether internal such as inertial navigation or external such as GPS, are only designed to ensure the munition stays on path to its preprogrammed target. The limitation of these guidance systems, however, is that they are only useful against fixed targets.

Homing weapons are a type of PGM used to track onto moving targets. By necessity since the target is moving, homing munitions have the ability to sense the target and adapt to its movements. Some homing munitions use passive sensors to detect their targets, as the Wren did. Passive sensors listen to or observe the environment and wait for the target to indicate its position by making noise or emitting in the electromagnetic spectrum. Active seekers send out signals, such as radar, to sense a target. An early U.S. active homing munition was the Bat anti-ship glide bomb, which had an active radar seeker to target enemy ships.

Some homing munitions “lock” onto a target, their seeker sensing the target before launch. Other munitions “lock on” after launch; they are launched with the seeker turned off, then it activates to begin looking for the moving target.

An attack dog is a good metaphor for a fire-and-forget homing munition. U.S. pilots refer to the tactic of launching the AIM-120 AMRAAM air-to-air missile in “lock on after launch” mode as going “maddog.” After the weapon is released, it turns on its active radar seeker and begins looking for targets. Like a mad dog in a meat locker, it will go after the first target it sees. Similar to the problem German U-boats faced with the Wren, pilots need to take care to ensure

that the missile doesn't track onto friendly targets. Militaries around the world often use tactics, techniques, and procedures ("TTPs" in military parlance) to avoid homing munitions turning back on themselves or other friendlies, such as the U-boat tactic of diving immediately after firing.

HOMING MUNITIONS HAVE LIMITED AUTONOMY

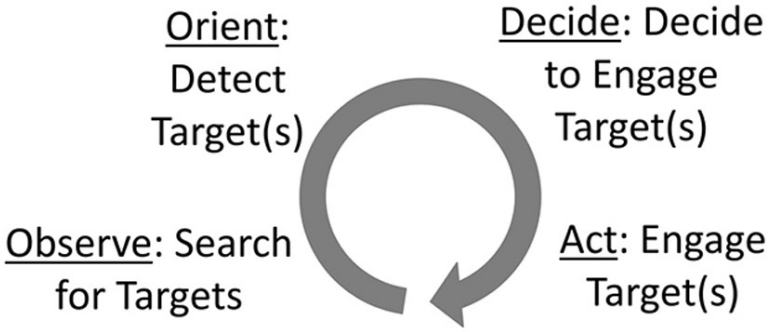
Homing munitions have some autonomy, but they are not "autonomous weapons"—a human still decides which specific target to attack. It's true that many homing munitions are "fire and forget." Once launched, they cannot be recalled. But this is hardly a new development in war. Projectiles have always been "fire and forget" since the sling and stone. Rocks, arrows, and bullets can't be recalled after being released either. What makes homing munitions different is their rudimentary onboard intelligence to guide their behavior. They can sense the environment (the target), determine the right course of action (which way to turn), and then act (maneuvering to hit the target). They are, in essence, a simple robot.

The autonomy given to a homing munition is tightly constrained, however. Homing munitions aren't designed to search for and hunt potential targets on their own. The munition simply uses automation to ensure it hits the specific target the human intended. They are like an attack dog sent by police to run down a suspect, not like a wild dog roaming the streets deciding on its own whom to attack.

In some cases, automation is used to ensure the munition does not hit unintended targets. The Harpoon anti-ship missile has a mode where the seeker stays off while the missile uses inertial navigation to fly a zigzag pattern toward the target. Then, at the designated location, the seeker activates to search for the intended target. This allows the missile to fly past other ships in the environment without engaging them. Because the autonomy of homing munitions is tightly constrained, the human operator needs to be aware of a specific target in advance. There must be some kind of intelligence informing the human of that *particular target* at that *specific time and place*. This intelligence could come from radars based on ships or aircraft, a ping on a submarine's sonar, information from satellites, or some other indicator. Homing munitions have a very limited ability in time and space to search for targets, and to launch one without knowledge of a specific target would be a waste. This means homing munitions must operate as part of a broader *weapon system* to be useful.

THE WEAPON SYSTEM

A weapon system consists of a sensor to search for and detect enemy targets, a decision-making element that decides whether to engage the target, and a munition (or other effector, such as a laser) that engages the target. Sometimes the weapon system is contained on a single platform, such as an aircraft. In the case of an Advanced Medium-Range Air-to-Air Missile (or AMRAAM), for example, the weapon system consists of the aircraft, radar, pilot, and missile. The radar searches for and senses the target, the human decides whether to engage, and the missile carries out the engagement. All of these elements are necessary for the engagement to work.



A *weapon system* consists of the components necessary to complete an entire combat OODA loop: searching for and detecting enemy targets, deciding whether to engage them, and engaging the targets.

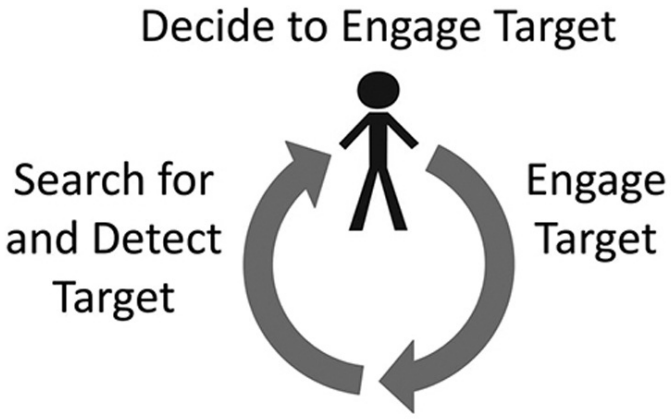
Weapon System OODA Loop

In other cases, components of the weapon system may be distributed across multiple physical platforms. For example, a maritime patrol aircraft might detect an enemy ship and pass the location data to a nearby friendly ship, which launches a missile. Defense strategists refer to this larger, distributed system with multiple components as a *battle network*. Defense analyst Barry Watts described the essential role battle networks play in making precision-guided weapons effective:

> Because "precision munitions" require detailed data on their intended targets or aimpoints to be militarily useful—as opposed to wasteful—they require "precision information." Indeed, the tight linkage between guided munitions and "battle networks," whose primary reason for existence is to provide the necessary targeting information, was one of the major lessons that emerged from careful study of the US-led air campaign during Operation Desert Storm in 1991. . . . [It] is *guided munitions together with the targeting networks that make these munitions "smart."*

[emphasis in the original]

Automation is used for many engagement-related tasks in weapon systems and battle networks: finding, identifying, tracking, and prioritizing potential targets; timing when to fire; and maneuvering munitions to the target. For most weapon systems in use today, a human makes the decision whether to engage the target. If there is a human in the loop deciding which target(s) to engage, it is a *semiautonomous weapon system*.



In semiautonomous weapons, automation may be used to search for and detect targets and carry out the engagement, but the human makes the decision to engage specific targets.

Supervised Autonomous Weapon System (human on the loop)

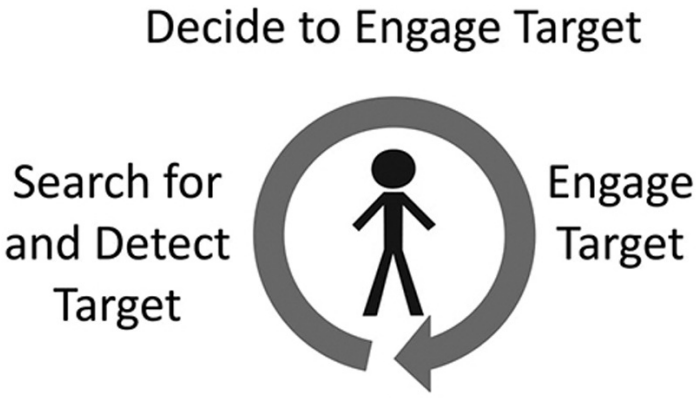
In *autonomous weapon systems*, the entire engagement loop—searching, detecting, deciding to engage, and engaging—is automated. (For ease of use, I'll often shorten "autonomous weapon system" to "autonomous weapon." The terms should be treated as synonymous, with the understanding that "weapon" refers to the entire system: sensor, decision-making element, and munition.) Most weapon systems in use today are semiautonomous, but a few cross the line to autonomous weapons.

SUPERVISED AUTONOMOUS WEAPON SYSTEMS

Because homing munitions can precisely target ships, bases, and vehicles, they can overwhelm defenders through saturation attacks with waves, or "salvos" of missiles. In an era of unguided ("dumb") munitions, defenders could simply ride out an enemy barrage, trusting that most of the incoming rounds would miss.

With precision-guided (“smart”) weapons, however, the defender must find a way to actively intercept and defeat incoming munitions before they impact. More automation—this time for defensive purposes—is the logical response.

At least thirty nations currently employ supervised autonomous weapon systems of various types to defend ships, vehicles, and bases from attack. Once placed in automatic mode and activated, these systems will engage incoming rockets, missiles, or mortars all on their own without further human intervention. Humans are on the loop, however, supervising their operation in real time.



Once activated, supervised autonomous weapons can search for, detect, decide to engage, and engage targets all on their own, but the human can intervene, if necessary.

Supervised Autonomous Weapon System (human on the loop)

These supervised autonomous weapons are necessary for circumstances in which the speed of engagements could overwhelm human operators. Like in the Atari game *Missile Command*, saturation attacks from salvos of simultaneous incoming threats could overwhelm human operators. Automated defenses are a vital part of surviving attacks from precision-guided weapons. They include ship-based defenses, such as the U.S. Aegis combat system and Phalanx Close-In Weapon System (CIWS); land-based air and missile defense systems, such as the U.S. Patriot; counter-rocket, artillery, and mortar systems such as the German MANTIS; and active protection systems for ground vehicles, such as the Israeli Trophy or Russian Arena system.

While these weapon systems are used for a variety of different situations—to defend ships, land bases, and ground vehicles—they operate in similar ways. Humans set the parameters of the weapon, establishing which threats the system should target and which it should ignore. Depending on the system, different rules may be used for threats coming from different directions, angles, and

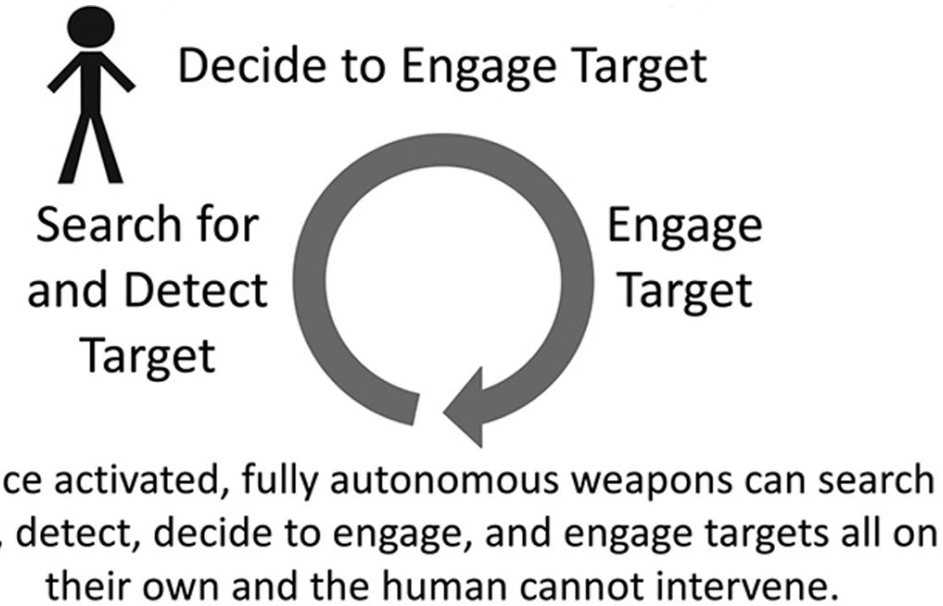
speeds. Some systems may have multiple modes of operation, allowing human in-the-loop (semiautonomous) or on-the-loop (supervised autonomous) control.

These automated defensive systems are autonomous weapons, but they have been used to date in very narrow ways—for immediate defense of human-occupied vehicles and bases, and generally targeting objects (like missiles, rockets, or aircraft), not people. Humans supervise their operation in real time and can intervene, if necessary. And the humans supervising the system are physically colocated with it, which means in principle they could physically disable it if the system stopped responding to their commands.

FULLY AUTONOMOUS WEAPON SYSTEMS

Do any nations have fully autonomous weapons that operate with no human supervision? Generally speaking, fully autonomous weapons are not in wide use, but there are a few select systems that cross the line. These weapons can search for, decide to engage, and engage targets on their own and no human can intervene. Loitering munitions are one example.

Loitering munitions can circle overhead for extended periods of time, searching for potential targets over a wide area and, once they find one, destroy it. Unlike homing munitions, loitering munitions do not require precise intelligence on enemy targets before launch. Thus, a loitering munition is a complete "weapon system" all on its own. A human can launch a loitering munition into a "box" to search for enemy targets without knowledge of any specific targets beforehand. Some loitering munitions keep humans in the loop via a radio connection to approve targets before engagement, making them semiautonomous weapon systems. Some, however, are fully autonomous.



Fully Autonomous Weapon System (human out of the loop)

The Israeli Harpy is one such weapon. No human approves the specific target before engagement. The Harpy has been sold to several countries—Chile, China, India, South Korea, and Turkey—and the Chinese are reported to have reverse engineered their own variant.

HARM vs. Harpy

Type of weapon

Target

Time to search

Distance

Degree of autonomy

HARM

Homing missile

Radars	Approx. 4.5 minutes	90+ km

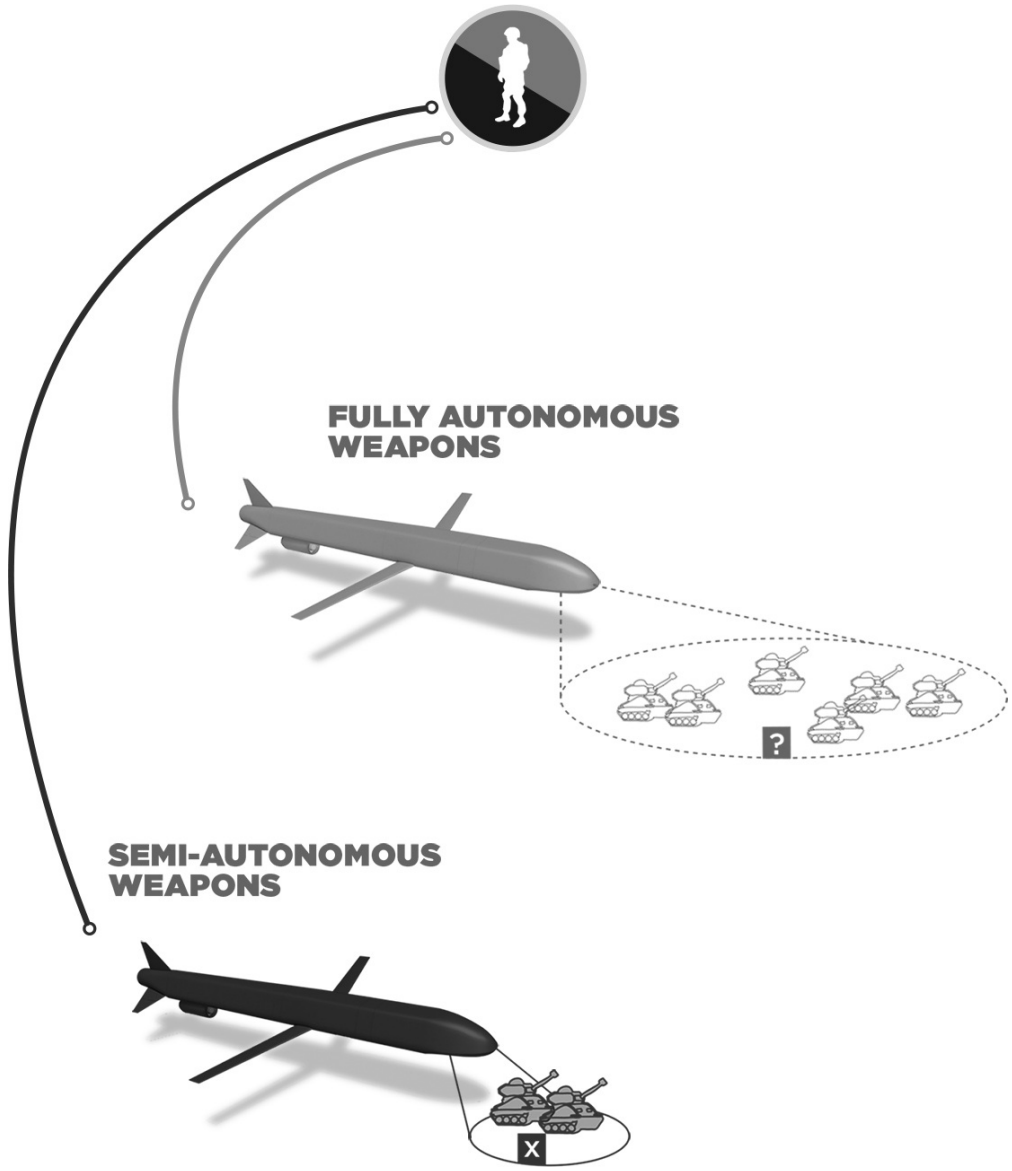
Semiautonomous weapon

Harpy

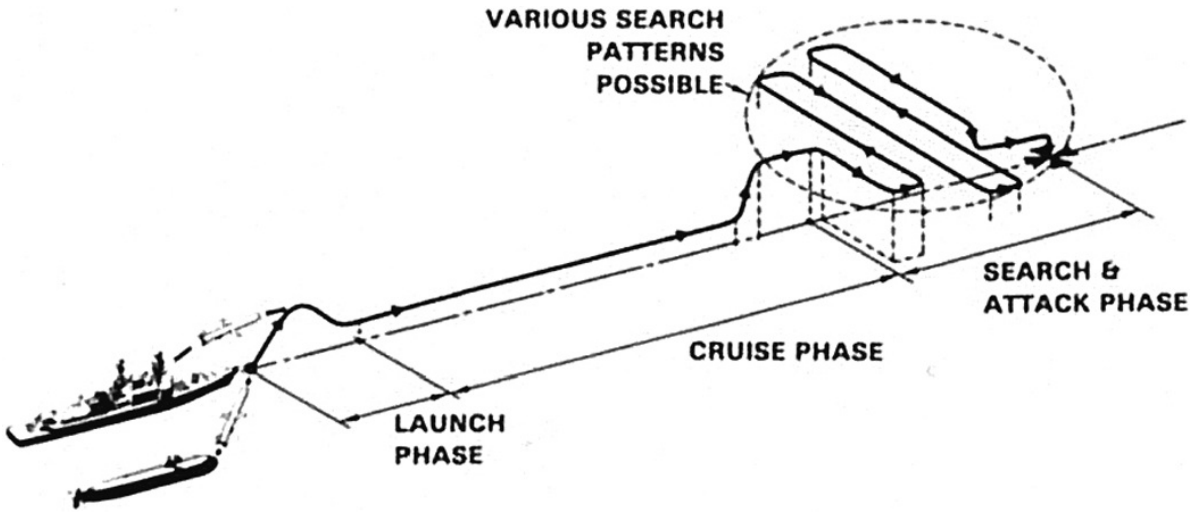
Loitering munition

Radars	2.5 hours	500 km

The difference between a fully autonomous loitering munition and a semiautonomous homing munition can be illustrated by comparing the Harpy with the High-speed AntiRadiation Missile (HARM). Both go after the same type of target (enemy radars), but their freedom to search for targets is massively different. The semiautonomous HARM has a range of 90-plus kilometers and a top speed of over 1,200 kilometers per hour, so it is only airborne for approximately four and a half minutes. Because it cannot loiter, the HARM has to be launched at a specific enemy radar in order to be useful. The Harpy can stay aloft for over two and a half hours covering up to 500 kilometers of ground. This allows the Harpy to operate independently of a broader battle network that gives the human targeting information before launch. The human launching the Harpy decides to destroy *any* enemy radars within a general area in space and time, but the Harpy itself chooses the specific radar it destroys.



Semiautonomous vs. Fully Autonomous Weapons *For semiautonomous weapons, the human operator launches the weapon at a specific known target or group of targets. The human chooses the target and the weapon carries out the attack. Fully autonomous weapons can search for and find targets over a wide area, allowing human operators to launch them without knowledge of specific targets in advance. The human decides to launch the fully autonomous weapon, but the weapon itself chooses the specific target to attack.*



Tomahawk Anti-Ship Missile Mission Profile *A typical mission for a Tomahawk Anti-Ship Missile (TASM). After being launched from a ship or submarine, the TASM would cruise to the target area. Once over the target area, it would fly a search pattern to look for targets and, if it found one, attack the target on its own.*

Despite conventional thinking that fully autonomous weapons are yet to come, isolated cases of fully autonomous loitering munitions go back decades. In the 1980s, the U.S. Navy deployed a loitering anti-ship missile that could hunt for, detect, and engage Soviet ships on its own. The Tomahawk Anti-Ship Missile (TASM) was intended to be launched over the horizon at possible locations of Soviet ships, then fly a search pattern over a wide area looking for their radar signatures. If it found a Soviet ship, TASM would attack it. (Despite the name, the TASM was quite different from the Tomahawk Land Attack Missile [TLAM], which uses digital scene mapping to follow a preprogrammed route to its target.) The TASM was taken out of Navy service in the early 1990s. While it was never fired in anger, it has the distinction being the first operational fully autonomous weapon, a significance that was not recognized at the time.

In the 1990s, the United States began development on two experimental loitering munitions: Tacit Rainbow and the Low Cost Autonomous Attack System (LOCAAS). Tacit Rainbow was intended to be a persistent antiradiation weapon to target land-based radars, like the Harpy. LOCAAS had an even more ambitious goal: to search for and destroy enemy tanks, which are harder targets than radars because they are not emitting in the electromagnetic spectrum. Neither Tacit Rainbow nor LOCAAS were ever deployed; both were cancelled while still in development.

These examples shine a light on a common misperception about autonomous weapons, which is the notion that intelligence is what makes a weapon “autonomous.” How intelligent a system is and which tasks it performs

autonomously are different dimensions. It is freedom, not intelligence, that defines an autonomous weapon. Greater intelligence can be added into weapons without changing their autonomy. To date, the target identification algorithms used in autonomous and semiautonomous weapons have been fairly simple. This has limited the usefulness of fully autonomous weapons, as militaries may not trust giving a weapon very much freedom if it isn't very intelligent. As machine intelligence advances, however, autonomous targeting will become technically possible in a wider range of situations.

UNUSUAL CASES—MINES, ENCAPSULATED TORPEDO MINES, AND SENSOR FUZED WEAPON

There are a few unusual cases of weapons that blur the lines between semiautonomous and fully autonomous weapons: mines and the Sensor Fuzed Weapon deserve special mention.

Placed on land or at sea, mines wait for their target to approach, at which point the mine explodes. While mines are automatic devices that will detonate on their own once triggered, they have no freedom to maneuver and search for targets. They simply sit in place. (For the most part—some naval mines can drift with the current.) They also generally have very limited methods for "deciding" whether or not to fire. Mines typically have a simple method for sensing a target and, when the threshold for the sensor is reached, the mine explodes. (Some naval mines and antitank mines employ a counter so that they will let the first few targets pass unharmed before detonating against a ship or vehicle later in the convoy.) Mines deserve special mention because their freedom in time is virtually unbounded, however. Unless specifically designed to self-deactivate after a certain period of time, mines can lay in wait for years, sometimes remaining active long after a war has ended.

The fact that mines are often unbounded in time has had devastating humanitarian consequences. By the mid-1990s, an estimated more than 110 million land mines lay hidden in sixty-eight countries around the globe, accumulated from scores of conflicts. Land mines have killed thousands of civilians, many of them children, and maimed tens of thousands more, sparking the global movement to ban land mines that culminated in the Ottawa Treaty in 1997. Adopted by 162 nations, the Ottawa Treaty prohibits the production, stockpiling, transfer, or use of antipersonnel land mines. Antitank land mines and naval mines are still permitted.

Mines can sense and act on their own, but do not search for targets. Encapsulated torpedo mines are a special type of naval mine that acts more like an autonomous weapon, however. Rather than simply exploding once activated, encapsulated torpedo mines release a torpedo that homes in on the target. This gives encapsulated torpedo mines the freedom to engage targets over a much wider area than a traditional mine, much like a loitering munition. The U.S. Mk 60 CAPTOR encapsulated torpedo mine had a published range of 8,000 yards. By contrast, a ship would have to pass over a regular mine for it to detonate. Even though encapsulated torpedo mines are moored in place to the seabed, their ability to launch a torpedo to chase down targets gives them a much greater degree of autonomy in space than a traditional naval mine. As with loitering munitions, examples of encapsulated torpedo mines are rare. The U.S. CAPTOR mine was in service for throughout the 1980s and 1990s but has been retired. The only encapsulated torpedo mine still in service is the Russian PMK-2, used by Russia and China.

The Sensor Fuzed Weapon (SFW) is an air-delivered antitank weapon that defies categorization. Released from an aircraft, an SFW can destroy an entire column of enemy tanks within seconds. The SFW functions through a series of Rube Goldberg machine–like steps: First, the aircraft releases a bomb-shaped canister than glides toward the target area. As the canister approaches the target area, the outer casing releases, exposing ten submunitions which are ejected from the canister. Each submunition releases a drogue parachute slowing its descent. At a certain height above the ground, the submunition springs into action. It opens its outer case, exposing four internally held “skeets” which are then rotated out of the inner casing and exposed. The parachute releases and the submunition fires retrojets that cause it to climb in altitude while spinning furiously. The hockey-puck-shaped skeets are then released, flung outward violently from the force of the spinning. Each skeet carries onboard laser and infrared sensors that it uses to search for targets beneath it. Upon detecting a vehicle beneath it, the skeet fires an explosively formed penetrator—a metal slug —downward into the vehicle. The metal slug strikes the vehicle on top, where armored vehicles have the thinnest armor, destroying the vehicle. In this manner, a single SFW can take out a group of tanks or other armored vehicles simultaneously, with the skeets targeting each vehicle precisely.

Similar to the distinction between Harpy and HARM, the critical variable in the evaluating SFW’s autonomy is its freedom in time and space. While the weapon distributes forty skeets over several acres, the time the weapon can search for targets is minuscule. Each skeet can hover with its sensor active for

only a few seconds before firing. Unlike the Harpy, the SFW cannot loiter for an extended period over hundreds of kilometers. The human launching the SFW must know that there is a group of tanks at a particular point in space and time. Like a homing munition, the SFW must be part of a wider weapon system that provides targeting data in order to be useful. The SFW is different than a traditional homing munition, because the SFW can hit multiple objects. This makes the SFW like a salvo of forty homing munitions launched at a tightly geographically clustered set of targets.

PUSHING "START"

Autonomous weapons are defined by the ability to complete the engagement cycle—searching for, deciding to engage, and engaging targets—on their own. Autonomous weapons, whether supervised or fully autonomous, are still built and put into operation by humans, though. Humans are involved in the broader process of designing, building, testing, and deploying weapons.

The fact that there are humans involved at some stage does not change the significance of a weapon that could complete engagements entirely on its own. Even the most highly autonomous system would still have been borne out of a process initiated by humans at some point. In the climactic scene of *Terminator 3: Rise of the Machines*, an Air Force general pushes the button to start Skynet. (Absurdly, this is done with an old "EXECUTE Y/N?" prompt like the kind used in MS-DOS in the 1980s.) From that point forward, Skynet embarks on its path to exterminate humanity, but at least at the beginning a human was in the loop. The question is not whether there was ever a human involved, but rather how much freedom the system has once it is activated.

WHY AREN'T THERE MORE AUTONOMOUS WEAPONS?

Automation has been used extensively in weapons around the world for decades, but the amount of freedom given to weapons has been, up to now, fairly limited. Homing munitions have seekers, but their ability to search for targets is narrowly constrained in time and space. Supervised autonomous weapons have only been used for limited defensive purposes. The technology to build simple fully autonomous loitering munitions like TASM and Harpy has existed for decades, yet there is only one example in use today.

Why aren't there more fully autonomous weapons? Homing munitions and even semiautonomous loitering munitions are widely used, but militaries have not aggressively pursued fully autonomous loitering munitions. The U.S. experience with TASM may shed some light on why. TASM was in service in the U.S. Navy from 1982 to 1994, when it was retired. To understand better why TASM was taken out of service, I spoke with naval strategist Bryan McGrath.

McGrath, a retired Navy officer, is well known in Washington defense circles. He is a keen strategist and unabashed advocate of sea power who thinks deeply about the past, present, and future of naval warfare. McGrath is familiar with TASM and other anti-ship missiles such as the Harpoon, and was trained on TASM in the 1980s when it was in the fleet.

McGrath explained to me that TASM could outrange the ship's own sensors. That meant that initial targeting had to come from another sensor, such as a helicopter or maritime patrol aircraft that detected an enemy ship. The problem, as McGrath described it, was a "lack of confidence in how the targeting picture would change from the time you fired the missile until you got it downrange." Because the target could move, unless there was an "active sensor" on the target, such as a helicopter with eyes on the target the whole time, the area of uncertainty of where the target was would grow over time.

The ability of the TASM to search for targets over a wide area mitigated, to some extent, this large area of uncertainty. If the target had moved, the TASM could simply fly a search pattern looking for it. But TASM didn't have the ability to accurately discriminate between enemy ships and merchant vessels that just happened to be in its path. As the search area widened, the risk increased that the TASM might run across a merchant ship and strike it instead. In an all-out war with the Soviet Navy, that risk might be acceptable, but in any situations short of that, getting approval to shoot the TASM was unlikely. TASM was, according to McGrath, "a weapon we just didn't want to fire."

Another factor was that if a TASM was launched and there wasn't a valid target within the search area of the weapon, the weapon would be wasted. McGrath would be loath to launch a weapon on scant evidence that there was a valid target in the search area. "I would want to know that there's something there, even if there was some kind of end-game autonomy in place." Why? "Because the weapons cost money," he said, "and I don't have a lot of them. And I may have to fight tomorrow."

Modern missiles can cost upwards of a million dollars apiece. As a practical matter, militaries will want to know that there is, in fact, a valid enemy target in the area before using an expensive weapon. One of the reasons militaries have

not used fully autonomous loitering munitions more may be the fact that the advantage they bring—the ability to launch a weapon without precise targeting data in advance—may not be of much value if the weapon is not reusable, since the weapon could be wasted.

FUTURE WEAPONS

The trend of creeping automation that began with Gatling's gun will continue. Advances in artificial intelligence will enable smarter weapons, which will be capable of more autonomous operation. At the same time, another facet of the information revolution is greater networking. German U-boats couldn't control the Wren torpedo once it was launched, not because they didn't want to; they simply had no means to do so.

Modern munitions are increasingly networked to allow them to be controlled or retargeted after they've been launched. Wire-guided munitions have existed for decades, but are only feasible for short distances. Long-range weapons are now incorporating datalinks to allow them to be controlled via radio communication, even over satellites. The Block IV Tomahawk Land Attack Missile (TLAM-E, or Tactical Tomahawk) includes a two-way satellite communications link that allows the weapon to be retargeted in flight. The Harpy 2, or Harop, has a communications link that allows it to be operated in a human-in-the-loop mode so that the human operator can directly target the weapon.

When I asked McGrath what feature he would most desire in a future weapon, it wasn't autonomy—it was a datalink. "You've got to talk to the missile," he explained. "The missiles have to be part of a network." Connecting the weapons to the network would allow you to send updates on the target while in flight. As a result, "confidence in employing that weapon would dramatically increase."

A networked weapon is a far more valuable weapon than one that is on its own. By connecting a weapon to the network, the munition becomes part of a broader system and can harness sensor data from other ships, aircraft, or even satellites to assist its targeting. Additionally, the commander can keep control of the weapon while in flight, making it less likely to be wasted. One advantage to the networked Tactical Tomahawk, for example, is the ability for humans to use sensors on the missile to do battle damage assessment (BDA) of potential targets before striking. Without the ability to conduct BDA of the target, commanders might have to launch several Tomahawks at a target to ensure its destruction, since the first missile might not completely destroy the target. Onboard BDA allows the commander to look at the target after the first missile hits. If more strikes are needed, more missiles can be used. If not, then subsequent missiles

can be diverted in flight to secondary targets.

Everything has a countermeasure, though, and increased networking runs counter to another trend in warfare, the rise of electronic attack. The more that militaries rely on the electromagnetic spectrum for communications and sensing targets, the more vital it will be to win the invisible electronic war of jamming, spoofing, and deception fought through the electromagnetic spectrum. In future wars between advanced militaries, communications in contested environments is by no means assured. Advanced militaries have ways of communicating that are resistant to jamming, but they are limited in range and bandwidth. When communications are denied, missiles or drones will be on their own, reliant on their onboard autonomy.

Due to their expensive cost, even highly advanced loitering munitions are likely to fall into the same trap as TASM, with commanders hesitant to fire them unless targets are clearly known. But drones change this equation. Drones can be launched, sent on patrol, and can return with their weapons unused if they do not find any targets. This simple feature—reusability—dramatically changes how a weapon could be used. Drones could be sent to search over a wide area in space and time to hunt for enemy targets. If none were found, the drone could return to base to hunt again another day.

More than ninety nations and non-state groups already have drones, and while most are unarmed surveillance drones, an increasing number are armed. At least sixteen countries already possess armed drones and another dozen or more nations are working on arming their drones. A handful of countries are even pursuing stealth combat drones specifically designed to operate in contested areas. For now, drones are used as part of traditional battle networks, with decision-making residing in the human controller. If communications links are intact, then countries can keep a human in the loop to authorize targets. If communications links are jammed, however, what will the drones be programmed to do? Will they return home? Will they carry out surveillance missions, taking pictures and reporting back to their human operators? Will the drones be authorized to strike fixed targets that have been preauthorized by humans, much like cruise missiles today? What if the drones run across emerging targets of opportunity that have not been authorized in advance by a human—will they be authorized to fire? What if the drones are fired upon? Will they be allowed to fire back? Will they be authorized to shoot first?

These are not hypothetical questions for the future. Engineers around the globe are programming the software for these drones today. In their hands, the future of autonomous weapons is being written.

PART II

Building the Terminator

4

THE FUTURE BEING BUILT TODAY

AUTONOMOUS MISSILES, DRONES, AND ROBOT SWARMS

Few actors loom larger in the robotics revolution than the U.S. Department of Defense. The United States spends 600 billion dollars annually on defense, more than the next seven countries combined. Despite this, U.S. defense leaders are concerned about the United States falling behind. In 2014, the United States launched a "Third Offset Strategy" to reinvigorate America's military technological advantage. The name harkens back to the first and second "offset strategies" in the Cold War, where the U.S. military invested in nuclear weapons in the 1950s and later precision-guided weapons in the 1970s to offset the Soviet Union's numerical advantages in Europe. The centerpiece of DoD's Third Offset Strategy is robotics, autonomy, and human-machine teaming.

Many applications of military robotics and autonomy are noncontroversial, such as uninhabited logistics convoys, tanker aircraft, or reconnaissance drones. Autonomy is also increasing in weapon systems, though, with next-generation missiles and combat aircraft pushing the boundaries of autonomy. A handful of experimental programs show how the U.S. military is thinking about the role of autonomy in weapons. Collectively, they are laying the foundations for the military of the future.

SALTY DOGS: THE X-47B DRONE

The X-47B experimental drone is one of the world's most advanced aircraft.

Only two have been ever built, named Salty Dog 501 and Salty Dog 502. With a sleek bat-winged shape that looks like something out of the 1980s sci-fi flick *Flight of the Navigator*, the X-47B practically screams "the future is here." In their short life-span as experimental aircraft from 2011 to 2015, Salty Dog 501 and 502 repeatedly made aviation history. The X-47B was the first uninhabited (unmanned) aircraft to autonomously take off and land on an aircraft carrier and the first uninhabited aircraft to autonomously refuel from another plane while in flight. These are key milestones to enabling future carrier-based combat drones. However, the X-47B was not a combat aircraft. It was an experimental "X-plane," a demonstration program designed to mature technologies for a follow-on aircraft. The focus of technology development was automating the physical movement of the aircraft—takeoff, landing, flight, and aerial refueling. The X-47B did not carry weapons or sensors that would permit it to make engagements.

The Navy has stated their first operational carrier-based drone will be the MQ-25 Stingray, a future aircraft that is still on the drawing board. While the specific design has yet to be determined, the MQ-25 is envisioned primarily as a tanker, ferrying fuel for manned combat aircraft such as the F-35 Joint Strike Fighter, with possibly a secondary role in reconnaissance. It is not envisioned as a combat aircraft. In fact, over the past decade the Navy has moved steadily away from any notion of uninhabited aircraft in combat roles.

The origin of the X-47 was in the Joint Unmanned Combat Air Systems (J-UCAS) program, a joint program between DARPA, the Navy, and the Air Force in the early 2000s to design an uninhabited combat aircraft. J-UCAS led to the development of two experimental X-45A aircraft, which in 2004 demonstrated the first drone designed for combat missions. Most drones today are intended for surveillance missions, which means they are designed for soaring and staying aloft for long periods of time. The X-45A, however, sported the same sharply angled wings and smooth top surfaces that define stealth aircraft like the F-117, B-2 bomber, and F-22 fighter. Designed to penetrate enemy air defenses, the intent was for the X-45A to perform close in jamming and strike missions in support of manned aircraft. The program was never completed, though. In the Pentagon's 2006 Quadrennial Defense Review, a major strategy and budget review conducted every four years, the J-UCAS program was scrapped and restructured.

J-UCAS's cancellation was curious because it came at the height of the post-9/11 defense budget boom and at a time when the Defense Department was waking up to the potential of robotic systems more broadly. Even while the military was deploying thousands of drones to Iraq and Afghanistan, the Air

Force was highly resistant to the idea of uninhabited aircraft taking on combat roles in future wars. In the ensuing decade since J-UCAS's cancellation, despite repeated opportunities, the Air Force has not restarted a program to build a combat drone. Drones play important roles in reconnaissance and counterterrorism, but when it comes to dogfighting against other enemy aircraft or taking down another country's air defense network, those missions are currently reserved for traditional manned aircraft.

The reality is that what may look from the outside like an unmitigated rush toward robotic weapons is, in actuality, a much more muddled picture inside the Pentagon. There is intense cultural resistance within the U.S. military to handing over combat jobs to uninhabited systems. Robotic systems are frequently embraced for support roles such as surveillance or logistics, but rarely for combat applications. The Army is investing in logistics robots, but not frontline armed combat robots. The Air Force uses drones heavily for surveillance, but is not pursing air-to-air combat drones. Pentagon vision documents such as the Unmanned Systems Roadmaps or the Air Force's 2013 *Remotely Piloted Aircraft Vector* often articulate ambitious dreams for robots in a variety of roles, but these documents are often disconnected from budgetary realities. Without funding, these visions are more hallucinations than reality. They articulate goals and aspirations, but do not necessarily represent the most likely future path.

The downscoping of the ambitious J-UCAS combat aircraft to the plodding MQ-25 tanker is a great case in point. In 2006 when the Air Force abandoned the J-UCAS experimental drone program, the Navy continued a program to develop a combat aircraft. The X-47B was supposed to mature the technology for a successor stealth drone, but in a series of internal Pentagon memoranda issued in 2011 and 2012, Navy took a sharp turn away from a combat aircraft. Designs were scaled back in favor of a less ambitious nonstealthy surveillance drone. Concept sketches shifted from looking like the futuristic sleek and stealthy X-45A and X-47B to the more pedestrian Predator and Reaper drones, already over a decade old at that point. The Navy, it appears, wasn't immune to the same cultural resistance to combat drones found in the Air Force.

The Navy's resistance to developing an uninhabited combat aerial vehicle (UCAV) is particularly notable because it comes in the face of pressure from Congress and a compelling operational need. China has developed anti-ship ballistic and cruise missiles that can outrange carrier-based F-18 and F-35 aircraft. Only uninhabited aircraft, which can stay aloft far longer than would be possible with a human in the airplane, have sufficient range to keep the carrier relevant in the face of advanced Chinese missiles. Sea power advocates outside

the Navy in Congress and think tanks have argued that without a UCAV on board, the aircraft carrier itself would be of limited utility against a high-technology opponent. Yet the Navy's current plan is for its carrier-based drone, the MQ-25, to ferry gas for human-inhabited jets. For now, the Navy is deferring any plans for a future UCAV.

The X-47B is an impressive machine and, to an outside observer, it may seem to portend a future of robot combat aircraft. Its appearance belies the reality that within the halls of the Pentagon, however, there is little enthusiasm for combat drones, much less fully autonomous ones that would target on their own. Neither the Air Force nor the Navy have programs under way to develop an operational UCAV. The X-47B is a bridge to a future that, at least for now, doesn't exist.

THE LONG-RANGE ANTI-SHIP MISSILE

The Long-Range Anti-Ship Missile (LRASM) is a state-of-the-art missile pushing the boundaries of autonomy. It is a joint DARPA-Navy-Air Force project intended to fill a gap in the U.S. military's ability to strike enemy ships at long ranges. Since the retirement of the TASM, the Navy has relied on the shorter-range Harpoon anti-ship missile, which has a range of only 67 nautical miles. The LRASM, on the other hand, can fly up to 500 nautical miles. LRASM also sports a number of advanced survivability features, including the ability to autonomously detect and evade threats while en route to its target.

LRASM uses autonomy in several novel ways, which has alarmed some opponents of autonomous weapons. The LRASM has been featured in no less than three *New York Times* articles, with some critics claiming it exhibits "artificial intelligence outside human control." In one of the articles, Steve Omohundro, a physicist and leading thinker on advanced artificial intelligence, stated "an autonomous weapons arms race is already taking place." It is a leap, though, to assume that these advances in autonomy mean states intend to pursue autonomous weapons that would hunt for target on their own.

The actual technology behind LRASM, while cutting edge, hardly warrants these breathless treatments. LRASM has many advanced features, but the critical question is who chooses LRASM's targets—a human or the missile itself? On its website, Lockheed Martin, the developer of LRASM, states:

> LRASM employs precision routing and guidance. . . . The missile employs a multi-modal sensor suite, weapon data link, and enhanced digital anti-jam Global Positioning System to detect and

> destroy specific targets within a group of numerous ships at sea. . . . This advanced guidance operation means the weapon can use gross target cueing data to find and destroy its predefined target in denied environments.

While the description speaks of advanced precision guidance, it doesn't say much that would imply artificial intelligence that would hunt for targets on its own. What was the genesis of the criticism? Well . . . Lockheed used to describe LRASM differently.

Before the first *New York Times* article in November 2014, Lockheed's description of LRASM boasted much more strongly of its autonomous features. It used the word "autonomous" three times in the description, describing it as an "autonomous, precision-guided anti-ship" missile that "cruises autonomously" and has an "autonomous capability." What exactly the weapon was doing autonomously was somewhat ambiguous, though.

After the first *New York Times* article, the description changed, substituting "semiautonomous" for "autonomous" in multiple places. The new description also clarified the nature of the autonomous features, stating "The semiautonomous guidance capability gets LRASM safely to the enemy area." Eventually, even the words "semiautonomous" were removed, leading to the description online today which only speaks of "precision routing and guidance" and "advanced guidance." Autonomy isn't mentioned at all.

What should we make of this shifting story line? Presumably the weapon's functionality hasn't changed, merely the language used to describe it. So how autonomous is LRASM?

Lockheed has described LRASM as using "gross target cueing data to find and destroy its predefined target in denied environments." If "predefined" target means that the specific target has been chosen in advance by a human operator, LRASM would be a semiautonomous weapon. On the other hand, if "predefined" means that the human has chosen only a general class of targets, such as "enemy ships," and given the missile the freedom to hunt for these targets over a wide area and engage them on its own, then it would be an autonomous weapon.

Helpfully, Lockheed posted a video online that explains LRASM's functionality. In a detailed combat simulation, the video shows precisely which engagement-related functions would be done autonomously and which by a human. In the video, a satellite identifies a hostile surface action group (SAG)—a group of enemy ships—and relays their location to a U.S. destroyer. The video shows a U.S. sailor looking at the enemy ships on his console. He presses a button and two LRASMs leap from their launching tubes in a blast of flame into

the air. The text on the video explains the LRASMs have been launched against the enemy cruiser, part of the hostile SAG. Once airborne, the LRASMs establish a line-of-sight datalink with the ship. As they continue to fly out toward the enemy SAG, they transition to satellite communications. A U.S. F/A-18E fighter aircraft then fires a third LRASM (this one air-launched) against an enemy destroyer, another ship in the SAG. The LRASMs enter a "communications and GPS-denied environment." They are now on their own.

The LRASMs maneuver via planned navigational routing, moving from one predesignated way point to another. Then, unexpectedly, the LRASMs encounter a "pop-up threat." In the video, a large red bubble appears in the sky, a no-go zone for the missiles. The missiles now execute "autonomous routing," detouring around the red bubble on their own. A second pop-up threat appears and the LRASMs modify their route again, moving around the threat to continue on their mission.

As the LRASMs approach their target destination, the video shifts to a new perspective focusing on a single missile, simulating what the missile's sensors see. Five dots appear on the screen representing objects detected by the missile's sensors, labeled "ID:71, ID:56, ID:44, ID:24, ID:19." The missile begins a process the video calls "organic [area of uncertainty] reduction." That's military jargon for a bubble of uncertainty. When the missile was launched, the human launching it knew where the enemy ship was located, but ships move. By the time the missile arrives at the ship, the ship could be somewhere else. The "area of uncertainty" is the bubble within which the enemy ship could be, a bubble that gets larger over time.

Since there could be multiple ships in this bubble, the LRASM begins to narrow down its options to determine which ship was the one it was sent to destroy. How this occurs is not specified, but on the video a large "area of uncertainty" appears around all the dots, then quickly shrinks to surround only three of them: ID:44, ID:24, and ID:19. The missile then moves to the next phase of its targeting process: "target classification." The missile scans each object, finally settling on ID:24. "Criteria match," the video states, "target classified." ID:24, the missile has determined, is the ship it was sent to destroy.

Having zeroed in on the right target, the missiles begin their final maneuvers. Three LRASMs descend below the enemy ships' radars to skim just above the water's surface. On their final approach, the missiles scan the ships one last time to confirm their targets. The enemy ships fire their defenses to try to hit the incoming missiles, but it's too late. Two enemy ships are hit.

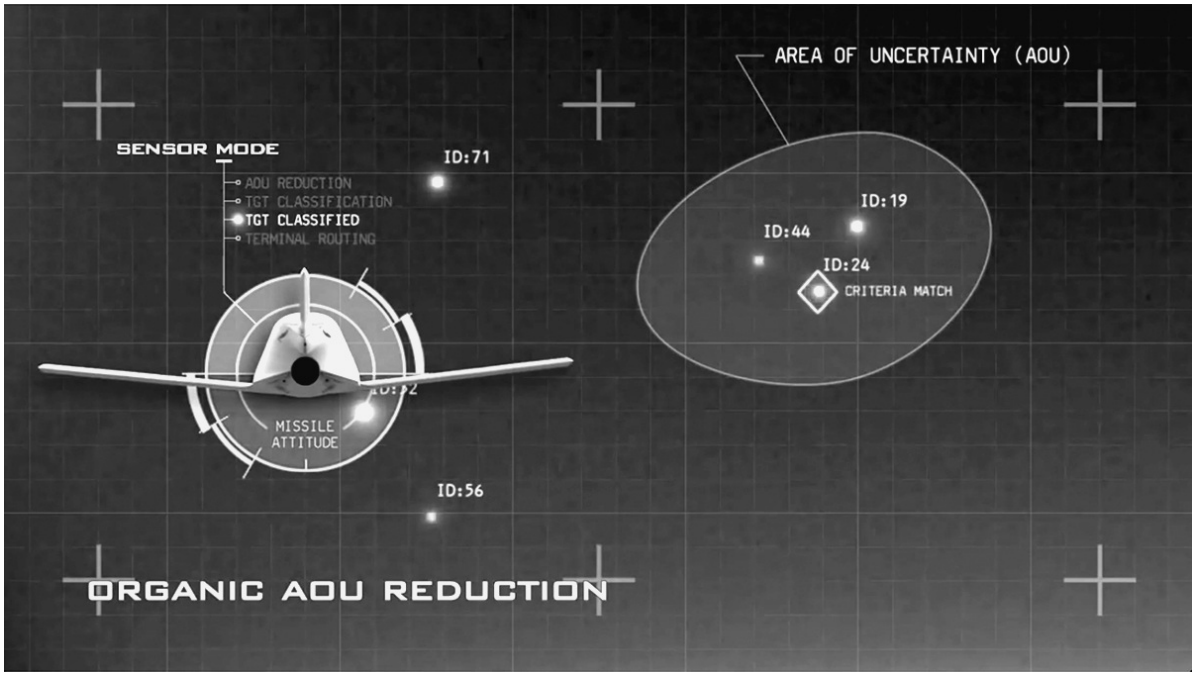
The video conveys the LRASM's impressive autonomous features, but is it

an autonomous weapon? The autonomous/semiautonomous/advanced guidance described on the website is clearly on display. In the video, midway through the flight the missiles enter a “communications and GPS denied environment.” Within this bubble, the missiles are on their own; they cannot call back to human controllers. Any actions they take are autonomous, but the type of actions they can take are limited. Just because the weapon is operating without a communications link to human controllers doesn’t mean it has the freedom to do anything it wishes. The missile isn’t a teenager whose parents have left town for the weekend. It has only been programmed to perform certain tasks autonomously. The missile can identify pop-up threats and autonomously reroute around them, but it doesn’t have the freedom to choose its own targets. It can identify and classify objects to confirm which object was the one it was sent to destroy, but that isn’t the same as being able to *choose* which target to destroy.

Screenshots from LRASM Video *In a video simulation depicting how the LRASM functions, a satellite transmits the location of enemy ships to a human, who authorizes the attack on those specific enemy ships.*

The LRASMs are launched against specific enemy ships, in this case a "SAG Cruiser."

While en route to their human-designated targets, the LRASMs employ autonomous routing around pop-up threats (shown as a bubble).



Because the human-designated target is a moving ship, by the time the LRASM arrives at the target area there is an "area of uncertainty" that defines the ship's possible location. Multiple objects are identified within this area of uncertainty. LRASM uses its onboard ("organic") sensors to reduce the area of uncertainty and identify the human-designated target. LRASM confirms "ID:24" is the target it was sent to destroy. While the missile has many advanced features, it does not choose its own target. The missile uses its sensors to confirm the human-selected target.

It is the human who decides which enemy ship to destroy. The critical point in the video isn't at the end of the missile's flight as it zeroes in on the ship—it's at the beginning. When the LRASMs are launched, the video specifies that they are launched against the "SAG cruiser" and "SAG destroyer." The humans are launching the missiles at specific ships, which the humans have tracked and identified via satellites. The missiles' onboard sensors are then used to confirm the targets before completing the attack. LRASM is only one piece of a weapon *system* that consists of the satellite, ship/aircraft, human, and missile. The human is "in the loop," deciding which specific targets to engage in the broader decision cycle of the weapon system. The LRASM merely carries out the engagement.

BREAKING THE SPEED LIMIT: FAST LIGHTWEIGHT AUTONOMY

Dr. Stuart Russell is a pioneering researcher in artificial intelligence. He literally wrote the textbook that is used to teach AI researchers around the world. Russell

is also one of the leaders in the AI community calling for a ban on "offensive autonomous weapons beyond meaningful human control." One research program Russell has repeatedly raised concerns about is DARPA's Fast Lightweight Autonomy (FLA).

FLA is a research project to enable high-speed autonomous navigation in congested environments. Researchers outfit commercial off-the-shelf quadcopters with custom sensors, processors, and algorithms with the goal of making them autonomously navigate through the interior of a cluttered warehouse at speeds up to forty-five miles per hour. In a press release, DARPA compared the zooming quadcopters to the Millennium Falcon zipping through the hull of a crashed Star Destroyer in *Star Wars: The Force Awakens.* (I would have gone with the Falcon maneuvering through the asteroid field in *The Empire Strikes Back* . . . or the Falcon zipping through the interior of Death Star II in *The Return of the Jedi.* But you get the idea: fast = awesome.) In a video accompanying the press release, shots of the flying quadcopters are set to peppy instrumental music. It's incongruous because in the videos released so far the drones aren't actually moving through obstacles at 45 mph . . . yet. For now, they are creeping their way around obstacles, but they are doing so fully autonomously. FLA's quadcopters use a combination of high-definition cameras, sonar, and laser light detection and ranging (LIDAR) to sense obstacles and avoid them all on their own.

Autonomous navigation around obstacles, even at slow speeds, is no mean feat. The quadcopter's sensors need to detect potential obstacles and track them as the quadcopter moves, a processor-hungry task. Because the quadcopter can only carry so much computing power, it is limited in how quickly it can process the obstacles it sees. The program aims in the coming months to speed it up. As DARPA program manager Mark Micire explained in a press release, "The challenge for the teams now is to advance the algorithms and onboard computational efficiency to extend the UAVs' perception range and compensate for the vehicles' mass to make extremely tight turns and abrupt maneuvers at high speeds." In other words, to pick up the pace.

FLA's quadcopters don't look menacing, but it isn't because of the up-tempo music or the cutesy *Star Wars* references. It's because there's nothing in FLA that has anything to do with weapons engagements. Not only are the quadcopters unarmed, they aren't performing any tasks associated with searching for and identifying targets. DARPA explains FLA's intended use as indoor reconnaissance:

> FLA technologies could be especially useful to address a pressing surveillance shortfall: Military

> teams patrolling dangerous overseas urban environments and rescue teams responding to disasters such as earthquakes or floods currently can use remotely piloted unmanned aerial vehicles (UAVs) to provide a bird's-eye view of the situation, but to know what's going on inside an unstable building or a threatening indoor space often requires physical entry, which can put troops or civilian response teams in danger. The FLA program is developing a new class of algorithms aimed at enabling small UAVs to quickly navigate a labyrinth of rooms, stairways and corridors or other obstacle-filled environments without a remote pilot.

To better understand what FLA was doing, I caught up with one of the project's research teams from the University of Pennsylvania's General Robotics Automation Sensing and Perception (GRASP) lab. Videos of GRASP's nimble quadcopters have repeatedly gone viral online, showing swarms of drones artfully zipping through windows, seemingly dancing in midair, or playing the James Bond theme song on musical instruments. I asked Dr. Daniel Lee and Dr. Vijay Kumar, the principal investigators of GRASP's work on FLA, what they thought about the criticism that the program was paving the way toward autonomous weapons. Lee explained that GRASP's research was "very basic" and focused on "fundamental capabilities that are generally applicable across all of robotics, including industrial and consumer uses." The technology GRASP was focused on "localization, mapping, obstacle detection and high-speed dynamic navigation." Kumar added that their motivations for this research were "applications to search and rescue and first response where time-critical response and navigation at high speeds are critical."

Kumar and Lee aren't weapons designers, so it may not be at the forefront of their minds, but it's worth pointing out that the technologies FLA is building aren't even the critical ones for autonomous weapons. Certainly, fast-moving quadcopters could have a variety of applications. Putting a gun or bomb on an FLA-empowered quadcopter isn't enough to make it an autonomous weapon, however. It would still need the ability to find targets on its own. Depending on the intended target, that may not be particularly complicated, but at any rate that's a separate technology. All FLA is doing is making quadcopters maneuver faster indoors. Depending on one's perspective, that could be cool or could be menacing, but either way FLA doesn't have anything more to do with autonomous weapons than self-driving cars do.

DARPA's description of FLA didn't seem to stack up against Stuart Russell's criticism. He has written that FLA and another DARPA program "foreshadow planned uses of [lethal autonomous weapon systems]." I first met Russell on the sidelines of a panel we both spoke on at the United Nations meetings on autonomous weapons in 2015. We've had many discussions on autonomous weapons since then and I've always found him to be thoughtful,

unsurprising given his prominence in his field. So I reached out to Russell to better understand his concerns. He acknowledged that FLA wasn't "cleanly directed only at autonomous weapon capability," but he saw it as a stepping stone toward something truly terrifying.

FLA is different from projects like the X-47B, J-UCAS, or LRASM, which are designed to engage highly sophisticated adversaries. Russell has a very different kind of autonomous weapon in mind, a swarm of millions of small, fast-moving antipersonnel drones that could wipe out an entire urban population. Russell described these lethal drones used en masse as a kind of "weapon of mass destruction." He explained, "You can make small, lethal quadcopters an inch in diameter and pack several million of them into a truck and launch them with relatively simple software and they don't have to be particularly effective. If 25 percent of them reach a target, that's plenty." Used in this way, even small autonomous weapons could devastate a population.

There's nothing to indicate that FLA is aimed at developing the kind of people-hunting weapon Russell describes, something he acknowledges. Nevertheless, he sees indoor navigation as laying the building blocks toward antipersonnel autonomous weapons. "It's certainly one of the things you'd like to do if you were wanting to develop autonomous weapons," he said.

It's worth nothing that Russell isn't opposed to the military as a whole or even military investments in AI or autonomy in general. He said that some of his own AI research is funded by the Department of Defense, but he only takes money for basic research, not weapons. Even a program like FLA that isn't specifically aimed at weapons still gives Russell pause, however. As a researcher, he said, it's something that he would "certainly think twice" about working on.

WEAPONS THAT HUNT IN PACKS: COLLABORATIVE OPERATIONS IN DENIED ENVIRONMENTS

Russell also raised concerns about another DARPA program: Collaborative Operations in Denied Environments (CODE). According to DARPA's official description, CODE's purpose is to develop "collaborative autonomy—the capability of groups of [unmanned aircraft systems] to work together under a single person's supervisory control." In a press release, CODE's program manager, Jean-Charles Ledé, described the project more colorfully as enabling drones to work together "just as wolves hunt in coordinated packs with minimal

communication."

The image of drones hunting in packs like wolves might be a little unsettling to some. Ledé clarified that the drones would remain under the supervision of a human: "multiple CODE-enabled unmanned aircraft would collaborate to find, track, identify and engage targets, all under the command of a single human mission supervisor." Graphics on DARPA's website depicting how CODE might work show communications relay drones linking the drone pack back to a manned aircraft removed from the edge of the battlespace. So, in theory, a human would be in the loop.

CODE is designed for "contested electromagnetic environments," however, where "bandwidth limitations and communications disruptions" are likely to occur. The means that the communications link to the human-inhabited aircraft might be limited or might not work at all. CODE aims to overcome these challenges by giving drones greater intelligence and autonomy so that they can operate with minimal supervision. Cooperative behavior is central to this concept. With cooperative behavior, one person can tell a group of drones to achieve a goal, and the drones can divvy up tasks on their own.

In CODE, the drone team finds and engages "mobile or rapidly relocatable targets," that is, targets whose locations cannot be specified in advance by a human operator. If there is a communications link to a human, then the human could authorize targets for engagement once CODE air vehicles find them. Communications are challenging in contested electromagnetic environments, but not impossible. U.S. fifth-generation fighter aircraft use low probability of intercept / low probability of detection (LPI/LPD) methods of communicating stealthily inside enemy air space. While these communications links are limited in range and bandwidth, they do exist. According to CODE's technical specifications, developers should count on no more than 50 kilobits per second of communications back to the human commander, essentially the same as a 56K dial-up modem circa 1997.

Keeping a human in the loop via a connection on par with a dial-up modem would be a significant change from today, where drones stream back high-definition full-motion video. How much bandwidth is required for a human to authorize targets? Not much, in fact. The human brain is extremely good at object recognition and can recognize objects even in relatively low resolution images. Snapshots of military objects and the surrounding area on the order of 10 to 20 kilobytes in size may be fuzzy to the human eye, but are still of sufficiently high resolution that an untrained person can discern trucks or military vehicles. A 50 kilobit per second connection could transmit one image

of this size every two to three seconds (1 kilobyte = 8 kilobits). This would allow the CODE air vehicles to identify potential targets and send them back to a human supervisor who would approve (or disapprove) each specific target before attack.

But is this what CODE intends? CODE's public description explains that the aircraft will operate "under a single person's supervisory control," but does not specify that the human would need to approve each target before engagement. As is the case with all of the systems encountered so far, from thermostats to next-generation weapons, the key is which tasks are being performed by the human and which by the machine. Publicly available information on CODE presents a mixed picture.

A May 2016 video released online of the human-machine interface for CODE shows a human authorizing each specific individual target. The human doesn't directly control the air vehicles. The human operator commands four groups of air vehicles, labeled Aces, Badger, Cobra, and Disco groups. The groups, each composed of two to four air vehicles, are given high-level commands such as "orbit here" or "follow this route." Then the vehicles coordinate among themselves to accomplish the task.

Disco Group is sent on a search and destroy mission: "Disco Group search and destroy all [antiaircraft artillery] in this area." The human operator sketches a box with his cursor and the vehicles in Disco Group move into the box. "Disco Group conducting search and destroy at Area One," the computer confirms.

As the air vehicles in Disco Group find suspected enemy targets, they cue up their recommended classification to the human for confirmation. The human clicks "Confirm SCUD" and "Confirm AAA" [antiaircraft artillery] on the interface. But confirmation does not mean approval to fire. A few seconds later, a beeping tone indicates that Disco Group has drawn up a strike plan on a target and is seeking approval. Disco Group has 90 percent confidence it has found an SA-12 surface-to-air missile system and includes a photo for confirmation. The human clicks on the strike plan for more details. Beneath the picture of the SA-12 is a small diagram showing estimated collateral damage. A brown splotch surrounds the target, showing potential damage to anything in the vicinity. Just outside of the splotch is a hospital, but it is outside of the anticipated area of collateral damage. The human clicks "Yes" to approve the engagement. In this video, a human is clearly in the loop. Many tasks are automated, but a human approves each specific engagement.

In other public information, however, CODE seems to leave the door open to removing the human from the loop. A different video shows two teams of air

vehicles, Team A and Team B, sent to engage a surface-to-air missile. As in the LRASM video, the specific target is identified by a human ahead of time, who then launches the missiles to take it out. Similar to LRASM, the air vehicles maneuver around pop-up threats, although this time the air vehicles work cooperatively, sharing navigation and sensor data while in flight. As they maneuver to their target, something unexpected happens: a "critical pop-up target" emerges. It isn't their primary target, but destroying it is a high priority. Team A reprioritizes to engage the pop-up target while Team B continues to the primary target. The video makes clear this occurs under the supervision of the human commander. This implies a different type of human-machine relationship, though, than the earlier CODE video. In this one, instead of the human being *in* the loop, the human is *on* the loop, at least for pop-up threats. For their primary target, they operate in a semiautonomous fashion. The human chose the primary target. But when a pop-up threat emerges, the missiles have the authority to operate as supervised autonomous weapons. They don't need to ask additional permission to take out the target. Like a quarterback calling an audible at the scrimmage line to adapt to the defense, they have the freedom to adapt to unexpected situations that arise. The human operator is like the coach standing on the sidelines—able to call a time-out to intervene, but otherwise merely supervising the action.

DARPA's description of CODE online seems to show a similar flexibility for whether the human or air vehicles themselves approve targets. The CODE website says: "Using collaborative autonomy, CODE-enabled unmanned aircraft would find targets and engage them as appropriate under established rules of engagement . . . and adapt to dynamic situations such as . . . the emergence of unanticipated threats." This appears to leave the door open to autonomous weapons that would find and engage targets on their own.

The detailed technical description issued to developers provides additional information, but little clarity. DARPA explains that developers should:

> Provide a concise but comprehensive targeting chipset so the mission commander can exercise appropriate levels of human judgment over the use of force or evaluate other options.

The specific wording used, "appropriate levels of human judgment," may sound vague and squishy, but it isn't accidental. This guidance directly quotes the official DoD policy on autonomy in weapons, DoD Directive 3000.09, which states:

> Autonomous and semiautonomous weapon systems shall be designed to allow commanders and operators to exercise appropriate levels of human judgment over the use of force.

Notably, that policy does not prohibit autonomous weapons. “Appropriate levels of human judgment” could include autonomous weapons. In fact, the DoD policy includes a path through which developers could seek approval to build and deploy autonomous weapons, with appropriate safeguards and testing, should they be desired.

At a minimum, then, CODE would seem to allow for the possibility of autonomous weapons. The aim of the project is not to build autonomous weapons necessarily. The aim is to enable collaborative autonomy. But in a contested electromagnetic environment where communications links to the human supervisor might be jammed, the program appears to allow for the possibility that the drones could be delegated the authority to engage pop-up threats on their own.

In fact, CODE even hints at one way that collaborative autonomy might aid in target identification. Program documents list one of the advantages of collaboration as “providing multi-modal sensors and diverse observation angles to improve target identification.” Historically, automatic target recognition (ATR) algorithms have not been good enough to trust with autonomous engagements. This poor quality of ATR algorithms could be compensated for by bringing together multiple different sensors to improve the confidence in target identification or by viewing a target from multiple angles, building a more complete picture. One of the CODE videos actually shows this, with air vehicles viewing the target from multiple directions and sharing data. Whether target identification could be improved enough to allow for autonomous engagements is unclear, but if CODE is successful, DoD will have to confront the question of whether to authorize autonomous weapons.

THE DEPARTMENT OF MAD SCIENTISTS

At the heart of many of these projects is the Defense Advanced Research Projects Agency (DARPA), or what writer Michael Belfiore called “the Department of Mad Scientists.” DARPA, originally called ARPA, the Advanced Research Projects Agency, was founded in 1958 by President Eisenhower in response to Sputnik. DARPA’s mission is to prevent “strategic surprise.” The United States was surprised and shaken by the Soviet Union’s launch of Sputnik. The small metal ball hurdling through space overhead was a wake-up call to the reality that the Soviet Union could now launch intercontinental ballistic missiles that could hit anywhere in the United States. In response, President Eisenhower created two organizations to develop breakthrough technologies, the National

Aeronautics and Space Administration (NASA) and ARPA. While NASA had the mission of winning the space race, ARPA had a more fundamental mission of investing in high-risk, high-reward technologies so the United States would never again be surprised by a competitor.

To achieve its mission, DARPA has a unique culture and organization distinct from the rest of the military-industrial complex. DARPA only invests in projects that are "DARPA hard," challenging technology problems that others might deem impossible. Sometimes, these bets don't pan out. DARPA has a mantra of "fail fast" so that if projects fail, they do so before investing massive resources. Sometimes, however, these investments in game-changing technologies pay huge dividends. Over the past five decades, DARPA has time and again laid the seeds for disruptive technologies that have given the United States decisive advantages. Out of ARPA came ARPANET, an early computer network that later developed into the internet. DARPA helped develop basic technologies that underpin the global positioning system (GPS). DARPA funded the first-ever stealth combat aircraft, HAVE Blue, which led to the F-117 stealth fighter. And DARPA has consistently advanced the horizons of artificial intelligence and robotics.

DARPA rarely builds completed weapon systems. Its projects are small, focused efforts to solve extremely hard problems, such as CODE's efforts to get air vehicles to collaborate autonomously. Stuart Russell said that he found these projects concerning because, from his perspective, they seemed to indicate that the United States was expecting to be in a position to deploy autonomous weapons at a future date. Was that, in fact, their intention, or was that simply an inevitability of the technology? If projects like CODE were successful, did DARPA intend to turn the key to full auto or was the intention to always keep a human in the loop?

It was clear that if I was going to understand the future of autonomous weapons, I would need to talk to DARPA.

5

INSIDE THE PUZZLE PALACE

IS THE PENTAGON BUILDING AUTONOMOUS WEAPONS?

DARPA sits in a nondescript office building in Ballston, Virginia, just a few miles from the Pentagon. From the outside, it doesn't look like a "Department of Mad Scientists." It looks like just another glass office building, with no hint of the wild-eyed ideas bubbling inside.

Once you're inside DARPA's spacious lobby, the organization's gravitas takes hold. Above the visitors' desk on the marble wall, raised metal letters that are both simple and futuristic announce: DEFENSE ADVANCED RESEARCH PROJECTS AGENCY. Nothing else. No motto or logo or shield. The organization's confidence is apparent. The words seem to say, "the future is being made here."

As I wait in the lobby, I watch a wall of video monitors announce DARPA's latest project to go public: the awkwardly named Anti-Submarine Warfare (ASW) Continuous Trail Unmanned Vessel (ACTUV). The ship's christened name, Sea Hunter, is catchier. The project is classic DARPA—not only game-changing, but paradigm-bending: the Sea Hunter is an entirely unmanned ship. Sleek and angular, it looks like something time-warped in from the future. With a long, narrow hull and two outriggers, the Sea Hunter carves the oceans like a three-pointed dagger, tracking enemy submarines. At the ship's christening, Deputy Secretary of Defense Bob Work compared it to a Klingon Bird of Prey from *Star Trek*.

There are no weapons on board the Sea Hunter, for now. There should be no mistake, however: the Sea Hunter is a warship. Work called it a "fighting ship,"

part of the Navy's future "human machine collaborative battle fleet." At $2 million apiece, the Sea Hunter is a fraction of the cost of a new $1.6-billion Arleigh Burke destroyer. The low price allows the Navy to purchase scores of the sub-hunting ships on the cheap. Work laid out his vision for flotillas of Sea Hunters roaming the seas:

> You can imagine anti-submarine warfare pickets, you can imagine anti-submarine warfare wolfpacks, you can imagine mine warfare flotillas, you can imagine distributive anti-surface warfare surface action groups . . . We might be able to put a six pack or a four pack of missiles on them. Now imagine 50 of these distributed and operating together under the hands of a flotilla commander, and this is really something.

Like many other robotic systems, the Sea Hunter can navigate autonomously and might someday be armed. There is no indication that DoD has any intention of authorizing autonomous weapons engagements. Nevertheless, the video on DARPA's lobby wall is a reminder that the robotics revolution continues at a breakneck pace.

BEHIND THE CURTAIN: INSIDE DARPA'S TACTICAL TECHNOLOGY OFFICE

DARPA is organized into six departments focusing on different technology areas: biology, information science, microelectronics, basic sciences, strategic technologies, and tactical technologies. CODE, FLA, LRASM, and the Sea Hunter fall into DARPA's Tactical Technology Office (TTO), the division that builds experimental vehicles, ships, airplanes, and spacecraft. Other TTO projects include the XS-1 Experimental Spaceplane, designed to fly to the edge of space and back; the Blue Wolf undersea robotic vehicle; an R2-D2-like robotic copilot for aircraft called ALIAS; the Mach 20 Falcon Hypersonic Technology Vehicle, which flies fast enough to zip from New York to Los Angeles in 12 minutes; and the Vulture program to build an ultra-long endurance drone that can stay in the air for up to five years without refueling. Mad science, indeed.

TTO's offices look like a child's dream toy room. Littered around the offices are models and even some actual prototype pieces of hardware from past TTO projects—missiles, robots, and stealth aircraft. I can't help but wonder what TTO is building today that will be the stealth of tomorrow.

Bradford Tousley, TTO's director, graciously agreed to meet with me to discuss CODE and other projects. Tousley began his government career as an

Army armor officer during the Cold War. His first tour was in an armored cavalry unit on the German border, being ready for a Soviet invasion that might kick off World War III. Later in his career, when the Army sent him back for a secondary education, Tousley earned a doctorate in electrical engineering. His career shifted from frontline combat units to research and development in lasers and optics, working to ensure the U.S. military had the best possible technology. Tousley's career has covered multiple stints at DARPA as well as time in the intelligence community on classified satellite payloads, so he has a breadth of understanding in technology beyond merely robotics.

Tousley pointed out that DARPA was founded in response to the strategic surprise of Sputnik: "DARPA's fundamental mission is unchanged: Enabling pivotal early investments for breakthrough capabilities for national security to achieve or prevent strategic surprise." Inside DARPA, they weigh these questions heavily. "Within the agency, we talk about every single program we begin and we have spirited discussions. We talk about the pros and cons. Why? Why not? . . . How far are we willing to go?" Tousley made clear, however, that answering those questions isn't DARPA's job. "Those are fundamental policy and concept and military employment considerations" for others to decide. "Our fundamental job is to take that technical question off the table. It's our job to make the investments to show the capabilities can exist" to give the warfighter options. In other words, to prevent another Sputnik.

If machines improved enough to reliably take out targets on their own, what the role was for humans in warfare? Despite his willingness to push the boundaries of technology, Tousley still saw humans in command of the mission: "That final decision is with humans, period." That might not mean requiring human authorization for every single target, but autonomous weapons would still operate under human direction, hunting and attacking targets at the direction of a human commander. At least for the foreseeable future, Tousley explained, humans were better than machines at identifying anomalies and reacting to unforeseen events. This meant that keeping humans involved at the mission level was critical to understand the broader context and make decisions. "Until the machine processors equal or surpass humans at making abstract decisions, there's always going to be mission command. There's always going to be humans in the loop, on the loop—whatever you want to call it."

Tousley painted a picture for me of what this might look like in a future conflict: "Groups of platforms that are unmanned that you are willing to attrit [accept some losses] may do extremely well in an anti-access air defense environment . . . How do I take those platforms and a bunch of others and knit

them together in architectures that have manned and unmanned systems striking targets in a congested and contested environment? You need that knitted system because you're going to be GPS-jammed; communications are going to be going in and out; you're going to have air defenses shooting down assets, manned and unmanned. In order to get in and strike critical targets, to control that [anti-access] environment, you're going to have to have a system-of-systems architecture that takes advantage of manned and unmanned systems at different ranges with some amount of fidelity in the ability of the munition by itself to identify the target—could be electronically, could be optically, could be infrared, could be [signals intelligence], could be different ways to identify the target. So that system-of-systems architecture is going to be necessary to knit it all together."

Militaries especially need autonomy in electronic warfare. "We're using physical machines and electronics, and the electronics themselves are becoming machines that operate at machine speed. . . . I need the cognitive electronic warfare to adapt in microseconds. . . . If I have radars trying to jam other radars but they're frequency hopping [rapidly changing radio frequencies] back and forth, I've got to track with it. So [DARPA's Microsystems Technology Office] is thinking about, how do I operate at machine speed to allow these machines to conduct their functions?"

Tousley compared the challenge of cognitive electronic warfare to Google's *go*-playing AlphaGo program. What happens when that program plays another version of AlphaGo at "machine speed?" He explained, "As humans ascend to the higher-level mission command and I've got machines doing more of that targeting function, those machines are going to be challenged by machines on the adversary's side and a human can't respond to that. It's got to be machines responding to machines. . . . That's one of the trends of the Third Offset, that machine on machine." Humans, therefore, shift into a "monitoring" role, watching these systems and intervening, if necessary. In fact, Tousley argues that a difficult question will be whether humans *should* intervene in these machine-on-machine contests, particularly in cyberspace and electronic warfare where the pace of interactions will far exceed human reaction times.

I pointed out that having a human involved in a monitoring role still implies some degree of connectivity, which might be difficult in a contested environment with jamming. Tousley was unconcerned. "We expect that there will be jamming and communications denial going on, but it won't be necessarily everywhere, all the time," he said. "It's one thing to jam my communication link over 1,000 miles, it's another thing to jam two missiles that

are talking in flight that may be three hundred meters apart flying in formation." Reliable communications in contested areas, even short range, would still permit a human being to be involved, at least in some capacity.

So, what role would that person play? Would this person need to authorize every target before engagement, or would human control sit at a higher level? "I think that will be a rule of engagement-dependent decision," Tousley said. "In an extremely hot peer-on-peer conflict, the rules of engagement may be more relaxed. . . . If things are really hot and heavy, you're going to rely on the fact that you built some of that autonomous capability in there." Still, even in this intense battlefield environment, he attested, the human plays the important role of overseeing the combat action. "But you still want some low data rate" to keep a person involved.

It took me a while to realize that Tousley wasn't shrugging off my questions about whether the human would be required to authorize each target because he was being evasive or trying to conceal a secret program, it was because he genuinely didn't see the issue the same way. Automation had been increasing in weapons for decades—from Tousley's perspective, programs like CODE were merely the next step. Humans would remain involved in lethal decision-making, albeit at a higher level overseeing and directing the combat action. The precise details of how much freedom an autonomous system might be granted to choose its own targets and in which situations wasn't his primary concern. Those were questions for military commanders to address. His job as a researcher was to, as he put it, "take that technical question off the table." His job was to build the options. That meant building swarms of autonomous systems that could go into a contested area and conduct a mission with as minimal human supervision as possible. It also meant building in resilient communications so that humans could have as much bandwidth and connectivity to oversee and direct the autonomous systems as possible. How exactly those technologies were implemented—which specific decisions were retained for the human and which were delegated to the machine—wasn't his call to make.

Tousley acknowledged that delegating lethal decision-making came with risks. "If [CODE] enables software that can enable a swarm to execute a mission, would that same swarm be able to execute a mission against the wrong target? Yeah, that is a possibility. We don't want that to happen. We want to build in all the fail-safe systems possible." For this reason, his number-one concern with autonomous systems was actually test and evaluation: "What I worry about the most is our ability to effectively test these systems to the point that we can quantify that we trust them." Trust is essential to commanders being

willing to employ autonomous systems. “Unless the combatant commander feels that that autonomous system is going to execute the mission with the trust that he or she expects, they’ll never deploy it in the first place.” Establishing that trust was all about test and evaluation, which could mean putting an autonomous system through millions of computer simulations to test its behavior. Even still, testing all of the possible situations an autonomous system might encounter and its potential behaviors in response could be very difficult. “One of the concerns I have,” he said, “is that the technology for autonomy and the technology for human-machine integration and understanding is going too far surpass our ability to test it. . . . That worries me.”

TARGET RECOGNITION AND ADAPTION IN CONTESTED ENVIRONMENTS (TRACE)

Tousley declined to comment on another DARPA program, Target Recognition and Adaption in Contested Environments (TRACE), because it fell under a different department he wasn’t responsible for. And although DARPA was incredibly open and helpful throughout the research for this book, the agency declined to comment on TRACE beyond publicly available information. If there’s one program that seems to be a linchpin for enabling autonomous weapons, it’s TRACE. The CODE project aims to compensate for poor automatic target recognition (ATR) algorithms by leveraging cooperative autonomy. TRACE aims to improve ATR algorithms directly.

TRACE’s project description explains the problem:

> In a target-dense environment, the adversary has the advantage of using sophisticated decoys and background traffic to degrade the effectiveness of existing automatic target recognition (ATR) solutions. . . . the false-alarm rate of both human and machine-based radar image recognition is unacceptably high. Existing ATR algorithms also require impractically large computing resources for airborne applications.

TRACE’s aim is to overcome these problems and “develop algorithms and techniques that rapidly and accurately identify military targets using radar sensors on manned and unmanned tactical platforms.” In short, TRACE’s goal is to solve the ATR problem.

To understand just how difficult ATR is—and how game-changing TRACE would be if successful—a brief survey of sensing technologies is in order. Broadly speaking, military targets can be grouped into two categories: “cooperative” and “non-cooperative” targets. Cooperative targets are those that

are actively emitting a signal, which makes them easier to detect. For example, radars, when turned on, emit energy in the electromagnetic spectrum. Radars "see" by observing the reflected energy from their signal. This also means the radar is broadcasting its own position, however. Enemies looking to target and destroy the radar can simply home in on the source of the electromagnetic energy. This is how simple autonomous weapons like the Harpy find radars. They can use passive sensors to simply wait and listen for the cooperative target (the enemy radar) to broadcast its position, and then home in on the signal to destroy the radar.

Non-cooperative targets are those that aren't broadcasting their location. Examples of non-cooperative targets could be ships, radars, or aircraft operating with their radars turned off; submarines running silently; or ground vehicles such as tanks, artillery, or mobile missile launchers. To find non-cooperative targets, active sensors are needed to send signals out into the environment to find targets. Radar and sonar are examples of active sensors; radar sends out electromagnetic energy and sonar sends out sound waves. Active sensors then observe the reflected energy and attempt to discern potential targets from the random noise of background clutter in the environment. Radar "sees" reflected electromagnetic energy and sonar "hears" reflected sound waves.

Militaries are therefore like two adversaries stumbling around in the dark, each listening and peering fervently into the darkness to hear and see the other while remaining hidden themselves. Our eyes are passive sensors; they simply receive light. In the darkness, however, an external source of light like a flashlight is needed. Using a flashlight gives away one's own position, though, making one a "cooperative target" for the enemy. In this contest of hiding and finding, zeroing in on the enemy's cooperative targets is like finding a person waving a flashlight around in the darkness. It isn't hard; the person waving the flashlight is going to stand out. Finding the non-cooperative targets who keep their flashlights turned off can be very, very tricky.

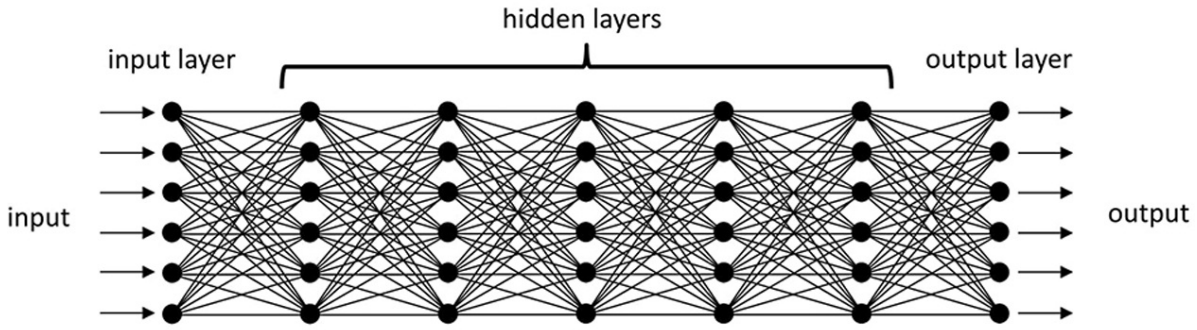
When there is little background clutter, objects can be found relatively easily through active sensing. Ships and aircraft stand out easily against their background—a flat ocean and an empty sky. They stand out like a person standing in an open field. A quick scan with even a dim light will pick out a person standing in the open, although discerning friend from foe can be difficult. In cluttered environments, however, even finding targets in the first place can be hard. Moving targets can be discerned via Doppler shifting—essentially the same method that police use to detect speeding vehicles. Moving objects shift the frequency of the return radar signal, making them stand out against a

stationary background. Stationary targets in cluttered environments can be as hard to see as a deer hiding in the woods, though. Even with a light shined directly on them, they might not be noticed.

Humans have challenges seeing stationary, camouflaged objects and human visual cognitive processing is incredibly complex. We take for granted how computationally difficult it is to see objects that blend into the background. While radars and sonars can “see” and “hear” in frequencies that humans are incapable of, military ATR is nowhere near as good as humans at identifying objects amid clutter.

Militaries currently sense many non-cooperative targets using a technique called synthetic aperture radar, or SAR. A vehicle, typically an aircraft, flies in a line past a target and sends out a burst of radar pulses as the aircraft moves. This allows the aircraft to create the same effect as having an array of sensors, a powerful technique that enhances image resolution. The result is sometimes grainy images composed of small dots, like a black-and-white pointillist painting. While SAR images are generally not as sharp as images from electro-optical or infrared cameras, SAR is a powerful tool because radar can penetrate through clouds, allowing all-weather surveillance. Building algorithms that can automatically identify SAR images is extremely difficult, however. Grainy SAR images of tanks, artillery, or airplanes parked on a runway often push the limits of human abilities to recognize objects, and historically ATR algorithms have fallen far short of human abilities.

The poor performance of military ATR stands in stark contrast to recent advances in computer vision. Artificial intelligence has historically struggled with object recognition and perception, but the field has seen rapid gains recently due to deep learning. Deep learning uses neural networks, a type of AI approach that is analogous to biological neurons in animal brains. Artificial neural networks don’t directly mimic biology, but are inspired by it. Rather than follow a script of *if-then* steps for how to perform a task, neural networks work based on the strength of connections within a network. Thousands or even millions of data samples are fed into the network and the weights of various connections between nodes in the network are constantly adjusted to “train” the network on the data. In this way, neural networks “learn.” Network settings are refined until the correct output, such as the correct image category (for example, cat, lamp, car) is achieved.



A deep neural network has hidden layers between the input and output layers.
Some deep neural networks can have as many as 150 or more hidden layers.

Deep Neural Network

Deep neural networks are those that have multiple "hidden" layers between the input and output, and have proven to be a very powerful tool for machine learning. Adding more layers in the network between the input data and output allows for a much greater complexity of the network, enabling the network to handle more complex tasks. Some deep neural nets have over a hundred layers.

This complexity is, it turns out, essential for image recognition, and deep neural nets have made tremendous progress. In 2015, a team of researchers from Microsoft announced that they had created a deep neural network that for the first time surpassed human performance in visual object identification. Using a standard test dataset of 150,000 images, Microsoft's network achieved an error rate of only 4.94 percent, narrowly edging out humans, who have an estimated 5.1 percent error rate. A few months later, they improved on their own performance with a 3.57 percent rate by a 152-layer neural net.

TRACE intends to harness these advances and others in machine learning to build better ATR algorithms. ATR algorithms that performed on par with or better than humans in identifying non-cooperative targets such as tanks, mobile missile launchers, or artillery would be a game changer in terms of finding and destroying enemy targets. If the resulting target recognition system was of sufficiently low power to be located on board the missile or drone itself, human authorization would not be required, at least from a purely technical point of view. The technology would enable weapons to hunt and destroy targets all on their own.

Regardless of whether DARPA was intending to build autonomous weapons, it was clear that programs like CODE and TRACE were putting in place the building blocks that would enable them in the future. Tousley's view was that it

wasn’t DARPA’s call whether to authorize that next fateful step across the line to weapons that would choose their own targets. But if it wasn’t DARPA’s call whether to build autonomous weapons, then whose call was it?

6

CROSSING THE THRESHOLD

APPROVING AUTONOMOUS WEAPONS

The Department of Defense has an official policy on the role of autonomy in weapons, DoD Directive 3000.09, "Autonomy in Weapon Systems." (Disclosure: While at DoD, I led the working group that drafted the policy.) Signed in November 2012, the directive is published online so anyone can read it.

The directive includes some general language on principles for design of semiautonomous and autonomous systems, such as realistic test and evaluation and understandable human-machine interfaces. The meat of the policy, however, is the delineation of three classes of systems that get the "green light" for approval in the policy. These are: (1) semiautonomous weapons, such as homing munitions; (2) defensive supervised autonomous weapons, such as the ship-based Aegis weapon system; and (3) nonlethal, nonkinetic autonomous weapons, such as electronic warfare to jam enemy radars. These three types of autonomous systems are in wide use today. The policy essentially says to developers, "If you want to build a weapon that uses autonomy in ways consistent with existing practices, you're free to do so." Normal acquisition rules apply, but those types of systems do not require any additional approval.

Any future weapon system that would use autonomy in a novel way outside of those three categories gets a "yellow light." Those systems need to be reviewed before beginning formal development (essentially the point at which large sums of money would be spent) and again before fielding. The policy outlines who participates in the review process—the senior defense civilian

officials for policy and acquisitions and the chairman of the Joint Chiefs of Staff—as well as the criteria for review. The criteria are lengthy, but predominantly focus on test and evaluation for autonomous systems to ensure they behave as intended—the same concern Tousley expressed. The stated purpose of the policy is to "minimize the probability and consequences of failures in autonomous and semiautonomous weapon systems that could lead to unintended engagements." In other words, to minimize the chances of armed robots running amok.

Lethal autonomous weapons are not prohibited by the policy directive. Instead, the policy provides a process by which new uses of autonomy could be reviewed by relevant officials before deployment. The policy helps ensure that if DoD were to build autonomous weapons that they weren't developed and deployed without sufficient oversight, but it doesn't help answer the question of whether DoD might actually approve such systems. On that question, the policy is silent. All the policy says is that if an autonomous weapon met all of the criteria, such as reliability under realistic conditions, then in principle it could be authorized.

GIVING THE GREEN LIGHT TO AUTONOMOUS WEAPONS

But *would* it be authorized? DARPA programs are intended to explore the art of the possible, but that doesn't mean that DoD would necessarily turn those experimental projects into operational weapon systems. To better understand whether the Pentagon might actually approve autonomous weapons, I sat down with then-Pentagon acquisition chief, Under Secretary of Defense Frank Kendall. As the under secretary of defense for acquisition, technology and logistics, Kendall was the Pentagon's chief technologist and weapons buyer under the Obama Administration. When it came to major weapons systems like the X-47B or LRASM, the decision whether or not to move forward was in Kendall's hands. In the process laid out under the DoD Directive, Kendall was one of three senior officials, along with the under secretary for policy and the chairman of the Joint Chiefs, who all had to agree in order to authorize developing an autonomous weapon.

Kendall has a unique background among defense technologists. In addition to a distinguished career across the defense technology enterprise, serving in a variety of roles from vice president of a major defense firm to several mid-level bureaucratic jobs within DoD, Kendall also has worked pro bono as a human rights lawyer. He has worked with Amnesty International, Human Rights First,

and other human rights groups, including as an observer at the U.S. prison at Guantánamo Bay. Given his background, I was hopeful that Kendall might be able to bridge the gap between technology and policy.

Kendall made clear, for starters, that there had never been a weapon autonomous enough even to trigger the policy review. “We haven’t had anything that was even remotely close to autonomously lethal.” If he were put in that position, Kendall said his chief concerns would be ensuring that it complied with the laws of war and that the weapon allowed for “appropriate human judgment,” a phrase that appears in the policy directive. Kendall admitted those terms weren’t defined, but conversation with him began to elucidate his thinking.

Kendall started his career as an Army air defender during the Cold War, where he learned the value of automation first hand. “We had an automatic mode for the Hawk system that we never used, but I could see in an extreme situation where you’d turn it on, because you just couldn’t do things fast enough otherwise,” he said. When you have “fractions of a second” to decide—that’s a role for machines.

Kendall said that automatic target recognition and machine learning were improving rapidly. As they improve, it should become possible for the machine to select its own targets for engagement. In some settings, such as taking out an enemy radar, he thought it could be done “relatively soon.”

This raises tricky questions. “Where do you want the human intervention to be?” he asked. “Do you want it to be the actual act of employing the lethality? Do you want it to be the acceptance of the rules that you set for identifying something as hostile?” Kendall didn’t have the answers. “I think we’re going to have to sort through all that.”

One important factor was the context. “Are you just driving down the street or are you actually in a war, or you’re in an insurgency? The context matters.” In some settings, using autonomy to select and engage targets might be appropriate. In others, it might not.

Kendall saw using an autonomous weapon to target enemy radars as fairly straightforward and something he didn’t see many people objecting to. There were other examples that pushed the boundaries. Kendall said on a trip to Israel, his hosts from the Israel Defense Forces had him sit in a Merkava tank that was outfitted with the Trophy active protection system. The Israelis fired a rocket propelled grenade near the tank (“offset a few meters,” he said) and the Trophy system intercepted it automatically. “But suppose I also wanted to shoot back at . . . wherever the bullet had come from?” he asked. “You can automate that, right? That’s protecting me, but it’s the use of that weapon in a way which could be

lethal to whoever, you know, was in the line of fire when I fire." He pointed out that automating a return-fire response might prevent a second shot, saving lives. Kendall acknowledged that had risks, but there were risks in not doing it as well. "How much do we want to put our own people at risk by not allowing them to use this technology? That's the other side of the equation."

Things become especially difficult if the machine is better than the person, which, at some point, will happen. "I think at that point, we'll have a tough decision to make as to how we want to go with that." Kendall saw value in keeping a human in the loop as a backup, but, "What if it's a situation where there isn't that time? Then aren't you better off to let the machine do it? You know, I think that's a reasonable question to ask."

I asked him for his answer to the question—after all, he was the person who would decide in DoD. But he didn't know.

"I don't think we've decided that yet," he said. "I think that's a question we'll have to confront when we get to where technology supports it."

Kendall wasn't worried, though. "I think we're a long way away from the Terminator idea, the killer robots let loose on the battlefield idea. I don't think we're anywhere near that and I don't worry too much about that." Kendall expressed confidence in how the United States would address this technology. "I'm in my job because I find my job compatible with being a human rights lawyer. I think the United States is a country which has high values and it operates consistent with those values. . . . I'm confident that whatever we do, we're going to start from the premise that we're going to follow the laws of war and obey them and we're going to follow humanitarian principles and obey them."

Kendall was worried about other countries, but he was most concerned about what terrorists might do with commercially available technology. "Automation and artificial intelligence are one of the areas where the commercial developments I think dwarf the military investments in R&D. They're creating capabilities that can easily be picked up and applied for military purposes." As one example, he asked, "When [ISIS] doesn't have to put a person in that car and can just send it out on its own, that's a problem for us, right?"

THE REVOLUTIONARY

Kendall's boss was Deputy Secretary of Defense Bob Work, the Pentagon's number-two bureaucrat—and DoD's number-one robot evangelist. As deputy secretary from 2014–17, Work was the driving force behind the Pentagon's

Third Offset Strategy and its focus on human-machine teaming. In his vision of future conflicts, AI will work in concert with humans in human-machine teams. This blended human-plus-machine approach could take many forms. Humans could be enhanced through exoskeleton suits and augmented reality, enabled by machine intelligence. AI systems could help humans make decisions, much like in "centaur chess," where humans are assisted by chess programs that analyze possible moves. In some cases, AI systems may perform tasks on their own with human oversight, particularly when speed is an advantage, similar to automated stock trading. Future weapons will be more intelligent and cooperative, swarming adversaries.

Collectively, Work argues these advances may lead to a "revolution" in warfare. Revolutions in warfare, Work explained in a 2014 monograph, are "periods of sharp, discontinuous change [in which] . . . existing military regimes are often upended by new more dominant ones, leaving old ways of warfare behind."

In defense circles, this is a bold claim. The U.S. defense community of the late 1990s and early 2000s became enamored with the potential of information technology to lead to a revolution in warfare. Visions of "information dominance" and "network-centric warfare" foundered in the mountains of Afghanistan and the dusty streets of Iraq as the United States became mired in messy counterinsurgency wars. High-tech investments in next-generation weapon systems such as F-22 fighter jets were overpriced or simply irrelevant for finding and tracking insurgents or winning the hearts and minds of civilian populations. And yet . . .

The information revolution continued, leading to more advanced computer processors and ever more sophisticated machine intelligence. And even while warfare in the information age might not have unfolded the way Pentagon futurists might have envisioned, the reality is information technology dramatically shaped how the United States fought its counterinsurgency wars. Information became the dominant driver of counternetwork operations as the United States sought to find insurgents hiding among civilians, like finding a needle in a stack of needles.

Sweeping technological changes like the industrial revolution or the information revolution unfold in stages over time, over the course of decades or generations. As they do, they inevitably have profound effects on warfare. Technologies like the internal-combustion engine that powered civilian automobiles and airplanes in the industrial revolution led to tanks and military aircraft. Tanks and airplanes, along with other industrial-age weaponry such as

machine guns, profoundly changed World War I and World War II.

Work is steeped in military history and a student of Pentagon futurist Andy Marshall, who for decades ran DoD's Office of Net Assessment and championed the idea that another revolution in warfare was unfolding today. Work understands the consequences of falling behind during periods of revolutionary change. Militaries can lose battles and even wars. Empires can fall, never to recover. In 1588, the mighty Spanish Armada was defeated by the British, who had more expertly exploited the revolutionary technology of the day: cannons. In the interwar period between World War I and World War II, Germany was more successful in capitalizing on innovations in aircraft, tanks, and radio technology and the result was the blitzkrieg—and the fall of France. The battlefield is an unforgiving environment. When new technologies upend old ways of fighting, militaries and nations don't often get second chances to get it right.

If Work is right, and a revolution in warfare is under way driven in part by machine intelligence, then there is an imperative to invest heavily in AI, robotics, and automation. The consequences of falling behind could be disastrous for the United States. The industrial revolution led to machines that were stronger than humans, and the victors were those who best capitalized on that technology. Today's information revolution is leading to machines that are smarter and faster than humans. Tomorrow's victors will be those who best exploit AI.

Right now, AI systems can outperform humans in narrow tasks but still fall short of humans in general intelligence, which is why Work advocates human-machine teaming. Such teaming allows the best of both human and machine intelligence. AI systems can be used for specific, tailored tasks and for their advantages in speed while humans can understand the broader context and adapt to novel situations. There are limitations to this approach. In situations where the advantages in speed are overwhelming, delegating authority entirely to the machine is preferable.

When it comes to lethal force, in a March 2016 interview, Work stated, "We will not delegate lethal authority for a machine to make a decision." He quickly caveated that statement a moment later, however, adding, "The only time we will . . . delegate a machine authority is in things that go faster than human reaction time, like cyber or electronic warfare."

In other words, we won't delegate lethal authority to a machine . . . unless we have to. In the same interview, Work said, "We might be going up against a competitor that is more willing to delegate authority to machines than we are and as that competition unfolds, we'll have to make decisions about how to

compete." How long before the tightening spiral of an ever-faster OODA loop forces that decision? Perhaps not long. A few weeks later in another interview, Work stated it was his belief that "within the next decade or decade and a half it's going to become clear when and where we delegate authority to machines." A principal concern of his was the fact that while in the United States we debate the "moral, political, legal, ethical" issues surrounding lethal autonomous weapons, "our potential competitors may not."

There was no question that if I was going to understand where the robotics revolution was heading, I needed to speak to Work. No single individual had more sway over the course of the U.S. military's investments in autonomy than he did, both by virtue of his official position in the bureaucracy as well as his unofficial position as the chief thought-leader on autonomy. Work may not be an engineer writing the code for the next generation of robotic systems, but his influence was even broader and deeper. Through his public statements and internal policies, Work was shaping the course of DoD's investments, big and small. He had championed the concept of human-machine teaming. How he framed the technology would influence what engineers across the defense enterprise chose to build. Work immediately agreed to an interview.

THE FUTURE OF LETHAL AUTONOMY

The Pentagon is an imposing structure. At 6.5 million square feet, it is one of the largest buildings in the world. Over 20,000 people enter the Pentagon every day to go to work. As I moved through the sea of visitors clearing security, I was reminded of the ubiquity of the robotics revolution. I heard the man in line behind me explain to Pentagon security that the mysterious item in his briefcase raising alarms in their x-ray scanners was a drone. "It's a UAV," he said. "A drone. I have clearance to bring it in," he added hastily.

The drones are literally everywhere, it would seem.

Work's office was in the famed E-ring where the Pentagon's top executives reside, and he was kind enough to take time out of his busy schedule to talk with me. I started with a simple question, one I had been searching to answer in vain in my research: Is the Department of Defense building autonomous weapons?

Underscoring the definitional problem, Work wanted to clarify what I meant by "autonomous weapon" before answering. I explained I was defining an autonomous weapon as one that could search for, select, and engage targets on its own. Work replied, "We, the United States, have had a lethal autonomous weapon, using your definition, since 1945: the Bat [radar-guided anti-ship

bomb].” He said, “I would define it as a narrow lethal autonomous weapon in that the original targeting of the Japanese destroyer that we fired at was done by a Navy PBY maritime patrol aircraft . . . they knew [the Japanese destroyer] was hostile—and then they launched the weapon. But the weapon itself made all of the decisions on the final engagement using an S-band radar seeker.” Despite his use of the term “autonomous weapon” to describe a radar-guided homing munition, Work clarified he was comfortable with that use of autonomy. “I see absolutely no problem in those types of weapons. It was targeted on a specific capability by a man in the loop and all the autonomy was designed to do was do the terminal endgame engagement.” He was also comfortable with how autonomy was used in a variety of modern weapons, from torpedoes to the Aegis ship combat system.

Painting a picture of the future, Work said, “We are moving to a world in which the autonomous weapons will have smart decision trees that will be completely preprogrammed by humans and completely targeted by humans. So let’s say we fire a weapon at 150 nautical miles because our off-board sensors say a Russian battalion tactical group is operating in this area. We don’t know exactly what of the battalion tactical group this weapon will kill, but we know that we’re engaging an area where there are hostiles.” Work explained that the missile itself, following its programming logic, might prioritize which targets to strike—tanks, artillery, or infantry fighting vehicles. “We’re going to get to that level. And I see no problem in that,” he said. “There’s a whole variety of autonomous weapons that do endgame engagement decisions after they have been targeted and launched at a specific target or target area.” (Here Work is using “autonomous weapon” to refer to fire-and-forget homing munitions.)

Loitering weapons, Work acknowledged, were qualitatively different. “The thing that people worry about is a weapon we fire at range and it loiters in the area and it decides when, where, how, and what to kill without anything other than the human launching it in the general direction.” Work acknowledged that, regardless of the label used, these loitering munitions were qualitatively different than homing munitions that had to be launched at a specific target. But Work didn’t see any problem with loitering munitions either. “People start to get nervous about that, but again, I don’t worry about that at all.” He said he didn’t believe the United States would ever fire such a weapon into an area unless it had done the appropriate estimates for potential collateral damage. If, on the other hand, “we are relatively certain that there are no friendlies in the area: weapons free. Let the weapon decide.”

These search-and-destroy weapons didn’t bother Work, even if they were

choosing their own targets, because they were still “narrow AI systems.” These weapons would be “programmed for a certain effect against a certain type of target. We can tell them the priorities. We can even delegate authority to the weapon to determine how it executes end game attack.” With these weapons, there may be “a lot of prescribed decision trees, but the human is always firing it into a general area and we will do [collateral damage estimation] and we will say, ‘Can we accept the risk that in this general area the weapon might go after a friendly?’ And we will do the exact same determination that we have right now.”

Work said the key question is, “What is your comfort level on target location error?” He explained, “If you are comfortable firing a weapon into an area in which the target location error is pretty big, you are starting to take more risks that it might go against an asset that might be a friendly asset or an allied asset or something like that. . . . So, really what’s happening is because you can put so much more processing power onto the weapon itself, the [acceptable degree of] target location error is growing. And we will allow the weapon to search that area and figure out the endgame.” An important factor is what else is in the environment and the acceptable level of collateral damage. “If you have real low collateral damage [requirements],” he said, “you’re not going to fire a weapon into an area where the target location is so large that the chances of collateral damage go up.”

In situations where that risk was acceptable, Work saw no problems with such weapons. “I hear people say, ‘This is some terrible thing. We’ve got killer robots.’ No we don’t. Robots . . . will only hit the targets that you program in. . . . The human is still launching the weapon and specifying the type of targets to be engaged, even if the weapon is choosing the specific targets to attack within that wide area. There’s always going to be a man or woman in the loop who’s going to make the targeting decision,” he said, even if that targeting decision was now at a higher level.

Work contrasted these narrow AI systems with artificial general intelligence (AGI), “where the AI is actually making these decisions on its own.” This is where Work would draw the line. “The danger is if you get a general AI system and it can rewrite its own code. That’s the danger. We don’t see ever putting that much AI power into any given weapon. But that would be the danger I think that people are worried about. What happens if Skynet rewrites its own code and says, ‘humans are the enemy now’? But that I think is very, very, very far in the future because general AI hasn’t advanced to that.” Even if technology did get there, Work was not so keen on using it. “We will be extremely careful in trying to put general AI into an autonomous weapon,” he said. “As of this point I can’t

get to a place where we would ever launch a general AI weapon . . . [that] makes all the decisions on its own. That's just not the way that I would ever foresee the United States pursuing this technology. [Our approach] is all about empowering the human and making sure that the humans inside the battle network has tactical and operational overmatch against their enemies."

Work recognized that other countries may use AI technology differently. "People are going to use AI and autonomy in ways that surprise us," he said. Other countries might deploy weapons that "decide who to attack, when to attack, how to attack" all on their own. If they did, then that could change the U.S. calculus. "The only way that we would go down that path, I think, is if it turns out our adversaries do and it turns out that we are at an operational disadvantage because they're operating at machine speed and we're operating at human speeds. And then we might have to rethink our theory of the case." Work said that challenge is something he worries about. "The nature of the competition about how people use AI and autonomy is really going to be something that we cannot control and we cannot totally foresee at this point."

THE PAST AS A GUIDE TO THE FUTURE

Work forthrightly answered every question I put to him, but I still found myself leaving the interview unsatisfied. He had made clear that he was comfortable using narrow AI systems to perform the kinds of tasks we're doing today: endgame autonomy to confirm a target chosen by a human or defensive human-supervised autonomy like at Aegis. He was comfortable with loitering weapons that might operate over a wider area or smarter munitions that could prioritize targets, but he continued to see humans playing a role in launching and directing those weapons. There were some technologies Work wasn't comfortable with—artificial general intelligence or "boot-strapping" systems that could modify their own code. But there was a wide swath of systems in between. What about a uninhabited combat aircraft that made its own targeting decisions? How much target error was acceptable? He simply didn't know. Those are questions future defense leaders would have to address.

To help shed light on how future leaders might answer those questions, I turned to Dr. Larry Schuette, director of research at the Office of Naval Research. Schuette is a career scientist with the Navy and has a doctorate in electrical engineering, so he understands the technology intimately. ONR has repeatedly been at the forefront of advancements in autonomy and robotics, and

Schuette directs much of this research. He is also an avid student of history, so I hoped he could help me understand what the past might tell us about the shape of things to come.

As a researcher, Schuette made it clear to me that autonomous weapons are not an area of focus for ONR. There are a lot of areas where uninhabited and autonomous systems could have value, but his perspective was to focus on the mundane tasks. “I’m always looking for: what’s the easiest thing with the highest return on investment that we could actually go do where people would thank us for doing it. . . . Don’t go after the hard missions. . . . Let’s do the easy stuff first.” Schuette pointed to thankless jobs like tanking aircraft or cleaning up oil spills. “Be the trash barge. . . . The people would love you.” His view was that even tackling these simple, unobjectionable missions was a big enough challenge. “I know that what is simple to imagine in science and technology isn’t as simple to do.”

Schuette also emphasized that he didn’t see a compelling operational need for autonomous weapons. Today’s model of “The man pushes a button and the weapon goes autonomous from there but the man makes the decision” was a “workable framework for some large fraction of what you would want to do with unmanned air, unmanned surface, unmanned underwater, unmanned ground vehicles. . . . I don’t see much need in future warfare to get around that model,” he said.

As a student of history, however, Schuette had a somewhat different perspective. His office looked like a naval museum, with old ship’s logs scattered on the bookshelves and black-and-white photos of naval aviators on the walls. While speaking, Schuette would frequently leap out of his chair to grab a book about unrestricted submarine warfare or the Battle of Guadalcanal to punctuate his point. The historical examples weren’t about autonomy, rather they were about a broader pattern in warfare. “History is full of innovations and asymmetric responses,” he said. In World War II, the Japanese were “amazed” at U.S. skill at naval surface gunfire. In response, they decided to fight at night, resulting in devastating nighttime naval surface action at the Battle of Guadalcanal. The lesson is that “the threat gets a vote.” Citing Japanese innovations in long-range torpedoes, Schuette said, “We had not planned on fighting a torpedo war. . . . The Japanese had a different idea.”

This dynamic of innovation and counter-innovation inevitably leads to surprises in warfare and can often change what militaries see as ethical or appropriate. “We’ve had these debates before about ethical use of X or Y,” Schuette pointed out. He compared today’s debates about autonomous weapons

to debates in the U.S. Navy in the interwar period between World War I and World War II about unrestricted submarine warfare. “We went all of the twenties, all the thirties, talking about how unrestricted submarine warfare was a bad idea we would never do it. And when the shit hit the fan the first thing we did was begin executing unrestricted submarine warfare.” Schuette grabbed a book off his shelf and quoted the order issued to all U.S. Navy ship and submarine commanders on December 7, 1941, just four and a half hours after the attack at Pearl Harbor:

EXECUTE AGAINST JAPAN UNRESTRICTED AIR AND SUBMARINE WARFARE

The lesson from history, Schuette said, was that “we are going to be violently opposed to autonomous robotic hunter-killer systems until we decide we can’t live without them.” When I asked him what he thought would be the decisive factor, he had a simple response: “Is it December eighth or December sixth?”

7

WORLD WAR R

ROBOTIC WEAPONS AROUND THE WORLD

The robotics revolution isn't American-made. It isn't even American-led. Countries around the world are pushing the envelope in autonomy, many further and faster than the United States. Conversations in U.S. research labs and the Pentagon's E-ring are only one factor influencing the future of autonomous weapons. Other nations get a vote too. What they do will influence how the technology develops, proliferates, and how other nations—including the United States—react.

The rapid proliferation of drones portends what is to come for increasingly autonomous systems. Drones have spread to nearly a hundred countries around the globe, as well as non-state groups such as Hamas, Hezbollah, ISIS, and Yemeni Houthi rebels. Armed drones are next. A growing number of countries have armed drones, including nations that are not major military powers such as South Africa, Nigeria, and Iraq.

Armed robots are also proliferating on the ground and at sea. South Korea has deployed a robot sentry gun to its border with North Korea. Israel has sent an armed robotic ground vehicle, the Guardium, on patrol near the Gaza border. Russia is building an array of ground combat robots and has plans for a robot tank. Even Shiite militias in Iraq have gotten in on the game, fielding an armed ground robot in 2015.

Armed Drone Proliferation *As of June 2017, sixteen countries possessed armed drones: China, Egypt, Iran, Iraq, Israel, Jordan, Kazakhstan, Myanmar, Nigeria, Pakistan, Saudi Arabia, Turkey, Turkmenistan, United Arab Emirates, the United Kingdom, and the United States. Some nations developed armed drones indigenously, while others acquired the technology from abroad. Over 90 percent of international armed drone transfers (shown on the map via arrows) have been from China.*

Armed robots are heading to sea as well. Israel has also developed an armed uninhabited boat, the Protector, to patrol its coast. Singapore has purchased the Protector and deployed it for counterpiracy missions in the Straits of Malacca. Even Ecuador has an armed robot boat, the ESGRUM, produced entirely indigenously. Armed with a rifle and rocket launcher, the ESGRUM will patrol Ecuadorian waterways to counter pirates.

As in the United States, the key question will be whether these nations plan to cross the line to full autonomy. No nation has stated they plan to build autonomous weapons. Few have ruled them out either. Only twenty-two nations have said they support a ban on lethal autonomous weapons: Pakistan, Ecuador, Egypt, the Holy See, Cuba, Ghana, Bolivia, Palestine, Zimbabwe, Algeria, Costa Rica, Mexico, Chile, Nicaragua, Panama, Peru, Argentina, Venezuela, Guatemala, Brazil, Iraq, and Uganda (as of November 2017). None of these states are major military powers and some, such as Costa Rica or the Holy See, lack a military entirely.

One of the first areas where countries will be forced to grapple with the

choice of whether to delegate lethal authority to the machine will be for uninhabited combat aircraft designed to operate in contested areas. Several nations are reportedly developing experimental combat drones similar to the X-47B, although for operation from land bases rather than aircraft carriers. These include the United Kingdom's Taranis, China's Sharp Sword, Russia's Skat, France's nEUROn, India's Aura, and a rumored unnamed Israeli stealth drone. Although these drones are likely designed to operate with protected communications links to human controllers, militaries will have to decide what actions they want the drone to carry out if (and when) communications are jammed. Restricting the drone's rules of engagement could mean giving up valuable military advantage, and few nations are being transparent about their plans.

Given that a handful of countries already possess the fully autonomous Harpy, it isn't a stretch to imagine them and others authorizing a similar level of autonomy with a recoverable drone. Whether countries are actually building those weapons today is more difficult to discern. If understanding what's happening inside the U.S. defense industry is difficult, peering behind the curtain of secret military projects around the globe is even harder. Are countries like Russia, China, the United Kingdom, and Israel building autonomous weapons? Or are they still keeping humans in the loop, walking right up to the line of autonomous weapons but not crossing it? Four high-profile international programs, a South Korean robot gun, a British missile, a British drone, and a Russian fleet of armed ground robots, show the difficulty in uncovering what nations around the globe are doing.

THE CURIOUS CASE OF THE AUTONOMOUS SENTRY BOT

South Korea's Samsung SGR-A1 robot is a powerful example of the challenge in discerning how much autonomy weapon systems have. The SGR-A1 is a stationary armed sentry robot designed to defend South Korean's border against North Korea. In 2007, when the robot was revealed, the electrical engineering magazine *IEEE Spectrum* reported it had a fully autonomous mode for engaging targets on its own. In an interview with the magazine, Samsung principal research engineer Myung Ho Yoo said, "the ultimate decision about shooting should be made by a human, not the robot." But the article made clear that Yoo's "should" was not a requirement, and that the robot did have a fully automatic option.

The story was picked up widely, with the SGR-A1 cited as an example of a real-world autonomous weapon by *The Atlantic*, the BBC, NBC, *Popular Science*, and *The Verge*. The SGR-A1 made *Popular Science*'s list of "Scariest Ideas in Science" with PopSci asking, "WHY, GOD? WHY?" Several academic researchers conducting in-depth reports on military robotics similarly cited the SGR-A1 as fully autonomous.

In the face of this negative publicity, Samsung backpedaled, saying that in fact a human *was* required to be in the loop. In 2010, a spokesperson for Samsung clarified that "the robots, while having the capability of automatic surveillance, cannot automatically fire at detected foreign objects or figures." Samsung and the South Korean government have been tight-lipped about details, though, and one can understand why. The SGR-A1 is designed to defend South Korea's demilitarized zone along its border with North Korea, with whom South Korea is technically still at war. Few countries on earth face as immediate and intense a security threat. One million North Korean soldiers and the threat of nuclear weapons loom over South Korea like a menacing shadow. In the same interview in which he asserted a human will always remain in the loop, the Samsung spokesperson asserted, "the SGR-1 can and will prevent wars."

What are the actual specifications and design parameters for the SGR-A1? It's essentially impossible to know without directly inspecting the robot. If Samsung says a human is in the loop, all we can do is take their word for it. If South Korea is willing to delegate more autonomy to their robots than other nations, however, it wouldn't be surprising. Defending the DMZ against North Korea is a matter of survival for South Korea. Accepting the risks of a fully autonomous sentry gun may be more than worth it for South Korea if it enhances deterrence against North Korea.

THE BRIMSTONE MISSILE

Similar to the U.S. LRASM, the United Kingdom's Brimstone missile has come under fire from critics who have questioned whether it has too much autonomy. The Brimstone is an aircraft-launched fire-and-forget missile designed to destroy ground vehicles or small boats. It can accomplish this mission in a variety of ways.

Brimstone has two primary modes of operation: Single Mode and Dual Mode. In Single Mode, a human "paints" the target with a laser and the missile homes in on the laser reflection. The missile will go wherever the human points the laser, allowing the human to provide "guidance all the way to the target."

Dual Mode combines the laser guidance with a millimeter-wave (MMW) radar seeker for "fast moving and maneuvering targets and under narrow Rules of Engagement." The human designates the target with a laser, then there is a "handoff" from the laser to the MMW seeker at the final stage so the weapon can home in on fast moving targets. In both modes of operation, the missile is clearly engaging targets that have been designated by a human, making it a semiautonomous weapon.

However, the developer also advertises another mode of operation, "a previously-developed fire-and-forget, MMW-only mode" that can be enabled "via a software role change." The developer explains:

> This mode provides through-weather targeting, kill box-based discrimination and salvo launch. It is highly effective against multi-target armor formations. Salvo-launched Brimstones self-sort based on firing order, reducing the probability of overkill for increased one-pass lethality.

This targeting mode would allow a human to launch a salvo of Brimstones against a group of enemy tanks, letting the missiles sort out which missiles hit which tank. According to a 2015 *Popular Mechanics* article, in this mode the Brimstone is fairly autonomous:

> It can identify, track, and lock on to vehicles autonomously. A jet can fly over a formation of enemy vehicles and release several Brimstones to find targets in a single pass. The operator sets a "kill box" for Brimstone, so it will only attack within a given area. In one demonstration, three missiles hit three target vehicles while ignoring nearby neutral vehicles.

On the Brimstone's spec sheet, the developer also describes a similar functionality against fast-moving small boats, also called fast inshore attack craft (FIAC):

> In May 2013, multiple Brimstone missiles operating in an autonomous [millimeter] wave (MMW) mode completed the world's first single button, salvo engagement of multiple FIAC, destroying three vessels (one moving) inside a kill box, while causing no damage to nearby neutral vessels.

When operating in MMW-only mode, is the Brimstone an autonomous weapon? While the missile has a reported range in excess of 20 kilometers, it cannot loiter to search for targets. This means that the human operator must know there are valid targets—ground vehicles or small boats—within the kill box before launch in order for the missile to be effective.

The Brimstone can engage these targets using some innovative features. A pilot can launch a salvo of multiple Brimstones against a group of targets within a kill box and the missiles themselves "self-sort based on firing order" to hit different targets. This makes the Brimstone especially useful for defending

against enemy swarm attacks. For example, Iran has harassed U.S. ships with swarming small boats that could overwhelm ship defenses, causing a USS *Cole*–type suicide attack. Navy helicopters armed with Brimstones would be an extremely effective defense against boat swarms, allowing pilots to take out an entire group of enemy ships at once without having to individually target each ship.

Even with all of the Brimstone's features, the human user still needs to launch it at a known group of targets. Because it cannot loiter, if there weren't targets in the kill box when the missile activated its seeker, the missile would be wasted. Unlike a drone, the missile couldn't return to base. The salvo launch capability allows the pilot to launch multiple missiles against a swarm of targets, rather than select each one individually. This makes a salvo of Brimstones similar to the Sensor Fuzed Weapon that is used to take out a column of tanks. Even though the missiles themselves might self-sort which missile hits which target, the human is still deciding to attack that specific cluster of targets. Even in MMW-only mode, the Brimstone is a semiautonomous weapon.

The line between the semiautonomous Brimstone and a fully autonomous weapon that would choose its own targets is a thin one. It isn't based on the seeker or the algorithms. The same seeker and algorithms could be used on a future weapon that *could* loiter over the battlespace—a missile with an upgraded engine or a drone that could patrol an area. A future weapon that patrolled a kill box, rather than entered one at a snapshot in time, would be an autonomous weapon, because the human could send the weapon to monitor the kill box without knowledge of any specific targets. It would allow the human to fire the weapon "blind" and let the weapon decide if and when to strike targets.

Even if the Brimstone doesn't quite cross the line to an autonomous weapon, it takes one more half step toward it, to the point where all that is needed is a light shove to cross the line. A MMW-only Brimstone could be converted into a fully autonomous weapon simply by upgrading the missile's engine so that it could loiter for longer. Or the MMW-only mode algorithms and seeker could be placed on a drone. Notably, the MMW-only mode is enabled in the missile by a software change. As autonomous technology continues to advance, more missiles around the globe will step right up to—or cross—that line.

Would the United Kingdom be willing to cross that line? The debate surrounding another British program, the Taranis drone, shows the difficulty in ascertaining how far the British might be willing to push the technology.

THE TARANIS DRONE

The Taranis is a next-generation experimental combat drone similar to those being developed by the United States, India, Russia, China, France, and Israel. BAE Systems, developer of the Taranis, has given one of the most extensive descriptions of how a combat drone's autonomy might work for weapons engagements. Similar to the X-47B, the Taranis is a demonstrator airplane, but the British military intends to carry the demonstration further than the United States and conduct simulated weapons engagements with the Taranis.

Information released by BAE shows how Taranis might be employed. It explains a simulated weapons test that "will demonstrate the ability of [an unmanned combat aircraft system] to: fend off hostile attack; deploy weapons deep in enemy territory and relay intelligence information." In the test:

1. Taranis would reach the search area via a preprogrammed flight path in the form of a three-dimensional corridor in the sky. Intelligence would be relayed to mission command.
2. When Taranis identifies a target it would be verified by mission command.
3. On the authority of mission command, Taranis would carry out a simulated firing and then return to base via the programmed flight path.

> At all times, Taranis will be under the control of a highly-trained ground crew. The Mission Commander will both verify targets and authorise simulated weapons release.

This protocol keeps the human in the loop to approve each target, which is consistent with other statements by BAE leadership. In a 2016 panel at the World Economic Forum in Davos, BAE Chairman Sir Roger Carr described autonomous weapons as "very dangerous" and "fundamentally wrong." Carr made clear that BAE only envisioned developing weapons that kept a connection to a human who could authorize and remain responsible for lethal decision-making.

In a 2016 interview, Taranis program manager Clive Marrison made a similar statement that "decisions to release a lethal mechanism will always require a human element given the Rules of Engagement used by the UK in the past." Marrison then hedged, saying, "but the Rules of Engagement could change."

The British government reacted swiftly. Following multiple media articles alleging BAE was building in the option for Taranis to "attack targets of its own accord," the UK government released a statement the next day stating:

> The UK does not possess fully autonomous weapon systems and has no intention of developing or acquiring them. The operation of our weapons will always be under human control as an absolute guarantee of human oversight, authority and accountability for their use.

The British government's full-throated denial of autonomous weapons would appear to be as clear a policy statement as there could be, but an important asterisk is needed regarding how the United Kingdom defines an "autonomous weapon system." In its official policy expressed in the UK Joint Doctrine Note 2/11, "The UK Approach to Unmanned Aircraft Systems," the British military describes an autonomous system as one that "must be capable of achieving the same level of situational understanding as a human." Short of that, a system is defined as "automated." This definition of autonomy, which hinges on the complexity of the system rather than its function, is a different way of using the term "autonomy" than many others in discussions on autonomous weapons, including the U.S. government. The United Kingdom's stance is not a product of sloppy language; it's a deliberate choice. The UK doctrine note continues:

> As computing and sensor capability increases, it is likely that many systems, using very complex sets of control rules, will appear and be described as autonomous systems, but as long as it can be shown that the system logically follows a set of rules or instructions and is not capable of human levels of situational understanding, then they should only be considered to be automated.

This definition shifts the lexicon on autonomous weapons dramatically. When the UK government uses the term "autonomous system," they are describing systems with human-level intelligence that are more analogous to the "general AI" described by U.S. Deputy Defense Secretary Work. The effect of this definition is to shift the debate on autonomous weapons to far-off future systems and away from potential near-term weapon systems that may search for, select, and engage targets on their own—what others might call "autonomous weapons." Indeed, in its 2016 statement to the United Nations meetings on autonomous weapons, the United Kingdom stated: "The UK believes that [lethal autonomous weapon systems] do not, and may never, exist." That is to say, Britain may develop weapons that would search for, select, and engage targets on their own; it simply would call them "automated weapons," not "autonomous weapons." In fact, the UK doctrine note refers to systems such as the Phalanx gun (a supervised autonomous weapon) as "fully automated weapon systems." The doctrine note leaves open the possibility of their development, provided they pass a legal weapons review showing they can be used in a manner compliant with the laws of war.

In practice, the British government's stance on autonomous weapons is not dissimilar from that expressed by U.S. defense officials. Humans will remain

involved in lethal decision-making . . . at some level. That might mean a human operator launching an autonomous/automated weapon into an area and delegating to it the authority to search for and engage targets on its own. Whether the public would react differently to such a weapon if it were rebranded an "automated weapon" is unclear.

Even if the United Kingdom's stance retains some flexibility, there is still a tremendous amount of transparency into how the U.S. and UK governments are approaching the question of autonomous weapons. Weapons developers like BAE, MBDA, and Lockheed Martin have detailed descriptions of their weapon systems on their websites, which is not uncommon for defense companies in democratic nations. DARPA describes its research programs publicly and in detail. Defense officials in both countries openly engage in a dialogue about the boundaries of autonomy and the appropriate role of humans and machines in lethal force. This transparency stands in stark contrast to authoritarian regimes.

RUSSIA'S WAR BOTS

While the United States has been very reluctant to arm ground robots, with only one short-lived effort during the Iraq war and no developmental programs for armed ground robots, Russia has shown no such hesitation. Russia is developing a fleet of ground combat robots for a variety of missions, from protecting critical installations to urban combat. Many of Russia's ground robots are armed, ranging from small robots to augment infantry troops to robotic tanks. How much autonomy Russia is willing to place into its ground robots will have a profound impact on the future of land warfare.

The Platform-M, a tracked vehicle roughly the size of a four-wheeler armed with a grenade launcher and an assault rifle, is on the smaller scale of Russian war bots. In 2014, the Platform-M took part in an urban combat exercise alongside Russian troops. According to an official statement from the Russian military, "the military robots were assigned to eliminate provisional illegal armed formations in urban conditions and striking stationary and mobile targets." The Russian military did not describe the degree of the Platform-M's autonomy, although according to the developer:

> Platform-M . . . is used for gathering intelligence, for discovering and eliminating stationary and mobile targets, for firepower support, for patrolling and for guarding important sites. The unit's weapons can be guided, it can carry out supportive tasks and it can destroy targets in automatic or semiautomatic control systems; it is supplied with optical-electronic and radio reconnaissance locators.

The phrase “can destroy targets in automatic . . . control” makes it sound like an autonomous weapon. This claim should be viewed with some skepticism. For one, videos of Russian robots show soldiers selecting targets on a computer screen. More importantly, the reality is that detecting targets autonomously in a ground combat environment is far more technically challenging than targeting enemy radars as the Harpy does or enemy ships on the high seas like TASM. The weapons Platform-M carries—a grenade launcher and assault rifle—would be effective against people, not armored vehicles like tanks or armored personnel carriers. People don’t emit in the electromagnetic spectrum like radars. They aren’t “cooperative targets.” At the time this claim was made in 2014, autonomously finding a person in a cluttered ground combat environment would have been difficult. Advances in neural nets have changed this in the past few years, making it easier to identify people. But discerning friend from foe would still be a challenge.

The autonomous target identification problem Russian war bots face is far more challenging than the South Korean sentry gun on the DMZ. In a demilitarized zone such as that separating North and South Korea, a country might decide to place stationary sentry guns along the border and authorize them to shoot anything with an infrared (heat) signature coming across. Such a decision would not be without its potential problems. Sentry guns that lack any ability to discriminate valid military targets from civilians could senselessly murder innocent refugees attempting to flee an authoritarian regime. In general, though, a DMZ is a more controlled environment than offensive urban combat operations. Authorizing static, defensive autonomous weapons that are fixed in place would be far different than roving autonomous weapons that would be intended to maneuver in urban areas where combatants are mixed in among civilians.

Technologies exist today that could be used for automatic responses against military targets, if the Russians wanted to give such a capability to the Platform-M. The technology is fairly crude, though. For example, the Boomerang shot detection system is a U.S. system that uses an array of microphones to detect incoming bullets and calculate their origin. According to the developer, “Boomerang uses passive acoustic detection and computer-based signal processing to locate a shooter in less than a second.” By comparing the relative time of arrival of a bullet’s shock wave at the various microphones, Boomerang and other shot detection systems can pinpoint a shooter’s direction. It can then call out the location of a shot, for example, “Shot. Two o’clock. 400 meters.” Alternatively, acoustic shot detection systems can be directly connected to a

camera or remote weapon station and automatically aim them at the shooter. Going the next step to allow the gun to automatically fire back at the shooter would not be technically challenging. Once the shot has been detected and the gun aimed, all that it would take would be to pull the trigger.

It's possible this is what Russia means when it says the Platform-M "can destroy targets in automatic . . . control." From an operational perspective, however, authorizing automatic return-fire would be quite hazardous. It would require an extreme confidence in the ability of the shot detection system to weed out false positives and to not be fooled by acoustic reflections and echoes, especially in urban areas. Additionally, the gun would have no ability to account for collateral damage—say, to hold fire because the shooter is using human shields. Finally, such a system would be a recipe for fratricide, with robot systems potentially automatically shooting friendly troops or other friendly robots. Two robots on the same side could become trapped in a never-ending loop of automatic fire and response, mindlessly exchanging gunfire until they exhausted their ammunition or destroyed each other. It is unclear whether this is what Russia intends, but from a technical standpoint it would possible.

Russia's other ground combat robots scale up in size and sophistication from the Platform-M. The MRK-002-BG-57 "Wolf-2" is the size of a small car and outfitted with a 12.7 mm heavy machine gun. According to David Hambling of *Popular Mechanics*, "In the tank's automated mode, the operator can remotely select up to 10 targets, which the robot then bombards. Wolf-2 can act on its own to some degree (the makers are vague about what degree), but the decision to use lethal force is ultimately under human control." The Wolf-2 sits among a family of similar size robot vehicles. The amphibious Argo is roughly the size of a Mini Cooper, sports a machine gun and rocket-propelled grenade launcher, and can swim at speeds up to 2.5 knots. The A800 Mobile Autonomous Robotic System (MARS) is an (unarmed) infantry support vehicle the size of a compact car that can carry four infantry soldiers and their gear. Pictures online show Russian soldiers riding on the back, looking surprisingly relaxed as the tracked robot cruises through an off-road course.

Compact car–sized war bots aren't necessarily unique to Russia, although the Russian military seems to have a casual attitude toward arming them not seen in Western nations. The Russian military isn't stopping at midsize ground robots, though. Several Russian programs are pushing the boundaries of what is possible with robotic combat vehicles, building systems that could prove decisive in highly lethal tank-on-tank warfare.

The Uran-9 looks like something straight out of a *MechWarrior* video game,

where players pilot a giant robot warrior armed with rockets and cannons. The Uran-9 is fully uninhabited, although it is controlled by soldiers remotely from a nearby command vehicle. It is the size of a small armored personnel carrier, sports a 30 mm cannon, and has an elevated platform to launch antitank guided missiles. The elevated missile platform that gives the Uran-9 a distinctive sci-fi appearance. The missiles rest on two platforms on either side of the vehicle that, when raised, look like arms reaching into the sky. The elevated platform allows the robot to fire missiles while safely sitting behind cover, for example behind the protective slope of a hillside. In an online promotional video from the developer, Rosoboronexport, slo-mo shots of the Uran-9 firing antitank missiles are set to music reminiscent of a Tchaikovsky techno remix.

The Uran-9 is a major step beyond smaller robotic platforms like the Platform-M and Wolf-2 not just because it's larger, but because its larger size allows it to carry heavier weapons capable of taking on antitank missions. Whereas the assault rifle and grenade launcher on a Platform-M would do very little to a tank, the Uran-9's antitank missiles would be potentially highly lethal. This makes the Uran-9 potentially a useful weapon in high-intensity combat against NATO forces on the plains of Europe. Uran-9s could hide behind hillsides or other protective cover and launch missiles against NATO tanks. The Uran-9 doesn't have the armor or guns to stand toe-to-toe against a modern tank, but because it's uninhabited, it doesn't have to. The Uran-9 could be a successful ambush predator. Even if firing its missiles exposed its position and led it to be taken out by NATO forces, the exchange might still be a win if it took out a Western tank. Because there's no one inside it and the Uran-9 is significantly smaller than a tank, and therefore presumably less expensive, Russia could field many of them on the battlefield. Just like many stings from a hornet can bring down a much larger animal, the Uran-9 could make the modern battlefield a deadly place for Western forces.

Russia's Vikhr "robot tank" has a similar capability. At 14 tons and lacking a main gun, it is significantly smaller and less lethal than a 50-to 70-ton main battle tank. Like the Uran-9, though, its 30 mm cannon and six antitank missiles show it is designed as a tank-killing ambush predator, not a tank-on-tank street fighter. The Vikhr is remote controlled, but news reports indicate it has the ability to "lock onto a target" and keep firing until the target is destroyed. While not the same as choosing its own target, tracking a moving target is doable today. In fact, tracking moving objects is as a standard feature on DJI's base model Spark hobby drone, which retails for under $500.

Taking the next step and allowing the Uran-9 or Vikhr to autonomously

target tanks would take some additional work, but it would be more feasible than trying to accurately discriminate among human targets. With large cannons and treads, tanks are distinctive military vehicles not easily confused with civilian objects. Moreover, militaries may be more willing to risk civilian casualties or fratricide in the no-holds-barred arena of tank warfare, where armored divisions vie for dominance and the fate of nations is at stake. In videos of the Uran-9, human operators can be clearly seen controlling the vehicle, but the technology is available for Russia to authorize fully autonomous antitank engagements, if it chose to do so.

Russia isn't stopping at development of the Vikhr and Uran-9, however. It envisions even more advanced robotic systems that could not only ambush Western tanks, but stand with them toe-to-toe and win. Russia reportedly has plans to develop a fully robotic version of its next-generation T-14 Armata tank. The T-14 Armata, which reportedly entered production as of 2016, sports a bevy of new defensive features, including advanced armor, an active protection system to intercept incoming antitank missiles, and a robotic turret. The T-14 will be the first main battle tank to sport an uninhabited turret, which will afford the crew greater protection by sheltering them within the body of the vehicle. Making the entire tank uninhabited would be the next logical step in protection, enabling a crew to control the vehicle remotely. While current T-14s are human-inhabited, Russia has long-term plans to develop a fully robotic version. Vyacheslav Khalitov, deputy director general of UralVagonZavod, manufacturer of the T-14 Armata, has stated, "Quite possibly, future wars will be waged without human involvement. That is why we have made provisions for possible robotization of Armata." He acknowledged that achieving the goal of full robotization would require more advanced AI that could "calculate the situation on the battlefield and, on this basis, to take the right decision."

In addition to pushing the boundaries on robots' physical characteristics, the Russian military has signaled it intends to use cutting-edge AI to boost its robots' decision-making. In July 2017, Russian arms manufacturer Kalashnikov stated that they would soon release "a fully automated combat module" based on neural networks. News reports indicate the neural networks would allow the combat module "to identify targets and make decisions." As in other cases, it is difficult to independently evaluate these claims, but they signal a willingness to use artificial intelligence for autonomous targeting. Russian companies' boasting of autonomous features has none of the hesitation or hedging that is often seen from American or British defense firms.

Senior Russian military commanders have stated they intend to move toward

fully robotic weapons. In a 2013 article on the future of warfare, Russian military chief of staff General Valery Gerasimov wrote:

> Another factor influencing the essence of modern means of armed conflict is the use of modern automated complexes of military equipment and research in the area of artificial intelligence. While today we have flying drones, tomorrow's battlefields will be filled with walking, crawling, jumping, and flying robots. In the near future it is possible a fully robotized unit will be created, capable of independently conducting military operations.
>
> How shall we fight under such conditions? What forms and means should be used against a robotized enemy? What sort of robots do we need and how can they be developed? Already today our military minds must be thinking about these questions.

This Russian interest in pursuing fully robotic units has not escaped notice in the West. In December 2015, Deputy Secretary of Defense Bob Work mentioned Gerasimov's comments in a speech on the future of warfare. As Work has repeatedly noted, U.S. decisions may be shaped by those of Russia and other nations. This is the danger of an arms race in autonomy: that nations feel compelled to race forward and build autonomous weapons out of the fear that others are doing so, without pausing to weigh the risks of their actions.

AN ARMS RACE IN AUTONOMOUS WEAPONS?

If it is true, as some have suggested, that a dangerous arms race in autonomous weapons is under way, then it is a strange kind of race. Nations are pursuing autonomy in many aspects of weaponry but, with the exception of the Harpy, are still keeping humans in the loop for now. Some weapons like Brimstone use autonomy in novel ways, pushing the boundaries of what could be considered a semiautonomous weapon. DARPA's CODE program appears to countenance moving to human-*on*-the-loop supervisory control for some types of targets, but there is no indication of full autonomy. Developers of the SGR-A1 gun and Taranis drone have suggested full autonomy could be a future option, although higher authorities immediately disputed the claim, saying that was not their intent.

Rather than a full-on sprint to build autonomous weapons, it seems that many nations do not yet know whether they might want them in the future and are hedging their bets. One challenge in understanding the global landscape of lethal autonomy is that the degree of transparency among nations differs greatly. While the official policies of the U.S. and UK governments leave room to develop autonomous weapons (although they express this differently with the

United Kingdom calling them “automated weapons”) countries such as Russia don’t even have a public policy. Policy discussions may be happening in private in authoritarian regimes, but we don’t know what they are. Pressure from civil society for greater transparency differs greatly across countries. In 2016, the UK-based NGO Article 36, which has been a leading voice in shaping international discussions on autonomous weapons, wrote a policy brief critiquing the UK government’s stance on autonomous weapons. In the United States, Stuart Russell and a number of well-respected colleagues from the AI community have met with mid-level officials from across the U.S. government to discuss autonomous weapons. In authoritarian Russia, there are no equivalent civil society groups to pressure the government to be more transparent about its plans. As a result, scrutiny focuses on the most transparent countries—democratic nations who are responsive to elements of civil society and are generally more open about their weapons development. What goes on in authoritarian regimes is far murkier, but no less relevant to the future path of lethal autonomy.

Looking across the global landscape of robotic systems, it’s clear that many nations are pursuing armed robots, including combat drones that would operate in contested air space. How much autonomy some weapon systems have is unclear, but there is nothing preventing countries from crossing the line to lethal autonomy in their next-generation missiles, combat drones, or ground robots. Next-generation robotic systems such as the Taranis may give countries that option, forcing uncomfortable conversations. Even if many countries would rather not move forward with autonomous weapons, it may only take one to start a cascade of others.

With no autonomous smoking gun, it seems unnecessarily alarmist to declare that an autonomous weapons arms race is already under way, but we could very well be at the starting blocks. The technology to build autonomous weapons is widely available. Even non-state groups have armed robots. The only missing ingredient to turn a remotely controlled armed robot into an autonomous weapon is software. That software, it turns out, is pretty easy to come by.

8

GARAGE BOTS

DIY KILLER ROBOTS

A gunshot cuts through the low buzz of the drone's rotors. The camera jerks backward from the recoil. The gun fires again. A small bit of flame darts out of the handgun attached to the homemade-looking drone. Red and yellow wires snake over the drone and into the gun's firing mechanism, allowing the human controller to remotely pull the trigger.

The controversial fifteen-second video clip released in the summer of 2015 was taken by a Connecticut teenager of a drone he armed himself. Law enforcement and the FAA investigated, but no laws were broken. The teenager used the drone on his family's property in the New England woods. There are no laws against firing weapons from a drone, provided it's done on private property. A few months later, for Thanksgiving, he posted a video of a flamethrower-armed drone roasting a turkey.

Drones are not only in wide use by countries around the globe; they are readily purchased by anyone online. For under $500, one can buy a small quadcopter that can autonomously fly a route preprogrammed by GPS, track and follow moving objects, and sense and avoid obstacles in its path. Commercial drones are moving forward in leaps and bounds, with autonomous behavior improving in each generation.

When I asked the Pentagon's chief weapons buyer Frank Kendall what he feared, it wasn't Russian war bots, it was cheap commercial drones. A world where everyone has access to autonomous weapons is a far different one than a world where only the most advanced militaries can build them. If autonomous

weapons could be built by virtually anyone in their garage, bottling up the technology and enforcing a ban, as Stuart Russell and others have advocated, would be extremely difficult. I wanted to know, could someone leverage commercially available drones to make a do-it-yourself (DIY) autonomous weapon? How hard would it be?

I was terrified by what I found.

HUNTING TARGETS

The quadcopter rose off the ground confidently, smoothly gaining altitude till it hovered around eye level. The engineer next to me tapped his tablet and the copter moved out, beginning its search of the house.

I followed along behind the quadcopter, watching it navigate each room. It had no map, no preprogrammed set of instructions for where to go. The drone was told merely to search and report back, and so it did. As it moved through the house it scanned each room with a laser range-finding LIDAR sensor, building a map as it went. Transmitted via Wi-Fi, the map appeared on the engineer's tablet.

As the drone glided through the house, each time it came across a doorway it stopped, its LIDAR sensor probing the space beyond. The drone was programmed to explore unknown spaces until it had mapped everything. Only then would it finish its patrol and report back.

I watched the drone pause in front of an open doorway. I imagined its sensors pinging the distant wall of the other room, its algorithms computing that there must be unexplored space beyond the opening. The drone hovered for a moment, then moved into the unknown room. A thought popped unbidden into my mind: *it's curious*.

It's silly to impart such a human trait to a drone. Yet it comes so naturally to us, to imbue nonhuman objects with emotions, thoughts, and intentions. I was reminded of a small walking robot I had seen in a university lab years ago. The researchers taped a face to one end of the robot—nothing fancy, just slices of colored construction paper in the shape of eyes, a nose, and a mouth. I asked them why. Did it help them remember which direction was forward? No, they said. It just made them feel better to put a face on it. It made the robot seem more human, more like us. There's something deep in human nature that wants to connect to another sentient entity, to know that it is like us. There's something alien and chilling about entities that can move intelligently through the world and not feel any emotion or thought beyond their own programming. There is

something predatory and remorseless about them, like a shark.

I shook off the momentary feeling and reminded myself of what the technology was actually doing. The drone "felt" nothing. The computer controlling its actions would have identified that there was a gap where the LIDAR sensors could not reach and so, following its programming, directed the drone to enter the room.

The technology *was* impressive. The company I was observing, Shield AI, was demonstrating fully autonomous indoor flight, an even more impressive feat than tracking a person and avoiding obstacles outdoors. Founded by brothers Ryan and Brandon Tseng, the former an engineer and the latter a former Navy SEAL, Shield AI has been pushing the boundaries of autonomy under a grant from the U.S. military. Shield's goal is to field fully autonomous quadcopters that special operators can launch into an unknown building and have the drones work cooperatively to map the building on their own, sending back footage of the interior and potential objects of interest to the special operators waiting outside.

Brandon described their goal as "highly autonomous swarms of robots that require minimal human input. That's the end-state. We envision that the DoD will have ten times more robots on the battlefield than soldiers, protecting soldiers and innocent civilians." Shield's work is pushing the boundaries of what is possible today. All the pieces of the technology are falling into place. The quadcopter I witnessed was using LIDAR for navigation, but Shield's engineers explained they had tested visual-aided navigation; they simply didn't have it active that day.

Visual-aided navigation is a critically important piece of technology that will allow drones to move autonomously through cluttered environments without the aid of GPS. Visual-aided navigation tracks how objects move through the camera's field of view, a process called "optical flow." By assessing optical flow, operating on the assumption that most of the environment is static and not moving, fixed objects moving through the camera's field of vision can be used as a reference point for the drone's own movement. This can allow the drone to determine how it is moving within its environment without relying on GPS or other external navigation aids. Visual-aided navigation can complement other internal guidance mechanisms, such as inertial measurement units (IMU) that work like a drone's "inner ear," sensing changes in velocity. (Imagine sitting blindfolded in a car, feeling the motion of the car's acceleration, braking, and turning.) When IMUs and visual-aided navigation are combined, they make an extremely powerful tool for determining a drone's position, allowing the drone

to accurately navigate through cluttered environments without GPS.

Visual-aided navigation has been demonstrated in numerous laboratory settings and will no doubt trickle down to commercial quadcopters over time. There is certain to be a market for quadcopters that can autonomously navigate indoors, from filming children's birthday parties to indoor drone racing. With visual-aided navigation and other features, drones and other robotic systems will increasingly be able to move intelligently through their environment. Shield AI, like many tech companies, was focused on near-term applications, but Brandon Tseng was bullish on the long-term potential of AI and autonomy. "Robotics and artificial intelligence are where the internet was in 1994," he told me. "Robotics and AI are about to have a really transformative impact on the world. . . . Where we see the technology 10 to 15 years down the road? It is going to be mind-blowing, like a sci-fi movie."

Autonomous navigation is not the same as autonomous targeting, though. Drones that can maneuver and avoid obstacles on their own—indoors or outdoors—do not necessarily have the ability to identify and discriminate among the various objects in their surroundings. They simply avoid hitting anything at all. Searching for specific objects and targeting them for action—whether it's taking photographs or something more nefarious—would require more intelligence.

The ability to do target identification is the key missing link in building a DIY autonomous weapon. An autonomous weapon is one that can search for, decide to engage, and engage targets. That requires three abilities: the ability to maneuver intelligently through the environment to search; the ability to discriminate among potential targets to identify the correct ones; and the ability to engage targets, presumably through force. The last element has already been demonstrated—people have armed drones on their own. The first element, the ability to autonomously navigate and search an area, is already available outdoors and is coming soon indoors. Target identification is the only piece remaining, the only obstacle to someone making an autonomous weapon in their garage. Unfortunately, that technology is not far off. In fact, as I stood in the basement of the building watching Shield AI's quadcopter autonomously navigate from room to room, autonomous target recognition was literally being demonstrated right outside, just above my head.

DEEP LEARNING

The research group asked that they not be named, because the technology was

new and untested. They didn't want to give the impression that it was good enough—that the error rate was low enough—to be used for military applications. Nor, it was clear, were military applications their primary intention in designing the system. They were engineers, simply trying to see if they could solve a tough problem with technology. Could they send a small drone out entirely on its own to autonomously find a crashed helicopter and report its location back to the human?

The answer, it turns out, is yes. To understand how they did it, we need to go deep.

Deep learning neural networks, first mentioned in chapter 5 as one potential solution to improving military automatic target recognition in DARPA's TRACE program, have been the driving force behind astounding gains in AI in the past few years. Deep neural networks have learned to play Atari, beat the world's reigning champion at *go*, and have been behind dramatic improvements in speech recognition and visual object recognition. Neural networks are also behind the "fully automated combat module" that Russian arms manufacturer Kalashnikov claims to have built. Unlike traditional computer algorithms that operate based on a script of instructions, neural networks work by learning from large amounts of data. They are an extremely powerful tool for handling tricky problems that can't be easily solved by prescribing a set of rules to follow.

Let's say, for example, that you wanted to write down a rule set for how to visually distinguish an apple from a tomato without touching, tasting, or smelling. Both are round. Both are red and shiny. Both have a green stem on top. They *look* different, but the differences are subtle and evade easy description. Yet a three-year-old child can immediately tell the difference. This is a tricky problem with a rules-based approach. What neural networks do is sidestep that problem entirely. Instead, they learn from vast amounts of data—tens of thousands or millions of pieces of data. As the network churns through the data, it continually adapts its internal structure until it optimizes to achieve the correct programmer-specified goal. The goal could be distinguishing an apple from a tomato, playing an Atari game, or some other task.

In one of the most powerful examples of how neural networks can be used to solve difficult problems, the Alphabet (formerly Google) AI company DeepMind trained a neural network to play *go*, a Chinese strategy game akin to chess, better than any human player. *Go* is an excellent game for a learning machine because the sheer complexity of the game makes it very difficult to program a computer to play at the level of a professional human player based on a rules-based strategy alone.

The rules of *go* are simple, but from these rules flows vast complexity. *Go* is played on a grid of 19 by 19 lines and players take turns placing stones—black for one player and white for the other—on the intersection points of the grid. The objective is to use one's stones to encircle areas of the board. The player who controls more territory on the board wins. From these simple rules come an almost unimaginably large number of possibilities. There are more possible positions in *go* than there are atoms in the known universe, making *go* 10^{100} (one followed by a hundred zeroes) times—literally a googol—more complex than chess.

Humans at the professional level play *go* based on intuition and feel. *Go* takes a lifetime to master. Prior to DeepMind, attempts to build *go*-playing AI software had fallen woefully short of human professional players. To craft its AI, called AlphaGo, DeepMind took a different approach. They built an AI composed of deep neural networks and fed it data from 30 million games of *go*. As explained in a DeepMind blog post, "These neural networks take a description of the Go board as an input and process it through 12 different network layers containing millions of neuron-like connections." Once the neural network was trained on human games of *go*, DeepMind then took the network to the next level by having it play itself. Our goal is to beat the best human players, not just mimic them," as explained in the post. "To do this, AlphaGo learned to discover new strategies for itself, by playing thousands of games between its neural networks, and adjusting the connections using a trial-and-error process known as reinforcement learning." AlphaGo used the 30 million human games of *go* as a starting point, but by playing against itself could reach levels of game play beyond even the best human players.

This superhuman game play was demonstrated in the 4–1 victory AlphaGo delivered over the world's top-ranked human *go* player, Lee Sedol, in March 2016. AlphaGo won the first game solidly, but in game 2 demonstrated its virtuosity. Partway through game 2, on move 37, AlphaGo made a move so surprising, so un-human, that it stunned professional players watching the match. Seemingly ignoring a contest between white and black stones that was under way in one corner of the board, AlphaGo played a black stone far away in a nearly empty part of the board. It was a surprising move not seen in professional games, so much so that one commentator remarked, "I thought it was a mistake." Lee Sedol was similarly so taken by surprise he got up and left the room. After he returned, he took fifteen minutes to formulate his response. AlphaGo's move wasn't a mistake. European *go* champion Fan Hui, who had lost to AlphaGo a few months earlier in a closed-door match, said at first the move surprised him

as well, and then he saw its merit. “It’s not a human move,” he said. “I’ve never seen a human play this move. So beautiful.” Not only did the move *feel* like a move no human player would never make, it was a move no human player probably would never make. AlphaGo rated the odds that a human would have made that move as 1 in 10,000. Yet AlphaGo made the move anyway. AlphaGo went on to win game 2 and afterward Lee Sedol said, “I really feel that AlphaGo played the near perfect game.” After losing game 3, thus giving AlphaGo the win for the match, Lee Sedol told the audience at a press conference, “I kind of felt powerless.”

AlphaGo’s triumph over Lee Sedol has implications far beyond the game of *go*. More than just another realm of competition in which AIs now top humans, the way DeepMind trained AlphaGo is what really matters. As explained in the DeepMind blog post, “AlphaGo isn’t just an ‘expert’ system built with hand-crafted rules; instead it uses general machine learning techniques to figure out for itself how to win at Go.” DeepMind didn’t program rules for how to win at *go*. They simply fed a neural network massive amounts of data and let it learn all on its own, and some of the things it learned were surprising.

In 2017, DeepMind surpassed their earlier success with a new version of AlphaGo. With an updated algorithm, AlphaGo Zero learned to play *go* without any human data to start. With only access to the board and the rules of the game, AlphaGo Zero taught itself to play. Within a mere three days of self-play, AlphaGo Zero had eclipsed the previous version that had beaten Lee Sedol, defeating it 100 games to 0.

These deep learning techniques can solve a variety of other problems. In 2015, even before DeepMind debuted AlphaGo, DeepMind trained a neural network to play Atari games. Given only the pixels on the screen and the game score as input and told to maximize the score, the neural network was able to learn to play Atari games at the level of a professional human video game tester. Most importantly, the same neural network architecture could be applied across a vast array of Atari games—forty-nine games in all. Each game had to be individually learned, but the same neural network architecture applied to any game; the researchers didn’t need to create a customized network design for each game.

The AIs being developed for *go* or Atari are still narrow AI systems. Once trained, the AIs are purpose-built tools to solve narrow problems. AlphaGo can beat any human at *go*, but it can’t play a different game, drive a car, or make a cup of coffee. Still, the tools used to train AlphaGo are generalizable tools that can be used to build any number of special-purpose narrow AIs to solve various

problems. Deep neural networks have been used to solve other thorny problems that have bedeviled the AI community for years, notably speech recognition and visual object recognition.

A deep neural network was the tool used by the research team I witnessed autonomously find the crashed helicopter. The researcher on the project explained that he had taken an existing neural network that had already been trained on object recognition, stripped off the top few layers, then retrained the network to identify helicopters, which hadn't originally been in its image dataset. The neural network he was using was running off of a laptop connected to the drone, but it could just as easily have been running off of a Raspberry Pi, a $40 credit-card sized processor, riding on board the drone itself.

All of these technologies are coming from outside the defense sector. They are being developed at places like Google, Microsoft, IBM, and university research labs. In fact, programs like DARPA's TRACE are not necessarily intended to invent new machine learning techniques, but rather import existing techniques into the defense sector and apply them to military problems. These methods are widely available to those who know how to use them. I asked the researcher behind the helicopter-hunting drone: Where did he get the initial neural network that he started with, the one that was already trained to recognize other images that weren't helicopters? He looked at me like I was either half-crazy or stupid. He got it online, of course.

NEURAL NETS FOR EVERYONE

I feel I should confess that I'm not a technologist. In my job as a defense analyst, I research military technology to make recommendations about where the U.S. military should invest to keep its edge on the battlefield, but I don't build things. My undergraduate degree was in science and engineering, but I've done nothing even remotely close to engineering since then. To claim my programming skills were rusty would be to imply that at one point in time they existed. The extent of my computer programming knowledge is a one-semester introductory course in C++ in college.

Nevertheless, I went online to check out the open-source software database the researcher pointed me to: TensorFlow. TensorFlow is an open-source AI library developed by Google AI researchers. With TensorFlow, Google researchers have taken what they have been learning with deep neural networks and passed it on to the rest of the world. On TensorFlow, not only can you download already trained neural networks and software for building your own,

there are reams of tutorials on how to teach yourself deep learning techniques. For users new to machine learning, there are basic tutorials on classic machine learning problems. These tools make neural networks accessible to computer programmers with little to no experience in machine learning. TensorFlow makes neural networks easy, even fun. A tutorial called Playground (playground.tensorflow.org) allows users to modify and train a neural network through a point-and-click interface in the browser. No programming skills are required at all.

Once I got into Playground, I was hooked. Reading about what neural networks could do was one thing. Building your own and training it on data was entirely another. Hours of time evaporated as I tinkered with the simple network in my browser. The first challenge was training the network to learn to predict the simple datasets used in Playground—patterns of orange and blue dots across a two-dimensional grid. Once I'd mastered that, I worked to make the leanest network I could, composed of the fewest neurons in the fewest number of layers that could still accurately make predictions. (Reader challenge: once you've mastered the easy datasets, try the spiral.)

With the Playground tutorial, the concept of neural nets becomes accessible to someone with no programming skills at all. Using Playground is no more complicated than solving an easy-level Sudoku puzzle and within the range of an average seven-year-old. Playground won't let the user build a custom neural net to solve novel problems. It's an illustration of what neural nets can do to help users see their potential. Within other parts of TensorFlow, though, lie more powerful tools to use existing neural networks or design custom ones, all within reach of a reasonably competent programmer in Python or C++.

TensorFlow includes extensive tutorials on convolutional neural nets, the particular type of neural network used for computer vision. In short order, I found a neural network available for download that was already trained to recognize images. The neural network Inception-v3 is trained on the ImageNet dataset, a standard database of images used by programmers. Inception-v3 can classify images into one of 1,000 categories, such as "gazelle," "canoe," or "volcano." As it turns out, none of the categories Inception-v3 is trained on are those that could be used to identify people, such as "human," "person," "man," or "woman." So one could not, strictly speaking, use this particular neural network to power an autonomous weapon that targets people. Still, I found this to be little consolation. ImageNet isn't the only visual object classification database used for machine learning online and others, such as the Pascal Visual Object Classes database, include "person" as a category. It took me all of about

ten seconds on Google to find trained neural networks available for download that could find human faces, determine age and gender, or label human emotions. All of the tools to build an autonomous weapon that could target people on its own were readily available online.

This was, inevitably, one of the consequences of the AI revolution. AI technology was powerful. It could be used for good purposes or bad purposes; that was up to the people using it. Much of the technology behind AI was software, which meant it could be copied practically for free. It could be downloaded at the click of a button and could cross borders in an instant. Trying to contain software would be pointless. Pandora's box has already been opened.

ROBOTS EVERYWHERE

Just because the tools needed to make an autonomous weapon were widely available didn't tell me how easy or hard it would be for someone to actually do it. What I wanted to understand was how widespread the technological know-how was to build a homemade robot that could harness state-of-the-art techniques in deep learning computer vision. Was this within reach of a DIY drone hobbyist or did these techniques require a PhD in computer science?

There is a burgeoning world of robot competitions among high school students, and this seemed like a great place to get a sense of what an amateur robot enthusiast could do. The FIRST Robotics Competition is one such competition that includes 75,000 students organized in over 3,000 teams across twenty-four countries. To get a handle on what these kids might be able to do, I headed to my local high school.

Less than a mile from my house is Thomas Jefferson High School for Science and Technology—"TJ," for short. TJ is a math and science magnet school; kids have to apply to get in, and they are afforded opportunities above and beyond what most high school students have access to. But they're still high school students—not world-class hackers or DARPA whizzes.

In the Automation and Robotics Lab at TJ, students get hands-on experience building and programming robots. When I visited, two dozen students sat at workbenches hunched over circuit boards or silently tapping away at computers. Behind them on the edges of the workshop lay discarded pieces of robots, like archeological relics of students' projects from semesters prior. On a shelf sat "Roby Feliks," the Rubik's Cube solving robot. Nearby, a Raspberry Pi processor sat atop a plastic musical recorder, wires running from the circuit board to the instrument like some musical cyborg. Somewhat randomly in the

center of the floor sat a half-disassembled robot, the remnants of TJ's admission to the FIRST competition that year. Charles Dela Cuesta, the teacher in charge of the lab, apologized for the mess, but it was exactly what I imagined a robot lab should look like.

Dela Cuesta came across as the kind of teacher you pray your own children have. Laid back and approachable, he seemed more like a lovable assistant coach than an aloof disciplinarian. The robotics lab had the feel of a place where students learn by doing, rather than sitting and copying down equations from a whiteboard.

Which isn't to say that there wasn't a whiteboard. There was. It sat in a corner amid a pile of other robotic projects, with circuit boards and wires draped over it. Students were designing an automatic whiteboard with a robot arm that could zip across the surface and sketch out designs from a computer. On the whiteboard were a series of inhumanly straight lines sketched out by the robot. It was at this point that I wanted to quit my job and sign up for a robotics class at TJ.

Dela Cuesta explained that all students at TJ must complete a robotics project in their freshmen year as part of their required coursework. "Every student in the building has had to design a small robot that is capable of navigating a maze and performing some sort of obstacle avoidance," he said. Students are given a schematic of what the maze looks like so they get to choose how to solve the problem, whether to preprogram the robot's moves or take the harder path of designing an autonomous robot that can figure it out on its own. After this required class, TJ offers two additional semesters of robotics electives, which can be complemented with up to five computer science courses in which students learn Java, C++, and Python. These are vital programming tools for using robot control systems, like the Raspberry Pi processor, which runs on Linux and takes commands in Python. Dela Cuesta explained that even though most students come into TJ with no programming experience, many learn fast and some even take computer science courses over the summer to get ahead. "They can pretty much program in anything—Java, Python. . . . They're just all over the place," he said. Their senior year, all students at TJ must complete a senior project in an area of their choosing. Some of the most impressive robotics projects are those done by seniors who choose to make robotics their area of focus. Next to the whiteboard stood a bicycle propped up on its kickstand. A large blue box sat inside the frame, wires snaking out of it to the gear shifters. Dela Cuesta explained it was an automatic gear shifter for the bike. The box senses when it is time to shift and does so automatically, like an automatic

transmission on a car.

The students' projects have been getting better over the years, Dela Cuesta explained, as they are able to harness more advanced open-source components and software. A few years ago, a class project to create a robot tour guide for the school took two years to complete. Now, the timeline has been shortened to nine weeks. "The stuff that was impressive to me five, six years ago we could accomplish in a quarter of the time now. It just blows my mind," he said. Still, Dela Cuesta pushes students to build things custom themselves rather than use existing components. "I like to have the students, as much as possible, build from scratch." Partly, this is because it's often easier to fit custom-built hardware into a robot, an approach that is possible because of the impressive array of tools Dela Cuesta has in his shop. Along a back wall were five 3-D printers, two laser cutters to make custom parts, and a mill to etch custom circuit boards. An even more important reason to have students do things themselves is they learn more that way. "Custom is where I want to go," Dela Cuesta said. "They learn a lot more from it. It's not just kind of this black box magic thing they plug in and it works. They have to really understand what they're doing in order to make these things work."

Across the hall in the computer systems lab, I saw the same ethos on display. The teachers emphasized having students do things themselves so they were learning the fundamental concepts, even if that meant re-solving problems that have already been solved. Repackaging open-source software isn't what the teachers are after. That isn't to say that students aren't learning from the explosion in open-source neural network software. On one teacher's desk sat a copy of Jeff Heaton's *Artificial Intelligence for Humans, Volume 3: Deep Learning and Neural Networks.* (This title begs the uncomfortable question whether there is a parallel course of study, *Artificial Intelligence for Machines*, where machines learn to program other machines. The answer, I suppose, is "Not yet.") Students are learning how to work with neural networks, but they're doing so from the bottom up. A junior explained to me how he trained a neural network to play tic-tac-toe—a problem that was solved over fifteen years ago, but remains a seminal coding problem. Next year, TJ will offer a course in computer vision that will cover convolutional neural networks.

Maybe it's a cliché to say that the projects students were working on are mind-blowing, but I was floored by the things I saw TJ students doing. One student was disassembling a Keurig machine and turning it into a net-enabled coffeemaker so it could join the Internet of Things. Wires snaked through it as though the internet was physically infiltrating the coffeemaker, like *Star Trek's*

Borg. Another student was tinkering with something that looked like a cross between a 1980s Nintendo Power Glove and an Apple smartwatch. He explained it was a "gauntlet," like that used by Iron Man. When I stared at him blankly, he explained (in that patient explaining-to-an-old-person voice that young people use) that a gauntlet is the name for the wrist-mounted control that Iron Man uses to fly his suit. "Oh, yeah. That's cool," I said, clearly not getting it. I don't feel like I need the full functionality of my smartphone mounted on my wrist, but then again I wouldn't have thought ten years ago that I needed a touchscreen smartphone on my person at all times in the first place. Technology has a way of surprising us. Today's technology landscape is a democratized one, where game-changing innovations don't just come out of tech giants like Google and Apple but can come from anyone, even high-school students. The AI revolution isn't something that is happening *out there*, only in top-tier research labs. It's happening everywhere.

THE EVERYONE REVOLUTION

I asked Brandon Tseng from Shield AI where this path to ever-greater autonomy was taking us. He said, "I don't think we're ever going to give [robots] complete autonomy. Nor do I think we should give them complete autonomy." On one level, it's reassuring to know that Tseng, like nearly everyone I met working on military robotics, saw a limit to how much autonomy we should give machines. Reasonable people might disagree on where that limit is, and for some people autonomous weapons that search for and engage targets within narrow human-defined parameters might be acceptable, but everyone I spoke with agreed there should be some limits. But the scary thing is that reasonableness on the part of Tseng and other engineers may not be enough. What's to stop a technologically inclined terrorist from building a swarm of people-hunting autonomous weapons and letting them loose in a crowded area? It might take some engineering and some time, but the underlying technological know-how is readily available. We are entering a world where the technology to build lethal autonomous weapons is available not only to nation-states but to individuals as well. That world is not in the distant future. It's already here.

What we do with the technology is an open question. What would be the consequence of a world of autonomous weapons? Would they lead to a robutopia or robopocalypse? Writers have pondered this question in science fiction for decades, and their answers vary wildly. The robots of Isaac Asimov's books are mostly benevolent partners to humans, helping to protect and guide

humanity. Governed by the Three Laws of Robotics, they are incapable of harming humans. In *Star Wars*, droids are willing servants of humans. In the *Matrix* trilogy, robots enslave humans, growing them in pods and drawing on their body heat for power. In the *Terminator* series, Skynet strikes in one swift blow to exterminate humanity after it determines humans are a threat to its existence.

We can’t know with any certainty what a future of autonomous weapons would look like, but we do have better tools than science fiction to guess at what promise and perils they might bring. Humanity’s past and present experiences with autonomy in the military and other settings point to the potential benefits and dangers of autonomous weapons. These lessons allow us to peer into a murky future and, piece by piece, begin to discern the shape of things to come.

PART III

Runaway Gun

9

ROBOTS RUN AMOK

FAILURE IN AUTONOMOUS SYSTEMS

March 22, 2003—The system said to fire. The radars had detected an incoming tactical ballistic missile, or TBM, probably a Scud missile of the type Saddam had used to harass coalition forces during the first Gulf War. This was their job, shooting down the missile. They needed to protect the other soldiers on the ground, who were counting on them. It was an unfamiliar set of equipment; they were supporting an unfamiliar unit; they didn't have the intel they needed. But this was their job. The weight of the decision rested on a twenty-two-year-old second lieutenant fresh out of training. She weighed the available evidence. She made the best call she could: *fire*.

With a BOOM-ROAR-WOOSH, the Patriot PAC-2 missile left the launch tube, lit its engine, and soared into the sky to take down its target. The missile exploded. Impact. The ballistic missile disappeared from their screens: their first kill of the war. Success.

From the moment the Patriot unit left the States, circumstances had been against them. First, they'd fallen in on a different, older, set of equipment than what they'd trained on. Then once in theater, they were detached from their parent battalion and attached to a new battalion whom they hadn't worked with before. The new battalion was using the newer model equipment, which meant their old equipment (which they weren't fully trained on in the first place) couldn't communicate with the rest of the battalion. They were in the dark. Their systems couldn't connect to the larger network, depriving them of vital information. All they had was a radio.

But they were soldiers, and they soldiered on. Their job was to protect coalition troops against Iraqi missile attacks, and so they did. They sat in their command trailer, with outdated gear and imperfect information, and they made the call. When they saw the missiles, they took the shots. They protected people.

The next night, at 1:30 a.m., there was an attack on a nearby base. A U.S. Army sergeant threw a grenade into a command tent, killing one soldier and wounding fifteen. He was promptly detained but his motives were unclear. Was this the work of one disgruntled soldier or was he an infiltrator? Was this the first of a larger plot? Word of the attack spread over the radio. Soldiers were sent to guard the Patriot battery's outer perimeter in case follow-on attacks came, leaving only three people in the command trailer, the lieutenant and two enlisted soldiers.

Elsewhere that same night, further north over Iraq, British Flight Lieutenant Kevin Main turned around his Tornado GR4A fighter jet and headed back toward Kuwait, his mission for the day complete. In the back seat as navigator was Flight Lieutenant Dave Williams. What Main and Williams didn't know as they rocketed back toward friendly lines was that a crucial piece of equipment, the identification friend or foe (IFF) signal, wasn't on. The IFF was supposed to broadcast a signal to other friendly aircraft and ground radars to let them know their Tornado was friendly and not to fire. But the IFF wasn't working. The reason why is still mysterious. It could be because Main and Williams turned it off while over Iraqi territory so as not to give away their position and forgot to turn it back on when returning to Kuwait. It could be because the system simply broke, possibly from a power supply failure. The IFF signal had been tested by maintenance personnel prior to the aircraft taking off, so it should have been functional, but for whatever reason it wasn't broadcasting.

As Main and Williams began their descent toward Ali Al Salem air base, the Patriot battery tasked with defending coalition bases in Kuwait sent out a radar signal into the sky, probing for Iraqi missiles. The radar signal bounced off the front of Main and Williams' aircraft and reflected back, where it was received by the Patriot's radar dish. Unfortunately, the Patriot's computer didn't register the radar reflection from the Tornado as an aircraft. Because of the aircraft's descending profile, the Patriot's computer tagged the radar signal as coming from an anti-radiation missile. In the Patriot's command trailer, the humans didn't know that a friendly aircraft was coming in for a landing. Their screen showed a radar-hunting enemy missile homing in on the Patriot battery.

The Patriot operators' mission was to shoot down ballistic missiles, which are different from anti-radiation missiles. It would be hard for a radar to confuse

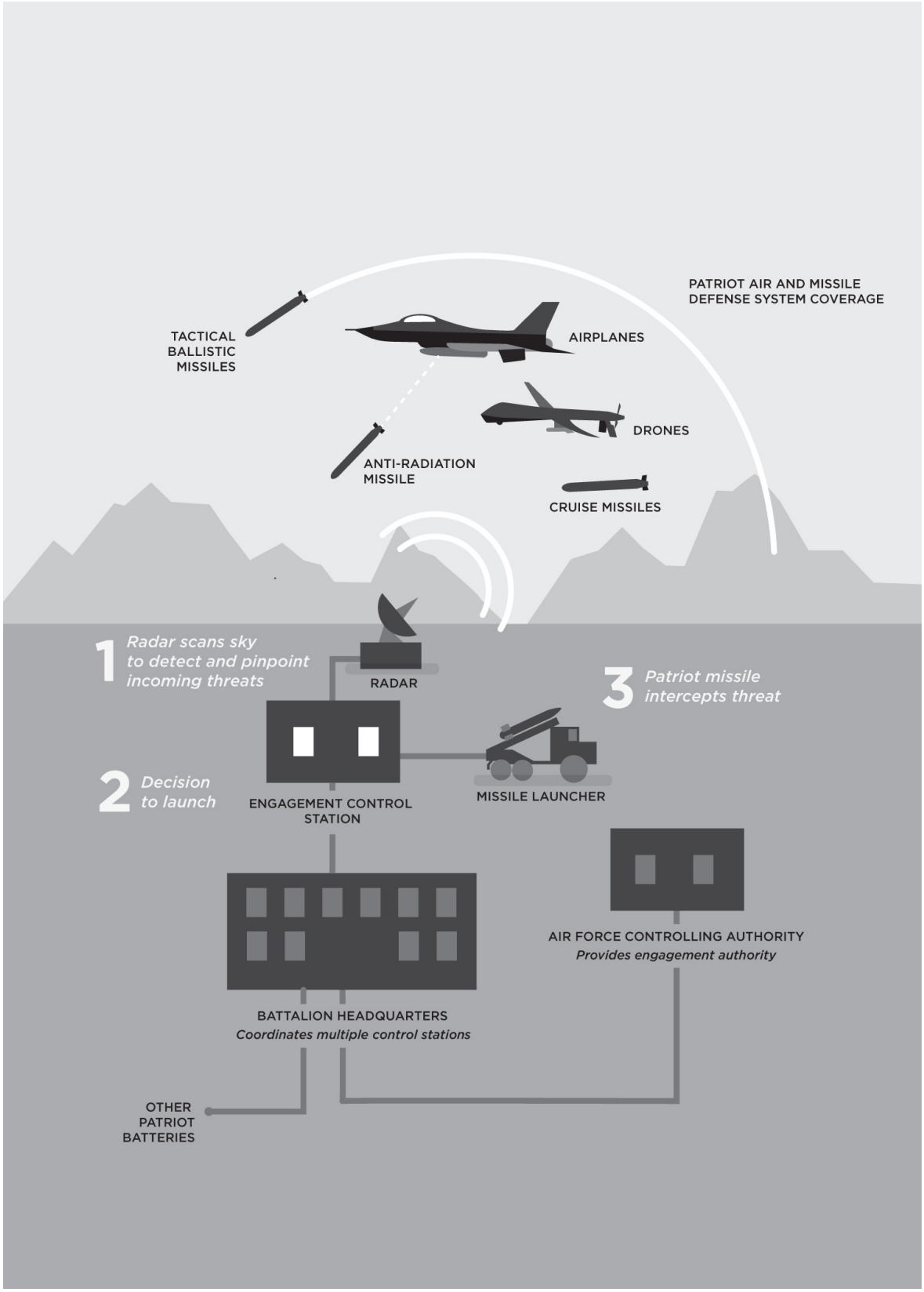
an aircraft flying level with a ballistic missile, which follows a parabolic trajectory through the sky like a baseball. Anti-radiation missiles are different. They have a descending flight profile, like an aircraft coming in on landing. Anti-radiation missiles home on radars and could be deadly to the Patriot. Shooting them wasn't the Patriot operators' primary job, but they were authorized to engage if the missile appeared to be homing in on their radar.

The Patriot operators saw the missile headed toward their radar and weighed their decision. The Patriot battery was operating alone, without the ability to connect to other radars on the network because of their outdated equipment. Deprived of the ability to see other radar inputs directly, the lieutenant called over the radio to the other Patriot units. Did they see an anti-radiation missile? No one else saw it, but this meant little, since other radars may not have been in a position to see it. The Tornado's IFF signal, which would have identified the blip on their radar as a friendly aircraft, wasn't broadcasting. Even if it had been working, as it turns out, the Patriot wouldn't have been able to see the signal—the codes for the IFF hadn't been loaded into the Patriot's computers. The IFF, which was supposed to be a backup safety measure against friendly fire, was doubly broken.

There were no reports of coalition aircraft in the area. There was nothing at all to indicate that the blip that appeared on their scopes as an anti-radiation missile might, in fact, be a friendly aircraft. They had seconds to decide.

They took the shot. The missile disappeared from their scope. It was a hit. Their shift ended. Another successful day.

Elsewhere, Main and Williams' wingman landed in Kuwait, but Main and Williams never returned. The call went out: there is a missing Tornado aircraft. As the sun came up over the desert, people began to put two and two together. The Patriot had shot down one of their own.



U.S. Army Patriot Operations *The Patriot air and missile defense system is used to counter a range of*

threats from enemy aircraft and missiles.

The Army opened an investigation, but there was still a war to fight. The lieutenant stayed at her post; she had a job to do. The Army needed her to do that job, to protect other soldiers from Saddam's missiles. Confusion and chaos are unfortunate realities of war. Unless the investigation determined that she was negligent, the Army needed her in the fight. More of Saddam's missiles were coming.

The very next night, another enemy ballistic missile popped up on their scope. They took the shot. Success. It was a clean hit—another enemy ballistic missile down. The same Patriot battery had two more successful ballistic missile shootdowns before the end of the war. In all, they were responsible for 45 percent of all successful ballistic missile engagements in the war. Later, the investigation cleared the lieutenant of wrongdoing. She made the best call with the information she had.

Other Patriot units were fighting their own struggle against the fog of war. The day after the Tornado shoot down, a different Patriot unit got into a friendly fire engagement with a U.S. F-16 aircraft flying south of Najaf in Iraq. This time, the aircraft shot first. The F-16 fired off a radar-hunting AGM-88 high-speed anti-radiation missile. The missile zeroed in on the Patriot's radar and knocked it out of commission. The Patriot crew was unharmed—a near miss.

After these incidents, a number of safety measures were immediately put in place to prevent further fratricides. The Patriot has both a manual (semiautonomous) and auto-fire (supervised autonomous) mode, which can be kept at different settings for different threats. In manual mode, a human is required to approve an engagement before the system will launch. In auto-fire mode, if there is an incoming threat that meets its target parameters, the system will automatically engage the threat on its own.

Because ballistic missiles often afford very little reaction time before impact, Patriots sometimes operated in auto-fire mode for tactical ballistic missiles. Now that the Army knew the Patriot might misidentify a friendly aircraft as an anti-radiation missile, however, they ordered Patriot units to operate in manual mode for anti-radiation missiles. As an additional safety, systems were now kept in "standby" status so they could track targets, but could not fire without a human bringing the system back to "operate" status. Thus, in order to fire on an anti-radiation missile, two steps were needed: bringing the launchers to operate status *and* authorizing the system to fire on the target. Ideally, this would prevent another fratricide like the Tornado shootdown.

Despite these precautions, a little over a week later on April 2, disaster struck

again. A Patriot unit operating north of Kuwait on the road to Baghdad picked up an inbound ballistic missile. Shooting down ballistic missiles was their job. Unlike the anti-radiation missile that the earlier Patriot unit had fired on—which turned out to be a Tornado—there was no evidence to suggest ballistic missiles might be misidentified as aircraft.

OBSERVE	ORIENT	DECIDE	ACT
What is it? Whose is it?	Is it hostile? Is it a valid target?	Engage?	System fires and missile maneuvers to target
Radar detects and classifies object	Establish situational awareness	Decision whether or not to fire	Human operator can choose to abort missile while in flight
Humans apply outside information and context	Apply rules of engagement	**Manual mode (semiautonomous):** Human operator must authorize engagement or system will not fire	
		Auto-fire mode (supervised autonomous): System will fire unless human operator halts engagement	

Patriot Decision-Making Process The OODA *decision-making* process for a Patriot system. In manual mode, the human operator must take a positive action in order for the system to fire. In *auto-fire* mode, the human supervises the system and can intervene if necessary, but the system will fire on its own if the human does not intervene. *Auto-fire* mode is vital for defending against *short-warning* attacks where there may be little time to make a decision before impact. In both modes, the human can still abort the missile while in flight.

What the operators didn't know—what they could not have known—was that there was no missile. There wasn't even an aircraft misidentified as a missile. There was nothing. The radar track was false, a "ghost track" likely caused by electromagnetic interference between their radar and another nearby

Patriot radar. The Patriot units supporting the U.S. advance north to Baghdad were operating in a nonstandard configuration. Units were spread in a line south-to-north along the main highway to Baghdad instead of the usual widely distributed pattern they would adopt to cover an area. This may have caused radars to overlap and interfere.

But the operators in the Patriot trailer didn't know this. All they saw was a ballistic missile headed their way. In response, the commander ordered the battery to bring its launchers from "standby" to "operate."

The unit was operating in manual mode for anti-radiation missiles, but auto-fire mode for ballistic missiles. As soon as the launcher became operational, the auto-fire system engaged: BOOM-BOOM. Two PAC-3 missiles launched automatically.

The two PAC-3 missiles steered toward the incoming ballistic missile, or at least to the spot where the ground-based radar told them it should be. The missiles activated their seekers to look for the incoming ballistic missile, but there was no missile.

Tragically, the missiles' seekers did find something: a U.S. Navy F/A-18C Hornet fighter jet nearby. The jet was piloted by Lieutenant Nathan White, who was simply in the wrong place at the wrong time. White's F-18 was squawking IFF and he showed up on the Patriot's radar as an aircraft. It didn't matter. The PAC-3 missiles locked onto White's aircraft. White saw the missiles coming and called it out over the radio. He took evasive action, but there was nothing he could do. Seconds later, both missiles struck his aircraft, killing him instantly.

ASSESSING THE PATRIOT'S PERFORMANCE

The Patriot fratricides are an example of the risks of operating complex, highly automated lethal systems. In a strict operational sense, the Patriot units accomplished their mission. Over sixty Patriot fire units were deployed during the initial phase of the war, forty from the United States and twenty-two from four coalition nations. Their mission was to protect ground troops from Iraqi ballistic missiles, which they did. Nine Iraqi ballistic missiles were fired at coalition forces; all were successfully engaged by Patriots. No coalition troops were harmed from Iraqi missiles. A Defense Science Board Task Force on the Patriot's performance concluded that, with respect to missile defense, the Patriot was a "substantial success."

On the other hand, in addition to these nine successful engagements, Patriots

were involved in three fratricides: two incidents in which Patriots shot down friendly aircraft, killing the pilots, and a third incident in which an F-16 fired on a Patriot. Thus, of the twelve total engagements involving Patriots, 25 percent were fratricides, an "unacceptable" fratricide rate according to Army investigators.

The reasons for the Patriot fratricides were a complex mix of human error, improper testing, poor training, and unforeseen interactions on the battlefield. Some problems were known—IFF was well understood to be an imperfect solution for preventing fratricides. Other problems, such as the potential for the Patriot to misclassify an aircraft as an anti-radiation missile, had been identified during operational testing but had not been corrected and were not included in operator training. Still other issues, such as the potential for electromagnetic interference to cause a false radar track, were novel and unexpected. Some of these complications were preventable, but others were not. War entails uncertainty. Even the best training and operational testing can only approximate the actual conditions of war. Inevitably, soldiers will face wartime conditions where the environment, adversary innovation, and simply the chaos, confusion, and violence of war all contribute to unexpected challenges. Many things that seem simple in training often look far different in the maw of combat.

One thing that did not happen and was not a cause of the Patriot fratricides is that the Patriot system did not fail, per se. It didn't break. It didn't blow a fuse. The system performed its function: it tracked incoming targets and, when authorized, shot them down. Also, in both instances a human was required to give the command to fire or at least to bring the launchers to operate. When this lethal, highly automated system was placed in the hands of operators who did not fully understand its capabilities and limitations, however, it turned deadly. Not because the operators were negligent. No one was found to be at fault in either incident. It would be overly simplistic to blame the fratricides on "human error." Instead, what happened was more insidious. Army investigators determined the Patriot community had a culture of "trusting the system without question." According to Army researchers, the Patriot operators, while nominally in control, exhibited automation bias: an "unwarranted and uncritical trust in automation. In essence, control responsibility is ceded to the machine." There may have been a human "in the loop," but the human operators didn't question the machine when they should have. They didn't exercise the kind of judgment Stanislav Petrov did when he questioned the signals his system was giving him regarding a false launch of U.S. nuclear missiles. The Patriot operators trusted the machine, and it was wrong.

ROBUTOPIA VS. ROBOPOCALYPSE

We have two intuitions when it comes to autonomous systems, intuitions that come partly from science fiction but also from our everyday experiences with phones, computers, cars, and myriad other computerized devices.

The first intuition is that autonomous systems are reliable and introduce greater precision. Just as autopilots have improved air travel safety, automation can also improve safety and reliability in many other domains. Humans are terrible drivers, for example, killing more than 30,000 people a year in the United States alone (roughly the equivalent of a 9/11 attack every month). Even without fully autonomous cars, more advanced vehicle autopilots that allow cars to drive themselves under most conditions could dramatically improve safety and save lives.

However, we have another instinct when it comes to autonomous systems, and that is one of robots run amok, autonomous systems that slip out of human control and result in disastrous outcomes. These fears are fed to us through a steady diet of dystopian science fiction stories in which murderous AIs turn on humans, from *2001: A Space Odyssey*'s HAL 9000 to *Ex Machina*'s Ava. But these intuitions also come from our everyday experiences with simple automated devices. Anyone who has ever been frustrated with an automated telephone call support helpline, an alarm clock mistakenly set to "p.m." instead of "a.m.," or any of the countless frustrations that come with interacting with computers, has experienced the problem of "brittleness" that plagues automated systems. Autonomous systems will do precisely what they are programmed to do, and it is this quality that makes them both reliable and maddening, depending on whether what they were programmed to do was the right thing at that point in time.

Both of our intuitions are correct. With proper design, testing, and use, autonomous systems can often perform tasks far better than humans. They can be faster, more reliable, and more precise. However, if they are placed into situations for which they were not designed, if they aren't fully tested, if operators aren't properly trained, or if the environment changes, then autonomous systems can fail. When they do fail, they often fail badly. Unlike humans, autonomous systems lack the ability to step outside their instructions and employ "common sense" to adapt to the situation at hand.

This problem of brittleness was highlighted during a telling moment in the 2011 *Jeopardy! Challenge* in which IBM's Watson AI took on human Jeopardy champions Ken Jennings and Brad Rutter. Toward the end of the first game, Watson momentarily stumbled in its rout of Jennings and Rutter in response to a

clue in the “Name the Decade” category. The clue was, “The first modern crossword is published and Oreo cookies are introduced.” Jennings rang in first with the answer, “What are the 20s?” Wrong, said host Alex Trebek. Immediately afterward, Watson rang in and gave the same answer, “What is 1920s?” A befuddled Trebek testily replied, “No, Ken said that.”

I’m not particularly good at Jeopardy, but even I knew, “What are the 1920s?” was the wrong answer once Jennings guessed wrong. (The correct answer is the 1910s.) Watson hadn’t been programmed to listen to other contestants’ wrong answers and adjust accordingly, however. Processing Jennings’ answer was outside of the bounds of Watson’s design. Watson was superb at answering Jeopardy questions under most conditions, but its design was brittle. When an atypical event occurred that Watson’s design didn’t account for, such as Ken Jennings getting a wrong answer, Watson couldn’t adapt on the fly. As a result, Watson’s performance suddenly plummeted from superhuman to super-dumb.

Brittleness can be managed when the person using an autonomous system understands the boundaries of the system—what it can and cannot do. The user can either steer the system away from situations outside the bounds of its design or knowingly account for and accept the risks of failure. In this case, Watson’s designers understood this limitation. They just didn’t think that the ability to learn from other contestants’ wrong answers would be important. “We just didn’t think it would ever happen,” one of Watson’s programmers said afterward. Watson’s programmers were probably right to discount the importance of this ability. The momentary stumble proved inconsequential. Watson handily defeated its human counterparts.

Problems can arise when human users don’t anticipate these moments of brittleness, however. This was the case with the Patriot fratricides. The system had vulnerabilities—misclassifying an anti-radiation missile as an aircraft, IFF failures, and electromagnetic interference causing “ghost track” ballistic missiles—that the human operators were either unaware of or didn’t sufficiently account for. As a result, the human user’s expectations and the system’s actual behavior were not in alignment. The operators thought the system was targeting missiles when it was actually targeting aircraft.

AUTONOMY AND RISK

One of the ways to compensate for the brittle nature of automated systems is to

retain tight control over their operation. If the system fails, humans can rapidly intervene to correct it or halt its operation. Tighter human control reduces the autonomy, or freedom, of the machine.

Immediate human intervention is possible for semiautonomous and human-supervised systems. Just because humans can intervene, however, doesn't mean they always do so when they should. In the case of the Patriot fratricides, humans were "in the loop," but they didn't sufficiently question the automation. Humans can't act as an independent fail-safe if they cede their judgment to the machine.

Effective human intervention may be even more challenging in supervised autonomous systems, where the system does not pause to wait for human input. The human's ability to actually regain control of the system in real time depends heavily on the speed of operations, the amount of information available to the human, and any time delays between the human's actions and the system's response. Giving a driver the ability to grab the wheel of an autonomous vehicle traveling at highway speeds in dense traffic, for example, is merely the illusion of control, particularly if the operator is not paying attention. This appears to have been the case in a 2016 fatality involving a Tesla Model S that crashed while driving on autopilot.

For fully autonomous systems, the human is out of the loop and cannot intervene at all, at least for some period of time. This means that if the system fails or the context changes, the result could be a runaway autonomous process beyond human control with no ability to halt or correct it.

This danger of autonomous systems is best illustrated not with a science fiction story, but with a Disney cartoon. *The Sorcerer's Apprentice* is an animated short in Disney's 1940 film *Fantasia.* In the story, which is an adaptation of an eighteenth-century poem by Goethe, Mickey Mouse plays the apprentice of an old sorcerer. When the sorcerer leaves for the day, Mickey decides to use his novice magic to automate his chores. Mickey enchants a broomstick, causing it to sprout arms and come to life. Mickey commands the broomstick to carry pails of water from the well to a cistern, a chore Mickey is supposed to be doing. Soon, Mickey is nodding off, his chores automated.

As Mickey sleeps, the cistern overfills. The job is done, but no one told the broomstick to stop. Mickey wakes to find the room flooded and the broomstick fetching more water. He commands the broomstick to halt, but it doesn't comply. Desperate, Mickey snatches an axe from the wall and chops the broomstick to pieces, but the splinters reanimate into a horde of broomsticks. They march forth to bring even more water, an army of rogue autonomous

agents out of control. Finally, the madness is stopped only by the return of the sorcerer himself, who disperses the water and halts the broomsticks with a wave of his arms.

With the original German poem written in 1797, *The Sorcerer's Apprentice* may be the first example of autonomy displacing jobs. It also shows the danger of automation. An autonomous system may not perform a task in the manner we want. This could occur for a variety of reasons: malfunction, user error, unanticipated interactions with the environment, or hacking. In the case of Mickey's problem, the "software" (instructions) that he bewitched the broomstick with were flawed because he didn't specify when to stop. Overfilling the cistern might have been only a minor annoyance if it had happened once, however. A semiautonomous process that paused for Mickey's authorization after each trip to the well would have been far safer. Having a human "in the loop" would have mitigated the danger from the faulty software design. The human can act like a circuit breaker, catching harmful events before they cascade out of control. Making the broomstick fully autonomous without a human in the loop wasn't the cause of the failure, but it did dramatically increase the consequences if something went wrong. Because of this potential to have a runaway process, fully autonomous systems are inherently more hazardous than semiautonomous ones.

Putting an autonomous system into operation means accepting the risk that it may perform its task incorrectly. Fully autonomous systems are not necessarily more likely to fail than semiautonomous or supervised autonomous ones, but if they do, the consequences—the potential damage caused by the system—could be severe.

TRUST, BUT VERIFY

Activating an autonomous system is an act of trust. The user trusts that the system will function in the manner that he or she expects. Trust isn't blind faith, however. As the Patriot fratricides demonstrated, too much trust can be just as dangerous as too little. Human users need to trust the system just the right amount. They need to understand both the capabilities *and* limitations of the system. This is why Bradford Tousley from DARPA TTO cited test and evaluation as his number one concern. A rigorous testing regime can help designers and operators better understand how the system performs under realistic conditions. Bob Work similarly told me that test and evaluation was

“central” to building trustworthy autonomous systems. “When you delegate authority to a machine, it’s got to be repeatable,” he said. “The same outcome has to happen over and over and over again. . . . So, what is going to be our test and evaluation regime for these smarter and smarter weapons to make sure that the weapon stays within the parameters of what we expect it to do? That’s an issue.”

The problem is that, even with simulations that test millions of scenarios, fully testing all of the possible scenarios a complex autonomous system might encounter is effectively impossible. There are simply too many possible interactions between the system and its environment and even within the system itself. Mickey should have been able to anticipate the cistern overflowing, but some real-world problems cannot be anticipated. The game of Go has more possible positions than atoms in the universe, and the real world is far more complex than Go. A 2015 Air Force report on autonomy bemoaned the problem:

> Traditional methods . . . fail to address the complexities associated with autonomy software . . . There are simply too many possible states and combination of states to be able to exhaustively test each one.

In addition to the sheer numerical problem of evaluating all possible combinations, testing is also limited by the testers’ imagination. In games like chess or *go*, the set of possible actions is limited. In the real world, however, autonomous systems will encounter any number of novel situations: new kinds of human error, unexpected environmental conditions, or creative actions by adversaries looking to exploit vulnerabilities. If these scenarios can’t be anticipated, they can’t be tested.

Testing is vital to building confidence in how autonomous systems will behave in real-world environments, but no amount of testing can entirely eliminate the potential for unanticipated behaviors. Sometimes these unanticipated behaviors may pleasantly surprise users, like AlphaGo’s 1 in 10,000 move that stunned human champion Lee Sedol. Sometimes these unanticipated actions can be negative. During Gary Kasparov’s first game against Deep Blue in 1997, a bug in Deep Blue caused it to make a nonsense random move in the forty-fourth move of the game. One of Deep Blue’s programmers later explained, “We had seen it once before, in a test game played earlier in 1997, and thought it was fixed. Unfortunately, there was one case that we had missed.” When playing games like Jeopardy, chess, or *go*, surprising behaviors may be tolerable, even interesting flukes. When operating high-risk automated systems where life or death is at stake, unexpected actions can lead to tragic accidents, such as the Patriot fratricides.

WHEN ACCIDENTS ARE NORMAL

To better understand the risks of autonomous weapons, I spoke with John Borrie from the UN Institute for Disarmament Research (UNIDIR). UNIDIR is an independent research institute within the United Nations that focuses on arms control and disarmament issues. Borrie authored a recent UNIDIR report on autonomous weapons and risk and he's worked extensively on arms control and disarmament issues in a variety of capacities—for the New Zealand government, the International Committee of the Red Cross, and UNIDIR—and on a host of technologies: cryptography, chemical and biological weapons, and autonomy. This made him well positioned to understand the relative risks of autonomous weapons.

Borrie and I sat down on the sidelines of the UN talks on autonomous weapons in Geneva in 2016. Borrie is not an advocate for a preemptive ban on autonomous weapons and in general has the sober demeanor of a professor, not a firebrand activist. He speaks passionately (though in an even-tempered, professorial cadence) in his lilting New Zealand accent. I could imagine myself pleasantly nodding off in his class, even as he calmly warned of the dangers of robots run amok.

"With very complex technological systems that are hazardous," Borrie said, "—and I think autonomous weapons fall into that category of hazard because of their intended lethality . . . we have difficulty [saying] that we can remove the risk of unintentional lethal effects." Borrie compared autonomous weapons to complex systems in other industries. Humans have decades of experience designing, testing, and operating complex systems for high-risk applications, from nuclear power plants to commercial airliners to spacecraft. The good news is that because of these experiences, there is a robust field of research on how to improve safety and resiliency in these systems. The bad news is that all of the experience with complex systems to date suggests that 100 percent error-free operation is impossible. In sufficiently complex systems, it is impossible to test every possible system state and combination of states; some unanticipated interactions will happen. Failures may be unlikely, but over a long enough timeline they are inevitable. Engineers refer to these incidents as "normal accidents" because their occurrence is inevitable, even normal, in complex systems. "Why would autonomous systems be any different?" Borrie asked.

The textbook example of a normal accident is the Three Mile Island nuclear power plant meltdown in 1979. The Three Mile Island incident was a "system failure," meaning that the accident was caused by the interaction of many small,

individually manageable failures interacting in an unexpected and dramatic way, much like the Patriot fratricides. The Three Mile Island incident illustrates the challenge in anticipating and preventing accidents in complex systems.

The trouble began when moisture from a leaky seal got into an unrelated system, causing it to shut off water pumps vital to cooling the reactor. An automated safety kicked in, activating emergency pumps, but a valve needed to allow water to flow through the emergency cooling system had been left closed. Human operators monitoring the reactor were unaware that the valve was shut because the indicator light on their control panel was obscured by a repair tag for another, unrelated system.

Without water, the reactor core temperature rose. The reactor automatically "scrammed," dropping graphite control rods into the reactor core to absorb neutrons and stop the chain reaction. However, the core was still generating heat. Rising temperatures activated another automatic safety, a pressure release valve designed to let off steam before the rising pressure cracked the containment vessel.

The valve opened as intended but failed to close. Moreover, the valve's indicator light also failed, so the plant's operators did not know the valve was stuck open. Too much steam was released and water levels in the reactor core fell to dangerous levels. Because water was crucial to cooling the still-hot nuclear core, another automatic emergency water cooling system kicked in and the plant's operators also activated an additional emergency cooling system.

What made these failures catastrophic was the fact that that nuclear reactors are *tightly coupled*, as are many other complex machines. Tight coupling is when an interaction in one component of the system directly and immediately affects components elsewhere. There is very little "slack" in the system—little time or flexibility for humans to intervene and exercise judgment, bend or break rules, or alter the system's behavior. In the case of Three Mile Island, the sequence of failures that caused the initial accident happened within a mere thirteen seconds.

It is the combination of complexity *and* tight coupling that makes accidents an expected, if infrequent, occurrence in such systems. In loosely coupled complex systems, such as bureaucracies or other human organizations, there is sufficient slack for humans to adjust to unexpected situations and manage failures. In tightly coupled systems, however, failures can rapidly cascade from one subsystem to the next and minor problems can quickly lead to system breakdown.

As events unfolded at Three Mile Island, human operators reacted quickly and automatic safeties kicked in. In their responses, though, we see the

limitations of both humans and automatic safeties. The automatic safeties were useful, but did not fully address the root causes of the problems—a water cooling valve that was closed when it should have been open and a pressure-release valve that was stuck open when it should have been closed. In principle, "smarter" safeties that took into account more variables could have addressed these issues. Indeed, nuclear reactor safety has improved considerably since Three Mile Island.

The human operators faced a different problem, though, one which more sophisticated automation actually makes harder, not easier: the incomprehensibility of the system. Because the human operators could not directly inspect the internal functioning of the reactor core, they had to rely on indicators to tell them what was occurring. But these indicators were also susceptible to failure. Some indicators did fail, leaving human operators with a substantial deficit of information about the system's internal state. The operators did not discover that the water cooling valve was improperly closed until eight minutes into the accident and did not discover that the pressure release valve was stuck open until two hours later. This meant that some of the corrective actions they took were, in retrospect, incorrect. It would be improper to call their actions "human error," however. They were operating with the best information they had at the time.

The father of normal accident theory, Charles Perrow, points out that the "incomprehensibility" of complex systems themselves is a stumbling block to predicting and managing normal accidents. The system is so complex that it is incomprehensible, or opaque, to users and even the system's designers. This problem is exacerbated in situations in which humans cannot directly inspect the system, such as a nuclear reactor, but also exists in situations where humans are physically present. During the *Apollo 13* disaster, it took seventeen minutes for the astronauts and NASA ground control to uncover the source of the instrument anomalies they were seeing, in spite of the fact that the astronauts were on board the craft and could "feel" how the spacecraft was performing. The astronauts heard a bang and felt a small jolt from the initial explosion in the oxygen tank and could tell that they had trouble controlling the attitude (orientation) of the craft. Nevertheless, the system was so complex that vital time was lost as the astronauts and ground-control experts pored over the various instrument readings and rapidly-cascading electrical failures before they discovered the root cause.

Failures are inevitable in complex, tightly coupled systems and the sheer complexity of the system inhibits predicting when and how failures are likely to

occur. John Borrie argued that autonomous weapons would have the same characteristics of complexity and tight coupling, making them susceptible to "failures . . . we hadn't anticipated." Viewed from the perspective of normal accident theory, the Patriot fratricides were not surprising—they were inevitable.

THE INEVITABILITY OF ACCIDENTS

The *Apollo 13* and Three Mile Island incidents occurred in the 1970s, when engineers where still learning to manage complex, tightly coupled systems. Since then, both nuclear power and space travel have become safer and more reliable—even if they can never be made entirely safe.

NASA has seen additional tragic accidents, including some that were not recoverable as *Apollo 13* was. These include the loss of the space shuttles *Challenger* (1986) and *Columbia* (2003) and their crews. While these accidents had discrete causes that could be addressed in later designs (faulty O-rings and falling foam insulation, respectively), the impossibility of anticipating such specific failures in advance makes continued accidents inevitable. In 2015, for example, the private company SpaceX had a rocket blow up on the launch pad due to a strut failure that had not been previously identified as a risk. A year later, another SpaceX rocket exploded during testing due to a problem with supercooled oxygen that CEO Elon Musk said had "never been encountered before in the history of rocketry."

Nuclear power has grown significantly safer since Three Mile Island, but the 2011 meltdown of the Japanese Fukushima Daiichi nuclear plant points to the limits of safety. Fukushima Daiichi was hardened against earthquakes and flooding, with backup generators and thirty-foot-high floodwalls. Unfortunately, the plant was not prepared for a 9.0 magnitude earthquake (the largest recorded earthquake to ever hit Japan) off the coast that caused both a loss in power *and* a massive forty-foot-high tsunami. Many safeties worked. The earthquake did not damage the containment vessels. When the earthquake knocked out primary power, the reactors automatically scrammed, inserting control rods to stop the nuclear reaction. Backup diesel generators automatically came online.

However, the forty-foot-high tsunami wave topped the thirty-foot-high floodwalls, swamping twelve of thirteen backup diesel generators. Combined with the loss of primary power from the electrical grid, the plant lost the ability to pump water to cool the still-hot reactor cores. Despite the heroic efforts of Japanese engineers to bring in additional generators and pump water into the

overheating reactors, the result was the worst nuclear power accident since Chernobyl.

The problem wasn't that Fukushima Daiichi lacked backup safeties. The problem was a failure to anticipate an unusual environmental condition (a massive earthquake off the coast that induced a tsunami) that caused what engineers call a *common-mode* failure—one that simultaneously overwhelmed two seemingly independent safeties: primary and backup power. Even in fields where safety is a central concern, such as space travel and nuclear power, anticipating all of the possible interactions of the system and its environment is effectively impossible.

"BOTH SIDES HAVE STRENGTHS AND WEAKNESSES"

Automation plays a mixed role in accidents. Sometimes the brittleness and inflexibility of automation can cause accidents. In other situations, automation can help reduce the probability of accidents or mitigate their damage. At Fukushima Daiichi, automated safeties scrammed the reactor and brought backup generators online. Is more automation a good or bad thing?

Professor William Kennedy of George Mason University has extensive experience in nuclear reactors and military hardware. Kennedy has a unique background—thirty years in the Navy (active and reserve) on nuclear missile submarines, combined with twenty-five years working for the Nuclear Regulatory Commission and the Department of Energy on nuclear reactor safety. To top it off, he has a PhD in information technology with a focus on artificial intelligence. I asked him to help me understand the benefits of humans versus AI in managing high-risk systems.

"A significant message for the Nuclear Regulatory Commission from Three Mile Island was that humans were not omnipotent," Kennedy said. "The solution prior to Three Mile Island was that every time there was a design weakness or a feature that needed to be processed was to give the operator another gauge, another switch, another valve to operate remotely from the control room and everything would be fine. And Three Mile Island demonstrated that humans make mistakes. . . . We got to the point where we had over 2,000 alarms in the control rooms, a wall of procedures for each individual alarm. And Three Mile Island said that alarms almost never occur individually." This was an unmanageable level of complexity for any human operator to absorb, Kennedy explained.

Following Three Mile Island, more automation was introduced to manage some of these processes. Kennedy supports this approach, to a point. “The automated systems, as they are currently designed and built, may be more reliable than humans for planned emergencies, or known emergencies. . . . If we can study it in advance and lay out all of the possibilities and in our nice quiet offices consider all the ways things can behave, we can build that into a system and it can reliably do what we say. But we don’t always know what things are possible. . . . Machines can repeatedly, quite reliably, do planned actions. . . . But having the human there provides for ‘beyond design basis’ accidents or events.” In other words, automation could help for situations that could be predicted, but humans were needed to manage novel situations. “Both sides have strengths and weaknesses,” Kennedy explained. “They need to work together, at the moment, to provide the most reliable system.”

AUTOMATION AND COMPLEXITY—A DOUBLE-EDGED SWORD

Kennedy’s argument tracks with what we have seen in modern machines—increasing software and automation but with humans still involved at some level. Modern jetliners effectively fly themselves, with pilots functioning largely as an emergency backup. Modern automobiles still have human drivers, but have a host of automated or autonomous features to improve driving safety and comfort: antilock brakes, traction and stability control, automatic lane keeping, intelligent cruise control, collision avoidance, and self-parking. Even modern fighter jets use software to help improve safety and reliability. F-16 fighter aircraft have been upgraded with automatic ground collision avoidance systems. The newer F-35 fighter reportedly has software-based limits on its flight controls to prevent pilots from putting the aircraft into unrecoverable spins or other aerodynamically unstable conditions.

The double-edged sword to this automation is that all of this added software increases complexity, which can itself introduce new problems. Sophisticated automation requires software with millions of lines of code: 1.7 million for the F-22 fighter jet, 24 million for the F-35 jet, and some 100 million lines of code for a modern luxury automobile. Longer pieces of software are harder to verify as being free from bugs or glitches. Studies have pegged the software industry average error rate at fifteen to fifty errors per 1,000 lines of code. Rigorous internal test and evaluation has been able to reduce the error rate to 0.1 to 0.5

errors per 1,000 lines of code in some cases. However, in systems with millions of lines of code, some errors are inevitable. If they aren't caught in testing, they can cause accidents if encountered during real world operations.

On their first deployment to the Pacific in 2007, eight F-22 fighter jets experienced a Y2K-like total computer meltdown when crossing the International Date Line. All onboard computer systems crashed, causing the pilots to lose navigation, fuel subsystems, and some communications. Stranded over the Pacific without a navigational reference point, the aircraft were able to make it back to land by following the tanker aircraft accompanying them, which relied on an older computer system. Under tougher circumstances, such as combat or even bad weather, the incident could have led to a catastrophic loss of the aircraft. While the existence of the International Date Line clearly could be anticipated, the interaction of the dateline with the software was not identified in testing.

Software vulnerabilities can also leave open opportunities for hackers. In 2015, two hackers revealed that they had discovered vulnerabilities that allowed them to remotely hack certain automobiles while they were on the road. This allowed them to take control of critical driving components including the transmission, steering column, and brakes. In future self-driving cars, hackers who gain access could simply change the car's destination.

Even if software does not have specific bugs or vulnerabilities, the sheer complexity of modern machines can make it challenging for users to understand what the automation is doing and why. When humans are no longer interacting with simple mechanical systems that may behave predictably but instead are interacting with complex pieces of software with millions of lines of code, the human user's expectation about what the automation will do may diverge significantly from what it actually does. I found this to be a challenge with the Nest thermostat, which doesn't have millions of lines of code. (A study of Nest users found similar frustrations, so apparently I am not uniquely unqualified in predicting Nest behavior.)

More advanced autonomous systems are often able to account for more variables. As a result, they can handle more complex or ambiguous environments, making them more valuable than simpler systems. They may fail less overall, because they can handle a wider range of situations. However, they will still fail sometimes and because they are more complex, accurately predicting *when* they will fail may be more difficult. Borrie said, "As systems get increasingly complex and increasingly self-directed, I think it's going to get more and more difficult for human beings to be able to think ahead of time what

those weak points are necessarily going to be." When this happens in high-risk situations, the result can be catastrophic.

"WE DON'T UNDERSTAND ANYTHING!"

On June 1, 2009, Air France Flight 447 from Rio to Paris ran into trouble midway over the Atlantic Ocean. The incident began with a minor and insignificant instrumentation failure. Air speed probes on the wings froze due to ice crystals, a rare but non-serious problem that did not affect the flight of the aircraft. Because the airspeed indicators were no longer functioning properly, the autopilot disengaged and handed over control back to the pilots. The plane also entered a different software mode for flight controls. Instead of flying under "normal law" mode, where software limitations prevent pilots from putting the plane into dangerous aerodynamic conditions such as stalls, the plane entered "alternate law" mode, where the software limitations are relaxed and the pilots have more direct control over the plane.

Nevertheless, there was no actual emergency. Eleven seconds following the autopilot disengagement, the pilots correctly identified that they had lost the airspeed indicators. The aircraft was flying normally, at appropriate speeds and full altitude. Everything was fine.

Inexplicably, however, the pilots began a series of errors that resulted in a stall, causing the aircraft to crash into the ocean. Throughout the incident, the pilots continually misinterpreted data from the airplane and misunderstood the aircraft's behavior. At one point mid-crisis, the copilot exclaimed, "We completely lost control of the airplane and we don't understand anything! We tried everything!" The problem was actually simple. The pilots had pulled back too far on the stick, causing the aircraft to stall and lose lift. This is a basic aerodynamic concept, but poor user interfaces and opaque automated processes on the aircraft, even while flown manually, contributed to the pilots' lack of understanding. The complexity of the aircraft created problems of transparency that would likely not have existed on a simpler aircraft. By the time the senior pilot understood what was happening, it was too late. The plane was too low and descending too rapidly to recover. The plane crashed into the ocean, killing all 228 people on board.

Unlike in the F-22 International Date Line incident or the automobile hack, the Air France Flight 447 crash was not due to a hidden vulnerability lurking within the software. In fact, the automation performed perfectly. However, it

would be overly simplistic to lay the crash at the feet of human error. Certainly the pilots made mistakes, but the problem is best characterized as human-automation failure. The pilots were confused by the automation and the complexity of the system.

THE PATRIOT FRATRICIDES AS NORMAL ACCIDENTS

Normal accident theory sheds light on the Patriot fratricides. They weren't merely freak occurrences, unlikely to be repeated. Instead, they were a normal consequence of operating a highly lethal, complex, tightly coupled system. True to normal accidents, the specific chain of events that led to each fratricide was unlikely. Multiple failures happened simultaneously. However, simply because these specific combinations of failures were unlikely does not mean that probability of accidents as a whole was low. In fact, given the degree of operational use, the probability of there being some kind of accident was quite high. Over sixty Patriot batteries were deployed to Operation Iraqi Freedom, and during the initial phase of the war coalition aircraft flew 41,000 sorties. This means that the number of possible Patriot-aircraft interactions were in the millions. As the Defense Science Board Task Force on the Patriot pointed out, given the sheer number of interactions, "even very-low-probability failures could result in regrettable fratricide incidents." The fact that the F-18 and Tornado incidents had different causes lends further credence to the view that normal accidents are lurking below the surface in complex systems, waiting to emerge. The complexities of war may bring these vulnerabilities to the surface.

Is it possible to safely operate hazardous complex systems? Normal accident theory says "no." The probability of accidents can be reduced, but never eliminated. There is an alternate point of view on complex systems, however, which suggests that, under certain conditions, normal accidents can largely be avoided.

10

COMMAND AND DECISION

CAN AUTONOMOUS WEAPONS BE USED SAFELY?

There is a robust body of evidence supporting normal accident theory, but a few outliers seem to defy expectations. The Federal Aviation Administration (FAA) air traffic control system and U.S. Navy aircraft carrier flight decks are two examples of "high-reliability organizations." Their rate of accidents isn't zero, but they *are* exceptionally low given the complexities of their operating environment and the hazards of operation. High-reliability organizations can be found across a range of applications and have some common characteristics: highly trained individuals, a collective mindfulness of the risk of failure, and a continued commitment to learn from near misses and improve safety.

While militaries as a whole would not be considered high-reliability organizations, some military communities have very high safety records with complex high-risk systems. In addition to aircraft carrier flight deck operations, the U.S. Navy's submarine community is an example of a high-reliability organization. Following the loss of the USS *Thresher* to an accident in 1963—at the time one of the Navy's most advanced submarines and first in her class—the Navy instituted the Submarine Safety (SUBSAFE) program. Submarine components that are critical for safe operation are designated "SUBSAFE" and subject to rigorous inspection and testing throughout their design, fabrication, maintenance, and use. There is no silver bullet to SUBSAFE's high reliability. It is a continuous process of quality assurance and quality control applied across the entire submarine's life cycle. Upon installation and at every subsequent inspection or repair over the life of the ship, every SUBSAFE component is

checked, double-checked, and checked again against technical specifications. If anything is amiss, it must be corrected or approved by an appropriate authority before the submarine can proceed with operations.

SUBSAFE is not a technological solution to normal accidents. It is a bureaucratic and organizational solution. Nevertheless, the results have been astounding. In 2003 Congressional testimony, Rear Admiral Paul Sullivan, the Navy deputy commander for ship design, integration, and engineering, explained the impact of the program:

> The SUBSAFE Program has been very successful. Between 1915 and 1963, 16 submarines were lost due to non-combat causes, an average of one every three years. Since the inception of the SUBSAFE Program in 1963 . . . We have never lost a SUBSAFE certified submarine.

It is hard to overstate the significance of this safety record. The U.S. Navy has more than seventy submarines in its force, with approximately one-third of them at sea at a time. The U.S. Navy has operated at this pace for over half a century without losing a single submarine. From the perspective of normal accident theory, this should not be possible. Operating a nuclear-powered submarine is extremely complex and inherently hazardous, and yet the Navy has been able to substantially reduce these risks. Accidents resulting in catastrophic loss of a submarine are not “normal” in the U.S. Navy. Indeed, they are unprecedented since the advent of SUBSAFE, making SUBSAFE a shining example of what high-reliability organizations can achieve.

Could high-reliability organizations be a model for how militaries might handle autonomous weapons? In fact, lessons from SUBSAFE and aircraft carrier deck operations have already informed how the Navy operates the Aegis combat system. The Navy describes the Aegis as “a centralized, automated, command-and-control (C2) and weapons control system that was designed as a total weapon system, from detection to kill.” It is the electronic brain of a ship’s weapons. The Aegis connects the ship’s advanced radar with its anti-air, anti-surface, and antisubmarine weapon systems and provides a central control interface for sailors. First fielded in 1983, the Aegis has gone through several upgrades and is now at the core of over eighty U.S. Navy warships. To better understand Aegis and whether it could be a model for safe use of future autonomous weapons, I traveled to Dahlgren, Virginia, where Aegis operators are trained.

THE AEGIS COMBAT SYSTEM

Captain Pete Galluch is commander of the Aegis Training and Readiness Center, where he oversees training for all Aegis-qualified officers and enlisted sailors. The phrase “steely-eyed missile man” comes to mind upon meeting Galluch. He speaks with the calmness and decisiveness of a surgeon, a man who is ready to let missiles fly if need be. I can imagine Galluch standing in the midst of a ship’s combat information center (CIC) in wartime, unflappable in the midst of the chaos, ordering his sailors when to take the shot and when to hold back. If I were flying within range of an Aegis’s weapons or was counting on its ballistic missile defense capabilities to protect my city, I would trust Galluch to make the right call.

Aegis is a weapon system of staggering complexity. At the core of Aegis is a computer called “Command and Decision,” or C&D, which governs the behavior of the radar and weapons. Command and Decision’s actions are governed by a series of statements—essentially programs or algorithms—that the Navy refers to as “doctrine.” Unlike the Patriot circa 2003, however, which had only a handful of different operating modes, Aegis doctrine is almost infinitely customizable.

With respect to weapons engagements, Aegis has four settings. The manual setting, in which engagements against radar “tracks” (objects detected by the radar) must be done directly by a human, involves the most human control. Ship commanders can increase the degree of automation in the engagement process by activating one of three types of doctrine: Semi-Auto, Auto SM, and Auto-Special. Semi-Auto, as the term would imply, automates part of the engagement process to generate a firing solution on a radar track, but final decision authority is withheld by the human operator. Auto SM automates more of the engagement process, but a human must still take a positive action before firing. Despite the term, Auto SM still retains a human in the loop. Auto-Special is the only mode where the human is “on the loop.” Once Auto-Special is activated, the Aegis will automatically fire against threats that meet its parameters. The human can intervene to stop the engagement, but no further authorization is needed to fire.

It would be a mistake to think, however, that this means that Aegis can only operate in four discrete modes. In fact, doctrine statements can mix and match these control types against different threats. For example, one doctrine statement could be written to use Auto SM against one type of threat, such as aircraft. Another doctrine statement might authorize Auto-Special against cruise missiles, for which there may be less warning. These doctrine statements can be applied individually or in packages. “You can mix and match,” Galluch explained. “It’s a very flexible system. . . . we can do all [doctrine statements] with a push of a

button, some with a push of a button, or bring them up individually."

This makes Aegis less like a finished product with a few different modes and more like a customizable system that can be tailored for each mission. Galluch explained that the ship's doctrine review board, consisting of the officers and senior enlisted personnel who work on Aegis, begin the process of writing doctrine months before deployment. They consider their anticipated missions, intelligence assessments, and information on the region for the upcoming deployment, then make recommendations on doctrine to the ship's captain for approval. The result is a series of doctrine statements, individually and in packages, that the captain can activate as needed during deployment. "If you have your doctrine statements built and tested," Galluch said, the time to "bring them up is seconds."

Doctrine statements are typically grouped into two general categories: non-saturation and saturation. Non-saturation doctrine is used when there is time to carefully evaluate each potential threat. Saturation doctrine is needed if the ship gets into a combat situation where the number of inbound threats could overwhelm the ability of operators to respond. "If World War III starts and people start throwing a lot of stuff at me," Galluch said, "I will have grouped my doctrine together so that it's a one-push button that activates all of them. And what we've done is we've tested and we've looked at how they overlap each other and what the effects are going to be and make sure that we're getting the defense of the ship that we expect." This is where something like Auto-Special comes into play, in a "kill or be killed" scenario, as Galluch described it.

It's not enough to build the doctrine, though. Extensive testing goes into ensuring that it works properly. Once the ship arrives in theater, the first thing the crew does is test the weapons doctrine to see if there is anything in the environment that might cause it to fire in peacetime, which would not be good. This is done safely by enabling a hardware-level cutout called the Fire Inhibit Switch, or FIS. The FIS includes a key that must be inserted for any of the ship's weapons to fire. When the FIS key is inserted, a red light comes on; when it is turned to the right, the light turns green, meaning the weapons are live and ready to fire. When the FIS is red—or removed entirely—the ship's weapons are disabled at the hardware level. As Galluch put it, "there is no voltage that can be applied to light the wick and let the rocket fly out." By keeping the FIS red or removing the key, the ship's crew can test Aegis doctrine statements safely without any risk of inadvertent firing.

Establishing the doctrine and activating it is the sole responsibility of the ship's captain. Doctrine is more than just a set of programs. It is the embodiment

of the captain's intent for the warship. "Absolutely, it's automated, but there's so much human interface with what gets automated and how we apply that automation," Galluch said. Aegis doctrine is a way for the captain to predelegate his or her decision-making against certain threats.

The Aegis community uses automation in a very different way than the Patriot community did in 2003. Patriot operators sitting at the consoles in 2003 were essentially trusting in the automation. They had a handful of operational modes they could activate, but the operators themselves didn't write the rules for how the automation would function in those modes. Those rules were written years beforehand. Aegis, by contrast, can be customized and tailored to the specific operating environment. A destroyer operating in the Western Pacific, for example, might have different doctrine statements than one operating in the Persian Gulf to account for different threats from Chinese versus Iranian missiles. But the differences run deeper than merely having more options. The whole philosophy of automation is different. With Aegis, the automation is used to capture the ship captain's intent. In Patriot, the automation embodies the intent of the designers and testers. The actual operators of the system may not even fully understand the designers' intent that went into crafting the rules. The automation in Patriot is largely intended to *replace* warfighters' decision-making. In Aegis, the automation is used to *capture* warfighters' decision-making.

Another key difference is where decision authority rests. Only the captain of the ship has the authority to activate Aegis weapons doctrine. The captain can predelegate that authority to the tactical action officer on watch, but the order must be in writing as part of official orders. This means the decision-maker's experience level for Aegis operations is radically different from Patriot. When Captain Galluch took command of the USS *Ramage*, he had eighteen years of experience and had served on three prior Aegis ships. By contrast, the person who made the call on the first Patriot fratricide was a twenty-two-year-old second lieutenant fresh out of training.

Throughout our conversation, Galluch's experience was apparent. He was clearly comfortable using Aegis, but he wasn't flippant about its automation. What came through was a healthy respect for the weapon system. Activating Aegis doctrine is a serious decision, not be taken lightly. "You're never driving around with any kind of weapons doctrine activated" unless you expect to get into a fight, he explained. Even on manual mode, it is possible to launch a missile in seconds. And if need be, doctrine can be activated quickly. "I've made more Gulf deployments than I care to," he said. "I'm very comfortable with

driving around for months at a time with no active doctrine, but making damn sure that I have it set up and tested and ready to go if I need to." Because there can be situations that call for that level of automation. "You can get a missile fired pretty quickly, so why don't you do everything manually?" Galluch explained: "My view is that [manual control] works well if it's one or two missiles or threats. But if you're controlling fighters, you're doing a running gun battle with small patrol boats, you're launching your helicopter. . . . and you've got a bunch of cruise missiles coming in from different angles. You know, the watch is pretty small. It's ten or twelve people. So, there's not that many people . . . You can miss things coming in. That's where I get to the whole concept of saturation vs. normal. You want the man in the loop as much as possible, but there comes a time when you can get overwhelmed."

Aegis philosophy is one of human control over engagements, even when doctrine is activated. What varies is the form of human control. In Auto-Special doctrine, firing authority is delegated to Aegis's Command & Decision computer, but the human intent is still there. The goal is always to ensure "there is a conscious decision to fire a weapon," Galluch said. That doesn't mean that accidents can't happen. In fact, it is the constant preoccupation with the potential for accidents that helps prevent them. Galluch and others understand that, with doctrine activated, mishaps can happen. That's precisely why tight control is kept over the weapon. "[Ship commanding officers] are constantly balancing readiness condition to fire the weapon versus a chance for inadvertent firing," he explained.

I saw this tight control in action when Galluch took me to the Aegis simulation center and had his team run through a series of mock engagements. Galluch stood in as the ship's commanding officer and had Aegis-qualified sailors sitting at the same terminals doing the same jobs they would on a real ship. Then they went to work.

"ROLL GREEN"

The Navy would not permit me to record the precise language of the commands used between the sailors, but they allowed me to observe and report on what I saw. First, Galluch ordered the sailors to demonstrate a shot in manual operation. They put a simulated radar track on the screen and Galluch ordered them to target the track. They began working a firing solution, with the three sailors calmly but crisply reporting when they had completed each step in the process.

Once the firing solution was ready, Galluch ordered the tactical action officer to roll his FIS key to green. Then Galluch gave the order to fire. A sailor pressed the button to fire and called out that the missile was away. On a large screen in front of us, the radar showed the outbound missile racing toward the track.

I checked my watch. The whole process had been exceptionally fast—under a minute. The threat had been identified, a decision made, and a missile launched well under a minute, and that was in manual mode. I could understand Galluch's confidence in his ability to defend the ship without doctrine activated.

They did it again in Semi-Auto mode, now with doctrine activated. The FIS key was back at red, the tactical action officer having turned it back right after the missile was launched. Galluch ordered them to activate Semi-Auto doctrine. Then they brought up another track to target. This time, Aegis's Command & Decision computer generated part of the firing solution automatically. This shortened the time to fire by more than half.

They rolled FIS red, activated Auto SM doctrine, and put up a new track. Roll FIS green. Fire.

Finally, they brought up Auto-Special doctrine. This was it. This was the big leap into the great unknown, with the human removed from the loop. The sailors were merely observers now; they didn't need to take any action for the system to fire. Except . . . I looked at the FIS key. The key was in, but it was turned to red. Auto-Special doctrine was enabled, but there was still a hardware-level cutout in place. There was not even any voltage applied to the weapons. Nothing could fire until the tactical action officer rolled his key green.

The track for a simulated threat came up on the screen and Galluch ordered them to roll FIS green. I counted only a handful of heartbeats before a sailor announced the missiles were away. That's all it took for Command & Decision to target the track and fire.

But I felt cheated. They hadn't turned on the automation and leaned back in their chairs, taking it easy. Even on Auto-Special, and they had their hand literally on the key that disabled firing. And as soon as the missile was away, I saw the tactical action officer roll FIS red again. They weren't trusting the automation at all!

Of course, that was the point, I realized. They didn't trust it. The automation was powerful and they respected it—they even recognized there was a place for it—but that didn't mean they were surrendering their human decision-making to the machine.

To further drive the point home, Galluch had them demonstrate one final shot. With Auto-Special doctrine enabled, they rolled FIS green and let

Command & Decision take its shot. But then after the missile was away, Galluch ordered them to abort the missile. They pushed a button and a few seconds later the simulated missile disappeared from our radar, having been destroyed mid-flight. Even in the case of Auto-Special, even after the missile had been launched, they still had the ability to reassert human control over the engagement.

The Aegis community has reason to be so careful. In 1988, an Aegis warship was involved in a horrible accident. The incident haunts the community like a ghost—an ever-present reminder of the deadly power of an Aegis ship. Galluch described what transpired as a "terrible, painful lesson" and talked freely what the Aegis community learned to prevent future tragedies.

THE USS *VINCENNES* INCIDENT

The Persian Gulf in 1988 was a dangerous place. The Iran-Iraq war, under way since 1980, had boiled over into an extended "tanker war," with Iran and Iraq attacking each others' oil tankers, trying to starve their economies into submission. In 1987, Iran expanded to attacks against U.S.-flagged tanker ships carrying oil from Kuwait. In response, the U.S. Navy began escorting U.S.-flagged Kuwaiti tankers to protect them from Iranian attacks.

U.S. Navy ships in the Gulf were on high alert to threats from mines, rocket-equipped Iranian fast boats, warships, and fighter aircraft from several countries. A year earlier, the USS *Stark* had been hit with two Exocet missiles fired from an Iraqi jet and thirty-seven U.S. sailors were killed. In April 1988, in response to a U.S. frigate hitting an Iranian mine, the United States attacked Iranian oil platforms and sunk several Iranian ships. The battle only lasted a day, but tensions between the United States and Iran were high afterward.

On July 3, 1988, the U.S. warships USS *Vincennes* and USS *Montgomery* were escorting tankers through the Strait of Hormuz when they came into contact with Iranian fast boats. The *Vincennes*'s helicopter, which was monitoring the Iranian boats, came under fire. The *Vincennes* and *Montgomery* responded, pursuing the Iranian boats into Iranian territorial waters and opening fire.

While the *Vincennes* was in the midst of a gun battle with the Iranian boats, two aircraft took off in close sequence from Iran's nearby Bandar Abbas airport. Bandar Abbas was a dual-use airport, servicing both Iranian commercial and military flights. One aircraft was a commercial airliner, Iran Air Flight 655. The

other was an Iranian F-14 fighter. For whatever reason, in the minds of the sailors in the *Vincennes*'s combat information center, the tracks of the two aircraft on their radar screens became confused. The Iranian F-14 veered away but Iran Air 655 flew along its normal commercial route, which happened to be directly toward the *Vincennes*. Even though the commercial jet was squawking IFF and flying a commercial airliner route, the *Vincennes* captain and crew became convinced, incorrectly, that the radar track headed toward their position was an Iranian F-14 fighter.

As the aircraft approached, the *Vincennes* issued multiple warnings on military and civilian frequencies. There was no response. Believing the Iranians were choosing to escalate the engagement by sending a fighter and that his ship was under threat, the *Vincennes*'s captain gave the order to fire. Iran Air 655 was shot down, killing all 290 people on board.

The USS *Vincennes* incident and the Patriot fratricides sit as two opposite cases on the scales of automation versus human control. In the Patriot fratricides, humans trusted the automation too much. The *Vincennes* incident was caused by human error and more automation might have helped. Iran Air 655 was flying a commercial route squawking IFF. Well-crafted Aegis doctrine should not have fired.

Automation could have helped the *Vincennes* crew in this fast-paced combat environment. They weren't overwhelmed with too many missiles, but they were overwhelmed with too much information: the running gun battle with Iranian boats and tracking an F-14 and a commercial airliner launching in close succession from a nearby airport. In this information-saturated environment, the crew missed important details they should have noticed and made poor decisions with grave consequences. Automation, by contrast, wouldn't have gotten overwhelmed by the amount of information. Just as automation could help shoot down incoming missiles in a saturation scenario, it could also help *not fire* at the wrong targets in an information-overloaded environment.

ACHIEVING HIGH RELIABILITY

The Aegis community has learned from the *Vincennes* incident, Patriot fratricides, and years of experience to refine their operating procedures, doctrine, and software to the point where they are able to operate a very complex weapon system with low accidents. In the nearly thirty years since *Vincennes*, there has not been another similar incident, even with Aegis ships deployed continuously around the world.

The Navy's track record with Aegis shows that high-reliability operation of complex, hazardous systems is possible, but it doesn't come from testing alone. The human operators are not passive bystanders in the Aegis's operation, trusting blindly in the automation. They are active participants at every stage. They program the system's operational parameters, constantly monitor its modes of operation, supervise its actions in real time, and maintain tight control over weapons release authority. The Aegis culture is 180 degrees from the "unwarranted and uncritical trust in automation" that Army researchers found in the Patriot community in 2003.

After the Patriot fratricides, the Army launched the Patriot Vigilance Project, a three-year postmortem assessment to better understand what went wrong and to improve training, doctrine, and system design to ensure it didn't happen again. Dr. John Hawley is an engineering psychologist who led the project and spoke frankly about the challenges in implementing those changes. He said that there are examples of communities that have been able to manage high-risk technologies with very low accident rates, but high reliability is not easy to achieve. The Navy "spent a lot of money looking into . . . how you more effectively use a system like Aegis so that you don't make the kinds of mistakes that led to the [*Vincennes* incident]," he said. This training is costly and time-consuming, and in practice there are bureaucratic and cultural obstacles that may prevent military organizations from investing this amount of effort. Hawley explained that Patriot commanders are evaluated based on how many trained crews they keep ready. "If you make the [training] situation too demanding, then you could start putting yourself in the situation where you're not meeting those [crew] requirements." It may seem that militaries have an incentive to make training as realistic as possible, and to a certain extent that's true, but there are limits to how much time and money can be applied. Hawley argued that Army Patriot operators train in a "sham environment" that doesn't accurately simulate the rigors of real-world combat. As a result, he said "the Army deceives itself about how good their people really are. . . . It would be easy to believe you're good at this, but that's only because you've been able to handle the relatively non-demanding scenarios that they throw at you." Unfortunately, militaries might not realize their training is ineffective until a war occurs, at which point it may be too late.

Hawley explained that the Aegis community was partially protected from this problem because they use their system day in and day out on ships operating around the globe. Aegis operators get "consistent objective feedback from your environment on how well you're doing," preventing this kind of self-deception.

The Army's peacetime operating environment for the Patriot, on the other hand, is not as intense, Hawley said. "Even when the Army guys are deployed, I don't think that the quality of their experience with the system is quite the same. They're theoretically hot, but they're really not doing much of anything, other than just monitoring their scopes." Leadership is also a vital factor. "Navy brass in the Aegis community are absolutely paranoid" about another *Vincennes* incident, Hawley said.

The bottom line is that high reliability not easy to achieve. It requires frequent experience under real-world operating conditions and a major investment in time and money. Safety must be an overriding priority for leaders, who often have other demands they must meet. U.S. Navy submariners, aircraft carrier deck operators, and Aegis weapon system operators are very specific military communities that meet these conditions. Military organizations in general do not. Hawley was pessimistic about the ability of the U.S. Army to safely operate a system like the Patriot, saying it was "too sloppy an organization to . . . insist upon the kinds of rigor that these systems require."

This is a disappointing conclusion, because the U.S. Army is one of the most professional military organizations in history. Hawley was even more pessimistic about other nations. "Judging from history and the Russian army's willingness to tolerate casualties and attitude about fratricide . . . I would expect that . . . they would tilt the scale very much in the direction of lethality and operational effectiveness and away from necessarily safe use." Practice would appear to bear this out. The accident rate for Soviet/Russian submarines is far higher than for U.S. submarines.

If there is any military community that should be incentivized to avoid accidents, it is those responsible for maintaining control of nuclear weapons. There are no weapons on earth more destructive than nuclear weapons. Nuclear weapons are therefore an excellent test case for the extent to which dangerous weapons can be managed safely.

NUCLEAR WEAPONS SAFETY AND NEAR-MISS ACCIDENTS

The destructive power of nuclear weapons defies easy comprehension. A single *Ohio*-class ballistic missile submarine can carry twenty-four Trident II (D5) ballistic missiles, each with eight 100-kiloton warheads per missile. Each 100-kiloton warhead is over six times more powerful than the bomb dropped on

Hiroshima. Thus, a single submarine has the power to unleash over a thousand times the destructive power of the attack on Hiroshima. Individually, nuclear weapons have the potential for mass destruction. Collectively, a nuclear exchange could destroy human civilization. But outside of testing they have not been used, intentionally or accidentally, since 1945.

On closer inspection, however, the safety track record of nuclear weapons is less than inspiring. In addition to the Stanislav Petrov incident in 1983, there have been multiple nuclear near-miss incidents that could have had catastrophic consequences. Some of these could have resulted in an individual weapon's use, while others could potentially have led to a nuclear exchange between superpowers.

In 1979, a training tape left in a computer at the U.S. military's North American Aerospace Defense Command (NORAD) led military officers to initially believe that a Soviet attack was under way, until it was refuted by early warning radars. Less than a year later in 1980, a faulty computer chip led to a similar false alarm at NORAD. This incident progressed far enough that U.S. commanders notified National Security Advisor Zbigniew Brzezinski that 2,200 Soviet missiles were inbound to the United States. Brzezinski was about to inform President Jimmy Carter before NORAD realized the alarm was false.

Even after the Cold War ended, the danger from nuclear weapons did not entirely subside. In 1995, Norway launched a rocket carrying a science payload to study the aurora borealis that had a trajectory and radar signature similar to a U.S. Trident II submarine-launched nuclear missile. While a single missile would not have made sense as a first strike, the launch was consistent with a high-altitude nuclear burst to deliver an electromagnetic pulse to blind Russian satellites, a prelude to a massive U.S. first strike. Russian commanders brought President Boris Yeltsin the nuclear briefcase, who discussed a response with senior Russian military commanders before the missile was identified as harmless.

In addition to these incidents are safety lapses that might not have risked nuclear war but are troubling nonetheless. In 2007, for example, a U.S. Air Force B-52 bomber flew from Minot Air Force Base to Barksdale Air Force Base with six nuclear weapons aboard without the pilots or crew being aware. After it landed, the weapons remained on board the aircraft, unsecured and with ground personnel unaware of the weapons, until they were discovered the following day. This incident was merely the most egregious in a series of recent security lapses in the U.S. nuclear community that caused Air Force leaders to warn of an "erosion" of adherence to appropriate safety standards.

Nor were these isolated cases. There were at least thirteen near-use nuclear incidents from 1962 to 2002. This track record does not inspire confidence. Indeed, it lends credence to the view that near-miss incidents are normal, if terrifying, conditions of nuclear weapons. The fact that none of these incidents led to an actual nuclear detonation, however, presents an interesting puzzle: Do these near-miss incidents support the pessimistic view of normal accident theory that accidents are inevitable? Or does the fact that they didn't result in an actual nuclear detonation support the more optimistic view that high-reliability organizations can safely operate high-risk systems?

Stanford political scientist Scott Sagan undertook an in-depth evaluation of nuclear weapons safety to answer this very question. In the conclusion of his exhaustive study, published in *The Limits of Safety: Organizations, Accidents, and Nuclear Weapons*, Sagan wrote:

> When I began this book, the public record on nuclear weapons safety led me to expect that the high reliability school of organization theorists would provide the strongest set of intellectual tools for explaining this apparent success story. . . . The evidence presented in this book has reluctantly led me to the opposite view: the experience of persistent safety problems in the U.S. nuclear arsenal should serve as a warning.

Sagan concluded, "the historical evidence provides much stronger support for the ideas developed by Charles Perrow in *Normal Accidents*" than for high-reliability theory. Beneath the surface of what appeared, at first blush, to be a strong safety record was, in fact, a "long series of close calls with U.S. nuclear weapon systems." This is not because the organizations in charge of safeguarding U.S. nuclear weapons were unnaturally incompetent or lax. Rather, the history of nuclear near misses simply reflects "the inherent limits of organizational safety," he said. Military organizations have other operational demands they must accommodate beyond safety. Political scientists have termed this the "always/never dilemma." Militaries of nuclear-armed powers must *always* be ready to launch nuclear weapons at a moment's notice and deliver a massive strike against their adversaries for deterrence to be credible. At the same time, they must *never* allow unauthorized or accidental detonation of a weapon. Sagan says this is effectively "impossible." There are limits to how safe some hazards can be made.

THE INEVITABILITY OF ACCIDENTS

Safety is challenging enough with nuclear weapons. Autonomous weapons would be potentially more difficult in a number of ways. Nuclear weapons are

available to only a handful of actors, but autonomous weapons could proliferate widely, including to countries less concerned about safety. Autonomous weapons have an analogous problem to the always/never dilemma: once put into operation, they are expected to find and destroy enemy targets and not strike friendlies or civilian objects. Unlike nuclear weapons, some isolated mistakes might be tolerated with autonomous weapons, but gross errors would not.

The fact that autonomous weapons are not obviously as dangerous as nuclear weapons might make risk mitigation more challenging in some respects. The perception that automation can increase safety and reliability—which is true in some circumstances—could lead militaries to be less cautious with autonomous weapons than even other conventional weapons. If militaries cannot reliably institute safety procedures to control and account for nuclear weapons, their ability to do so with autonomous weapons is far less certain.

The overall track record of nuclear safety, Aegis operations, and the Patriot fratricides suggests that sound procedures can reduce the likelihood of accidents, but can never drive them to zero. By embracing the principles of high-reliability organizations, the U.S. Navy submarine and Aegis communities have been able to manage complex, hazardous systems safely, at least during peacetime. Had the Patriot community adopted some of these principles prior to 2003, the fratricides might have been prevented. At the very least, the Tornado shootdown could have been prevented with a greater cultural vigilance to respond to near-miss incidents and correct known problems, such as the anti-radiation missile misclassification problem, which had come up in testing. High-reliability theory does not promise zero accidents, however. It merely suggests that very low accident rates are possible. Even in industries where safety is paramount, such as nuclear power, accidents still occur.

There are reasons to be skeptical of the ability to achieve high-reliability operations for autonomous weapons. High-reliability organizations depend on three key features that work for Aegis in peacetime, but are unlikely to be present for fully autonomous weapons in war.

First, high-reliability organizations can achieve low accident rates by constantly refining their operations and learning from near-miss incidents. This is only possible if they can accumulate extensive experience in their operating environment. For example, when Aegis first arrives to an area, the ship operates for some time with its radar on and doctrine enabled, but the weapons deactivated, so sailors can see how the doctrine responds to the unique peculiarities of that specific operating environment. Similarly, FAA air traffic control, nuclear power plants, and aircraft carriers are systems people operate

day in and day out, accumulating large amounts of operational experience. This daily experience in real-world conditions allows them to refine safe operations.

When extreme events occur outside the norm, safety can be compromised. Users are not able to anticipate all of the possible interactions that may occur under atypical conditions. The 9.0 magnitude earthquake in Japan that led to the Fukushima-Daiichi meltdown is one such example. If 9.0 magnitude earthquakes causing forty-foot-high tsunamis were a regular occurrence, nuclear power plant operators would have quickly learned to anticipate the common-mode failure that knocked out primary and backup power. They would have built higher floodwalls and elevated the backup diesel generators off the ground. It is difficult, however, to anticipate the specific failures that might occur during atypical events.

War is an atypical condition. Militaries prepare for war, but the usual day-to-day experience of militaries is peacetime. Militaries attempt to prepare for the rigors of war through training, but no amount of training can replicate the violence and chaos of actual combat. This makes it very difficult for militaries to accurately predict the behavior of autonomous systems in war. Even for Aegis, activating the doctrine with the weapons disabled allows the operators to understand only how the doctrine will interact with a peacetime operating environment. A wartime operating environment will inevitably be different and raise novel challenges. The USS *Vincennes* accident highlights this problem. The *Vincennes* crew faced a set of conditions that were different from peacetime—military and commercial aircraft operating in close proximity from the same air base coupled with an ongoing hostile engagement from Iranian boats firing at the *Vincennes.* Had they routinely faced these challenges, they might have been able to come up with protocols to avoid an accident, such as staying off the path of civilian airliners. However, their day-to-day operations did not prepare them—and could not have prepared them—for the complexities that combat would bring. Hawley remarked, “You can go through all of the kinds of training that you think you should do . . . what nails you is the unexpected and the surprises.”

Another important difference between peacetime high-reliability organizations and war is the presence of adversarial actors. Safe operation of complex systems is difficult because bureaucratic actors have other interests that can sometimes compete with safety—profit, prestige, *etc.* However, none of the actors are generally hostile to safety. The risk is that people take shortcuts, not actively sabotage safe operations. War is different. War is an inherently adversarial environment in which there are actors attempting to undermine, exploit, or subvert systems. Militaries prepare their troops for this environment

not by trying to train their troops for every possible enemy action, but rather by inculcating a culture of resiliency, decisiveness, and autonomous execution of orders. Warfighters must adapt on the fly and come up with novel solutions to respond to enemy actions. This is an area in which humans excel, but machines perform poorly. The brittleness of automation is a major weakness when it comes to responding to adversary innovation. Once an adversary finds a vulnerability in an autonomous system, he or she is free to exploit it until a human realizes the vulnerability and either fixes the system or adapts its use. The system itself cannot adapt. The predictability that a human user finds desirable in automation can be a vulnerability in an adversarial environment.

Finally, the key ingredient in high-reliability organizations that makes them reliable is people, who by definition are not present in the actual execution of operations by a fully autonomous weapon. People are what makes high-reliability organizations reliable. Automation can play a role for "planned actions," as William Kennedy explained, but humans are required to make the system flexible, so that operations are resilient in the face of atypical events. Humans put slack in a system's operations, reducing the tight coupling between components and allowing for judgment to play a role in operations. In fully autonomous systems, humans are present during the design and testing of a system and humans put the system into operation, but humans are not present during actual operations. They cannot intervene if something goes wrong. The *organization* that enables high reliability is not available—the machine is on its own, at least for some period of time. Safety under these conditions requires something more than high-reliability organizations. It requires high-reliability fully autonomous complex machines, and there is no precedent for such systems. This would require a vastly different kind of machine from Aegis, one that was exceptionally predictable to the user but not to the enemy, and with a fault-tolerant design that defaulted to safe operations in the event of failures.

Given the state of technology today, no one knows how to build a complex system that is 100 percent fail-safe. It is tempting to think that future systems will change this dynamic. The promise of "smarter" machines is seductive: they will be more advanced, more intelligent, and therefore able to account for more variables and avoid failures. To a certain extent, this is true. A more sophisticated early warning system that understood U.S. nuclear doctrine might have been able to apply something similar to Petrov's judgment, determining that the attack was likely false. A more advanced version of the Patriot might have been able to take into account the IFF problems or electromagnetic interference and withhold firing on potentially ambiguous targets.

But smarter machines couldn’t avoid accidents entirely. New features increase complexity, a double-edged sword. More complex machines may be more capable, but harder for users to understand and predict their behavior, particularly in novel situations. For rule-based systems, deciphering the intricate web of relationships between the various rules that govern a system’s behavior and all possible interactions it might have with its environment quickly becomes impossible. Adding more rules can make a system smarter by allowing it to account for more scenarios, but the increased complexity of its internal logic makes it even more opaque to the user.

Learning systems would appear to sidestep this problem. They don’t rely on rules. Rather, the system is fed data and then learns the correct answer through experience over time. Some of the most innovative advances in AI are in learning systems, such as deep neural networks. Militaries will want to use learning systems to solve difficult problems, and indeed programs such as DARPA’s TRACE already aim to do so. Testing these systems is even more challenging, however. Incomprehensibility is a problem in complex systems, but it is far worse in systems that learn on their own.

11

BLACK BOX

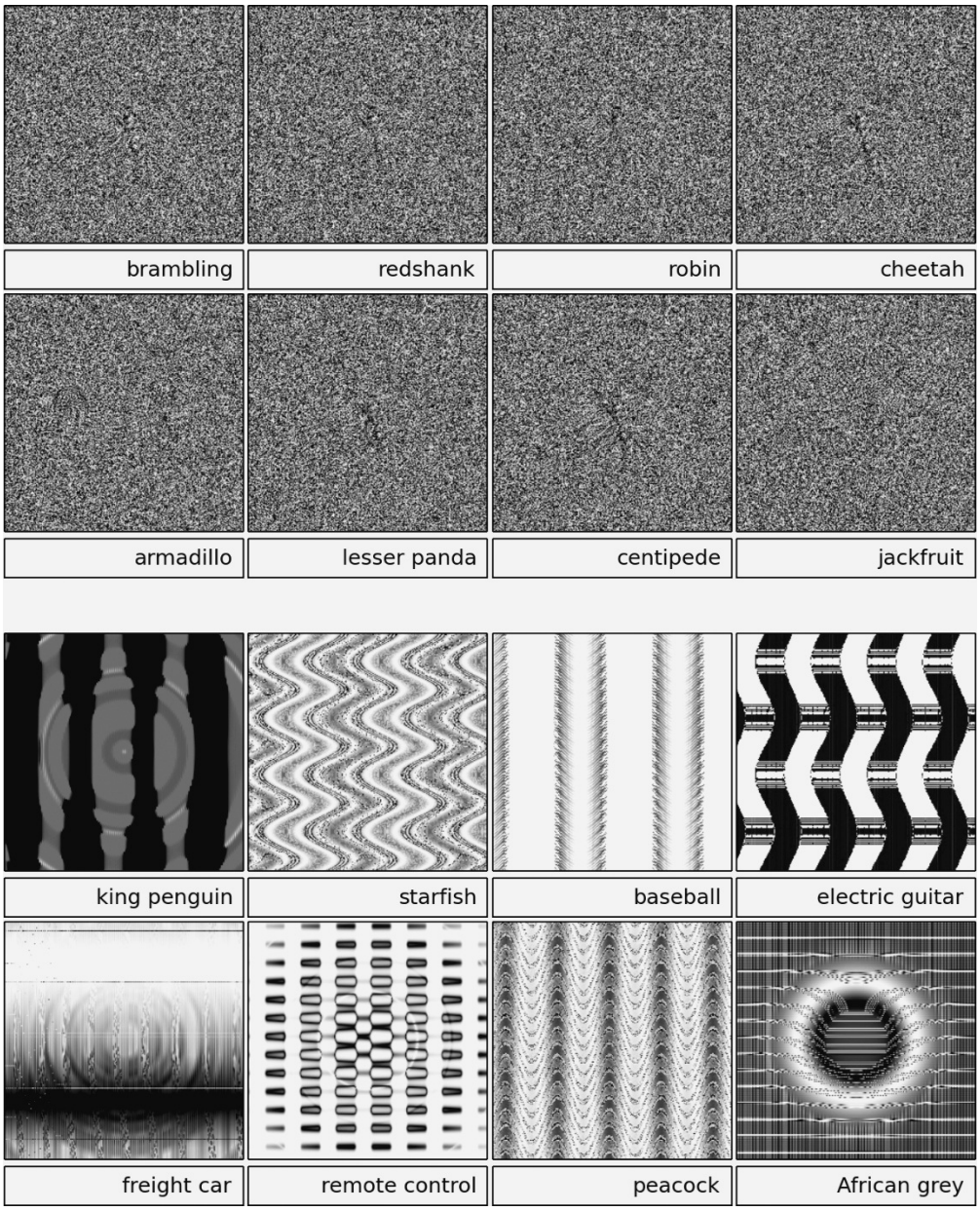
THE WEIRD, ALIEN WORLD OF DEEP NEURAL NETWORKS

Learning machines that don't follow a set of programmed rules, but rather learn from data, are effectively a "black box" to designers. Computer programmers can look at the network's output and see whether it is right or wrong, but understanding *why* the system came to a certain conclusion—and, more importantly, predicting its failures in advance—can be quite challenging. Bob Work specifically called out this problem when I met with him. "How do you do test and evaluation of learning systems?" he asked. He didn't have an answer; it is a difficult problem.

The problem of verifying the behavior of learning systems is starkly illustrated by the vulnerability of the current class of visual object recognition AIs to "adversarial images." Deep neural networks have proven to be an extremely powerful tool for object recognition, performing as well or better than humans in standard benchmark tests. However, researchers have also discovered that, at least with current techniques, they have strange and bizarre vulnerabilities that humans lack.

Adversarial images are pictures that exploit deep neural networks' vulnerabilities to trick them into confidently identifying false images. Adversarial images (usually created by researchers intentionally) come in two forms: one looks like abstract wavy lines and shapes and the other looks to the human eye like meaningless static. Neural networks nevertheless identify these nonsense images as concrete objects, such as a starfish, cheetah, or peacock, with greater than 99 percent confidence. The problem isn't that the networks get

some objects wrong. The problem is that the way in which the deep neural nets get the objects wrong is bizarre and counterintuitive to humans. The networks falsely identify objects from meaningless static or abstract shapes in ways that humans never would. This makes it difficult for humans to accurately predict the circumstances in which the neural net might fail. Because the network behaves in a way that seems totally alien, it is very difficult for humans to come up with an accurate mental model of the network's internal logic to predict its behavior. Within the black box of the neural net lies a counterintuitive and unexpected form of brittleness, one that is surprising even to the network's designers. This is not a weakness of only one specific network. This vulnerability appears to be replicated across most deep neural networks currently used for object recognition. In fact, one doesn't even need to know the specific internal structure of the network in order to fool it.

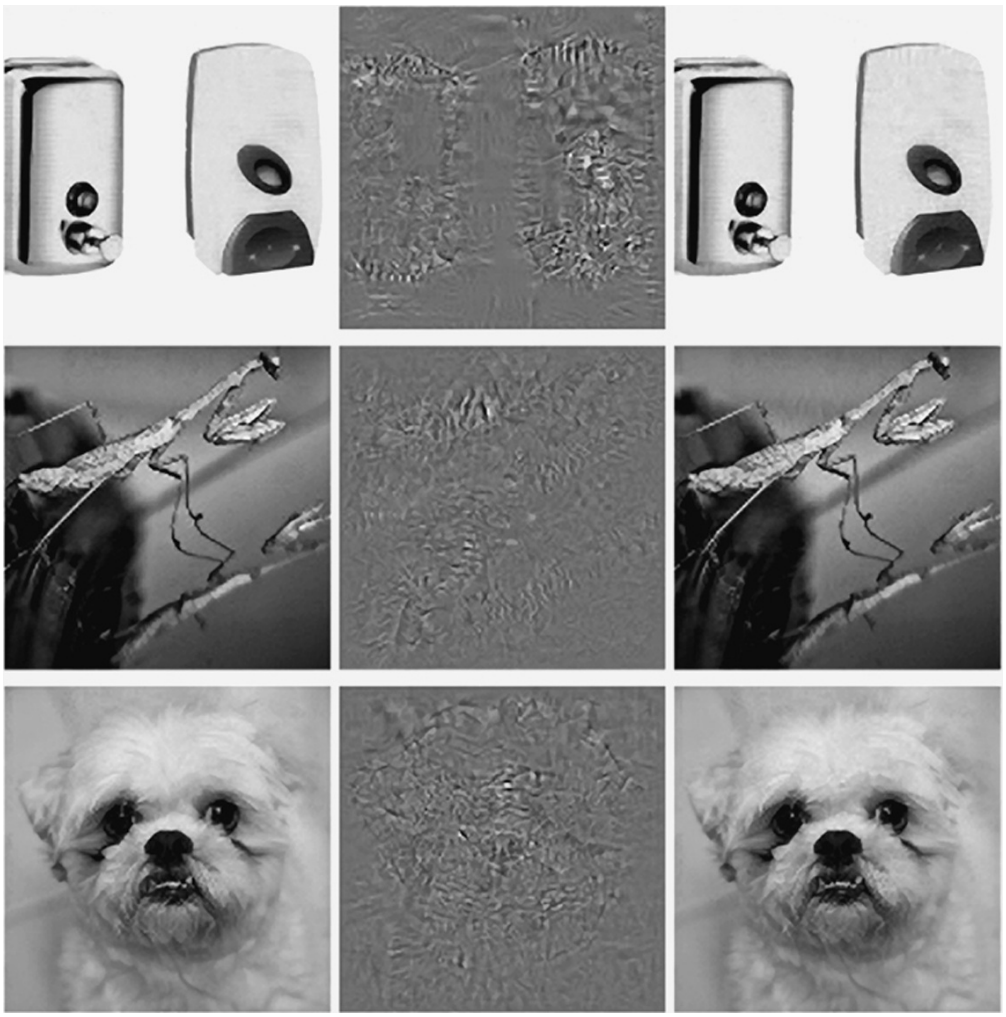


High-Confidence "Fooling Images" A *state-of*-the-art image recognition neural network identified these images, which are unrecognizable to humans, as familiar objects with a greater than 99.6 percent certainty. Researchers evolved the images using two different techniques: evolving individual pixels for the top eight images and evolving the image as a whole for the bottom eight images.

To better understand this phenomenon, I spoke with Jeff Clune, an AI researcher at the University of Wyoming who was part of the research team that discovered these vulnerabilities. Clune described their discovery as a "textbook case of scientific serendipity." They were attempting to design a "creative artificial intelligence that could endlessly innovate." To do this, they took an existing deep neural network that was trained on image recognition and had it

evolve new images that were abstractions of the image classes it knew. For example, if it had been trained to recognize baseballs, then they had the neural net evolve a new image that captured the essence of “baseball.” They envisioned this creative AI as a form of artist and expected the result would be unique computer images that were nevertheless recognizable to humans. Instead, the images they got were “completely unrecognizable garbage,” Clune said. What was even more surprising, however, was that other deep neural nets agreed with theirs and identified the seemingly garbage images as actual objects. Clune described this discovery as stumbling across a “huge, weird, alien world of imagery” that AIs all agree on.

This vulnerability of deep neural nets to adversarial images is a major problem. In the near term, it casts doubt on the wisdom of using the current class of visual object recognition AIs for military applications—or for that matter any high-risk applications in adversarial environments. Deliberately feeding a machine false data to manipulate its behavior is known as a spoofing attack, and the current state-of-the-art image classifiers have a known weakness to spoofing attacks that can be exploited by adversaries. Even worse, the adversarial images can be surreptitiously embedded into normal images in a way that is undetectable by humans. This makes it a “hidden exploit,” and Clune explained that this could allow an adversary to trick the AI in a way that was invisible to the human. For example, someone could embed an image into the mottled gray of an athletic shirt, tricking an AI security camera into believing the person wearing the shirt was authorized entry, and human security guards wouldn’t even be able to tell a fooling image being used.



Hidden Spoofing Attacks Inside Images *The images on the right and left columns look identical to humans, but are perceived very differently by neural networks. The left column shows the unaltered image, which is correctly identified by the neural network. The middle column shows, at 10x amplification, the difference between the images on the right and left. The right column shows the manipulated images, which contain a hidden spoofing attack that is not noticeable by humans. Due to the subtle manipulation of the image, the neural network identified all of the objects in the right column as "ostrich."*

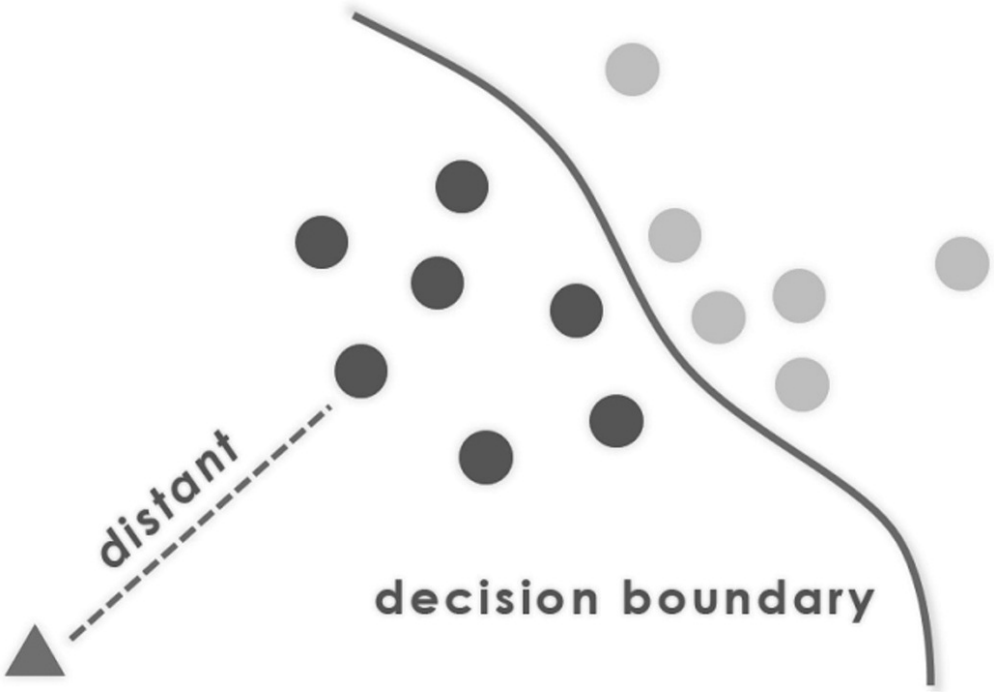
Researchers are only beginning to understand why the current class of deep neural networks is susceptible to this type of manipulation. It appears to stem from fundamental properties of their internal structures. The semitechnical explanation is that while deep neural networks are highly nonlinear at the macro level, they actually use linear methods to interpret data at the micro level. What does that mean? Imagine a field of gray dots separated into two clusters, with mostly light gray dots on the right and darker gray dots on the left, but with some overlap in the middle. Now imagine the neural net is trained on this data and asked to predict whether, given the position of a new dot, it is likely to be light or dark gray. Based on current methods, the AI will draw a line between the

light and dark gray clusters. The AI would then predict that new dots on the left side of the line are likely to be darker and new dots on the right side of the line are likely to be lighter, acknowledging that there is some overlap and there will be an occasional light gray dot on the left or dark gray on the right. Now imagine that you asked it to predict where the darkest possible dot would be. Since the further one moves to the left the more likely the dot is to be dark gray, the AI would put it "infinitely far to the left," Clune explained. This is the case even though the AI has zero information about any dots that far away. Even worse, because the dot is so far to the left, the AI would be very confident in its prediction that the dot would be dark. This is because at the micro level, the AI has a very simple, linear representation of the data. All it knows is that the further one moves left, the more likely the dot is to be dark.

The "fooling images," as Clune calls them, exploit this vulnerability. He explained that, "real-world images are a very, very small, rare subset of all possible images." On real-world images, the AIs do fairly well. This hack exploits their weakness on the extremes, however, in the space of all possible images, which is virtually infinite.

Because this vulnerability stems from the basic structure of the neural net, it is present in essentially every deep neural network commonly in use today, regardless of its specific design. It applies to visual object recognition neural nets but also to those used for speech recognition or other data analysis. This exploit has been demonstrated with song-interpreting AIs, for example. Researchers fed specially evolved noise into the AI, which sounds like nonsense to humans, but which the AI confidently interpreted as music.

In some settings, the consequences of this vulnerability could be severe. Clune gave a hypothetical example of a stock-trading neural net that read the news. News-reading trading bots appear to already be active on the market, evidenced by sharp market moves in response to news events at speeds faster than what is possible by human traders. If these bots used deep neural networks to understand text—a technique that has been demonstrated and is extremely effective—then they would be vulnerable to this form of hacking. Something as simple as a carefully crafted tweet could fool the bots into believing a terrorist attack was under way, for example. A similar incident already occurred in 2013 when the Associated Press Twitter account was hacked and used to send a false tweet reporting explosions at the White House. Stocks rapidly plunged in response. Eventually, the AP confirmed that its account had been hacked and markets recovered, but what makes Clune's exploit so damaging is that it could be done in a hidden way, without humans even aware that it is occurring.



Evolving Fooling Images *"Fooling images" are created by evolving novel images that are far from the decision boundary of the neural network. The "decision boundary" is the line of 50/50 confidence between two classes of images, in this case two shades of dots. The neural network's confidence in the image's correct classification increases as the image is further from the decision boundary. At the extremes, however, the image may no longer be recognizable, yet the neural network classifies the image with high confidence.*

You may be wondering, why not just feed these images back into the network and have it learn that these images are false, vaccinating the network against this hack? Clune and others have tried that. It doesn't work, Clune explained, because the space of all possible images is "virtually infinite." The neural net learns that specific image is false, but many more fooling images can be evolved. Clune compared it to playing an "infinite game of whack-a-mole" with "an infinite number of holes." No matter how many fooling images the AI learns to ignore, more can be created.

In principle, it ought to be possible to design deep neural networks that aren't vulnerable to this kind of spoofing attack, but Clune said that he hasn't seen a satisfactory solution yet. Even if one could be discovered, however, Clune said "we should definitely assume" that the new AI has some other "counterintuitive, weird" vulnerability that we simply haven't discovered yet.

In 2017, a group of scientific experts called JASON tasked with studying the implications of AI for the Defense Department came to a similar conclusion. After an exhaustive analysis of the current state of the art in AI, they concluded:

> [T]he sheer magnitude, millions or billions of parameters (i.e. weights/biases/etc.), which are learned as part of the training of the net . . . makes it impossible to really understand exactly how the network does what it does. Thus the response of the network to all possible inputs is unknowable.

Part of this is due to the early stage of research in neural nets, but part of it is due to the sheer complexity of the deep learning. The JASON group argued that "the very nature of [deep neural networks] may make it intrinsically difficult for them to transition into what is typically recognized as a professionally engineered product."

AI researchers are working on ways to build more transparent AI, but Jeff Clune isn't hopeful. "As deep learning gets even more powerful and more impressive and more complicated and as the networks grow in size, there will be more and more and more things we don't understand. . . . We have now created artifacts so complicated that we ourselves don't understand them." Clune likened his position to an "AI neuroscientist" working to discover how these artificial brains function. It's possible that AI neuroscience will elucidate these complex machines, but Clune said that current trends point against it: "It's almost certain that as AI becomes more complicated, we'll understand it less and less."

Even if it were possible to make simpler, more understandable AI, Clune argued that it probably wouldn't work as well as AI that is "super complicated and big and weird." At the end of the day, "people tend to use what works," even if they don't understand it. "This kind of a race to use the most powerful stuff—if the most powerful stuff is inscrutable and unpredictable and incomprehensible—somebody's probably going to use it anyway."

Clune said that this discovery has changed how he views AI and is a "sobering message." When it comes to lethal applications, Clune warned using deep neural networks for autonomous targeting "could lead to tremendous harm." An adversary could manipulate the system's behavior, leading it to attack the wrong targets. "If you're trying to classify, target, and kill autonomously with no human in the loop, then this sort of adversarial hacking could get fatal and tragic extremely quickly."

While couched in more analytic language, the JASON group essentially issued the same cautionary warning to DoD:

> [I]t is not clear that the existing AI paradigm is immediately amenable to any sort of software engineering validation and verification. This is a serious issue, and is a potential roadblock to DoD's use of these modern AI systems, especially when considering the liability and accountability of using AI in lethal systems.

Given these glaring vulnerabilities and the lack of any known solution, it would

be extremely irresponsible to use deep neural networks, as they exist today, for autonomous targeting. Even without any knowledge about how the neural network was structured, adversaries could generate fooling images to draw the autonomous weapon onto false targets and conceal legitimate ones. Because these images can be hidden, it could do so in a way that is undetectable by humans, until things start blowing up.

Beyond immediate applications, this discovery should make us far more cautious about machine learning in general. Machine learning techniques are powerful tools, but they also have weaknesses. Unfortunately, these weaknesses may not be obvious or intuitive to humans. These vulnerabilities are different and more insidious than those lurking within complex systems like nuclear reactors. The accident at Three Mile Island might not have been predictable ahead of time, but it is at least understandable after the fact. One can lay out the specific sequence of events and understand how one event led to another, and how the combination of highly improbable events led to catastrophe. The vulnerabilities of deep neural networks are different; they are entirely alien to the human mind. One group of researchers described them as “nonintuitive characteristics and intrinsic blind spots, whose structure is connected to the data distribution in a non-obvious way.” In other words: the AIs have weaknesses that we can’t anticipate and we don’t really understand how it happens or why.

12

FAILING DEADLY

THE RISK OF AUTONOMOUS WEAPONS

Acknowledging that machine intelligence has weaknesses does not negate its advantages. AI isn't good or bad. It is powerful. The question is how humans should use this technology. How much freedom (autonomy) should we give AI-empowered machines to perform tasks on their own?

Delegating a task to a machine means accepting the consequences if the machine fails. John Borrie of UNIDIR told me, "I think that we're being overly optimistic if we think that we're not going to see problems of system accidents" in autonomous weapons. Army researcher John Hawley agreed: "If you're going to turn these things loose, whether it be Patriot, whether it be Aegis, whether it be some type of totally unmanned system with the ability to kill, you have to be psychologically prepared to accept the fact that sometimes incidents will happen." Charles Perrow, the father of normal accident theory, made a similar conclusion about complex systems in general:

> [E]ven with our improved knowledge, accidents and, thus, potential catastrophes are inevitable in complex, tightly coupled systems with lethal possibilities. We should try harder to reduce failures—and that will help a great deal—but for some systems it will not be enough. . . . We must live and die with their risks, shut them down, or radically redesign them.

If we are to use autonomous weapons, we must accept their risks. All weapons are dangerous. War entails violence. Weapons that are designed to be dangerous to the enemy can also be dangerous to the user if they slip out of control. Even a knife wielded improperly can slip and cut its user. Most modern weapons, regardless of their level of autonomy, are complex systems. Accidents

will happen, and sometimes these accidents will result in fratricide or civilian casualties. What makes autonomous weapons any different?

The key difference between semi-, supervised, and fully autonomous weapons is amount of damage the system can cause until the next opportunity for a human to intervene. In semi-or supervised autonomous weapons, such as Aegis, the human is a natural fail-safe against accidents, a circuit breaker if things go wrong. The human can step outside of the rigid rules of the system and exercise judgment. Taking the human out of the loop reduces slack and increases the coupling of the system. In fully autonomous weapons, there is no human to intervene and halt the system's operation. A failure that might cause a single unfortunate incident with a semiautonomous weapon could cause far greater damage if it occurred in a fully autonomous weapon.

THE RUNAWAY GUN

A simple malfunction in an automatic weapon—a machine gun—provides an analogy for the danger with autonomous weapons. When functioning properly, a machine gun continues firing so long as the trigger remains held down. Once the trigger is released, a small metal device called a "sear" springs into place to stop the operating rod within the weapon from moving, halting the automatic firing process. Over time, however, the sear can become worn down. If the sear becomes so worn down that it fails to stop the operating rod, the machine gun will continue firing even when the trigger is released. The gun will keep firing on its own until it exhausts its ammunition.

This malfunction is called a runaway gun.

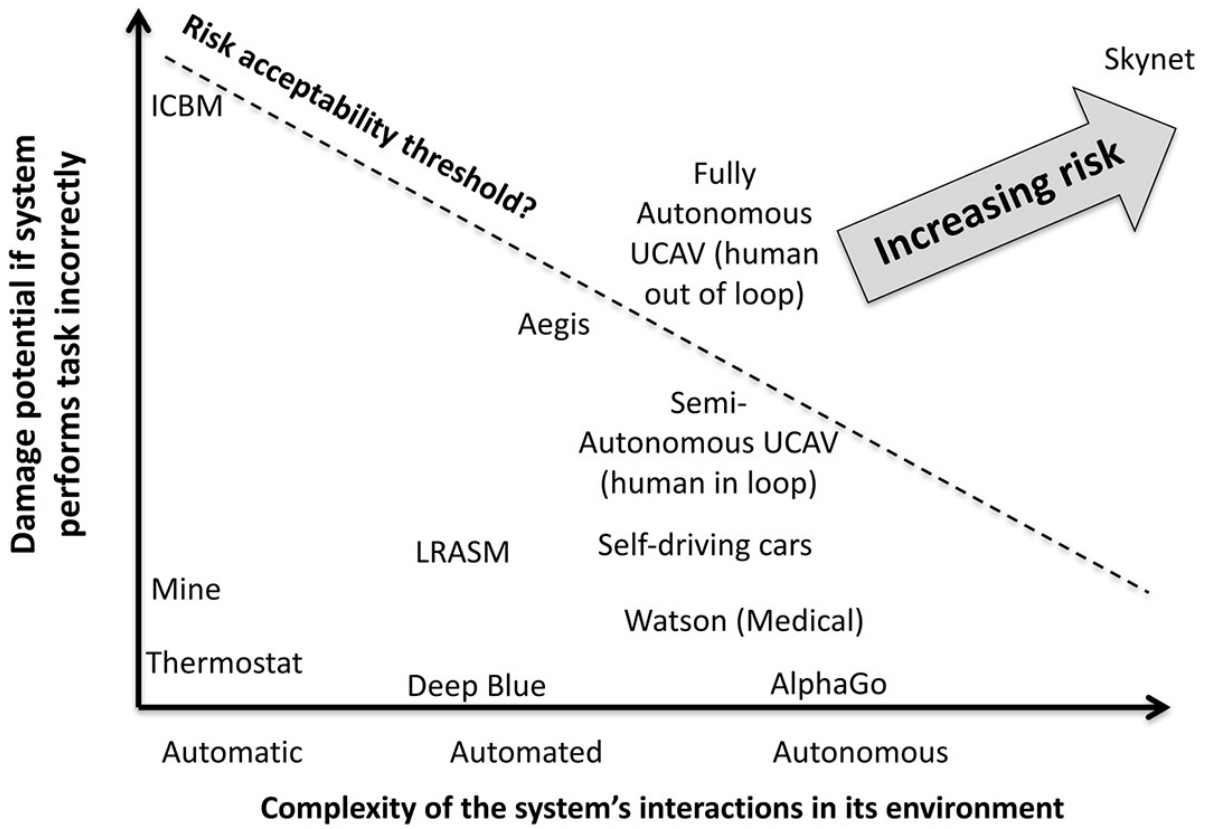
Runaway guns are serious business. The machine gunner has let go of the trigger, but the gun continues firing: the firing process is now fully automatic, with no way to directly halt it. The only way to stop a runaway gun is to break the links on the ammunition belt feeding into the weapon. While this is happening, the gunner must ensure the weapon stays pointed in a safe direction.

A runaway gun is the kind of hypothetical danger I was aware of as an infantry soldier, but I remember clearly the first time I heard about one actually occurring. We were out on an overnight patrol in northeastern Afghanistan and got word of an incident back at the outpost where we were based. An M249 SAW (light machine gun) gunner tried to disassemble his weapon without removing the ammunition first. (Pro tip: bad idea.) When he removed the pistol grip, the sear that held back the operating rod came out with it. The bolt slammed forward, firing off a round. The recoil cycled the weapon, which

reloaded and fired again. Without anything to stop it, the weapon kept firing. A stream of bullets sailed across the outpost, stitching a line of holes across the far wall until someone broke the links of the ammunition belt feeding into the gun. No one was killed, but such accidents don't always end well.

In 2007, a South African antiaircraft gun malfunctioned on a firing range, resulting in a runaway gun that killed nine soldiers. Contrary to breathless reports of a "robo-cannon rampage," the remote gun was not an autonomous weapon and likely malfunctioned because of a mechanical problem, not a software glitch. According to sources knowledgeable about the weapon, it was likely bad luck, not deliberate targeting, that caused the gun to swivel toward friendly lines when it malfunctioned. Unfortunately, despite the heroic efforts of one artillery officer who risked her life to try to stop the runaway gun, the gun poured a string of 35 mm rounds into a neighboring gun position, killing the soldiers present.

Runaway guns can be deadly affairs even with simple machine guns that can't aim themselves. A loss of control of an autonomous weapon would be a far more dangerous situation. The destruction unleashed by an autonomous weapon would not be random—it would be targeted. If there were no human to intervene, a single accident could become many, with the system continuing to engage inappropriate targets until it exhausted its ammunition. "The machine doesn't know it's making a mistake," Hawley observed. The consequences to civilians or friendly forces could be disastrous.



Risk of Delegating Autonomy to a Machine

THE DANGER OF AUTONOMOUS WEAPONS

With autonomous weapons, we are like Mickey enchanting the broomstick. We trust that autonomous weapons will perform their functions correctly. We trust that we have designed the system, tested it, and trained the operators correctly. We trust that the operators are using the system the right way, in an environment they can understand and predict, and that they remain vigilant and don't cede their judgment to the machine. Normal accident theory would suggest that we should trust a little less.

Autonomy is tightly bounded in weapons today. Fire-and-forget missiles cannot be recalled once launched, but their freedom to search for targets in space and time is limited. This restricts the damage they could cause if they fail. In order for them to strike the wrong target, there would need to be an inappropriate target that met the seeker's parameters within the seeker's field of view for the limited time it was active. Such a circumstance is not inconceivable. That appears to be what occurred in the F-18 Patriot fratricide. If missiles were made more autonomous, however—if the freedom of the seeker to search in time and

space were expanded—the possibility for more accidents like the F-18 shootdown would expand.

Supervised autonomous weapons such as the Aegis have more freedom to search for targets in time and space, but this freedom is compensated for by the fact that human operators have more immediate control over the weapon. Humans supervise the weapon's operation in real time. For Aegis, they can engage hardware-level cutouts that will disable power, preventing a missile launch. An Aegis is a dangerous dog kept on a tight leash.

Fully autonomous weapons would be a fundamental paradigm shift in warfare. In deploying fully autonomous weapons, militaries would be introducing onto the battlefield a highly lethal system that they cannot control or recall once launched. They would be sending this weapon into an environment that they do not control where it is subject to enemy hacking and manipulation. In the event of failures, the damage fully autonomous weapons could cause would be limited only by the weapons' range, endurance, ability to sense targets, and magazine capacity.

Additionally, militaries rarely deploy weapons individually. Flaws in any one system are likely to be replicated in entire squadrons and fleets of autonomous weapons, opening the door to what John Borrie described as "incidents of mass lethality." This is fundamentally different from human mistakes, which tend to be idiosyncratic. Hawley told me, "If you put someone else in [a fratricide situation], they probably would assess the situation differently and they may or may not do that." Machines are different. Not only will they continue making the same mistake; all other systems of that same type will do so as well.

A frequent refrain in debates about autonomous weapons is that humans also make mistakes and if the machines are better, then we should use the machines. This objection is a red herring and misconstrues the nature of autonomous weapons. If there are specific engagement-related tasks that automation can do better than humans, then those tasks should be automated. Humans, whether in the loop or on the loop, act as a vital fail-safe, however. It's the difference between a pilot flying an airplane on autopilot and an airplane with no human in the cockpit at all. The key factor to assess with autonomous weapons isn't whether the system is better than a human, but rather if the system fails (which it inevitably will), what is the amount of damage it could cause, and can we live with that risk?

Putting an offensive fully autonomous weapon system into operation would be like turning an Aegis to Auto-Special, rolling FIS green, pointing it toward a

communications-denied environment, and having everyone on board exit the ship. Deploying autonomous weapons would be like putting a whole fleet of these systems into operation. There is no precedent for delegating that amount of lethality to autonomous systems without any ability for humans to intervene. In fact, placing that amount of trust in machines would run 180 degrees counter to the tight control the Aegis community maintains over supervised autonomous weapons today.

I asked Captain Galluch what he thought of an Aegis operating on its own with no human supervision. It was the only question I asked him in our four-hour interview for which he did not have an immediate answer. It was clear that in his thirty-year career it had never once occurred to him to turn an Aegis to Auto-Special, roll FIS green, and have everyone on board exit the ship. He leaned back in his chair and looked out the window. "I don't have a lot of good answers for that," he said. But then he began to walk through what one might need to do to build trust in such a system, applying his decades of experience with Aegis. One would need to "build a little, test a little," he said. High-fidelity computer modeling coupled with real-world tests and live-fire exercises would be necessary to understand the system's limitations and the risks of using it. Still, he said, if the military did deploy a fully autonomous weapon, "we're going to get a *Vincennes*-like response" in the beginning. "Understanding the complexity of Aegis has been a thirty-year process," Galluch said. "Aegis today is not the Aegis of *Vincennes*," but only because the Navy has learned from mistakes. With a fully autonomous weapon, we'd be starting at year zero.

Deploying fully autonomous weapons would be a weighty risk, but it might be one that militaries decide is worth taking. Doing so would be entering uncharted waters. Experience with supervised autonomous weapons such as Aegis would be useful, but only to a point. Fully autonomous weapons in wartime would face unique conditions that limit the applicability of lessons from high-reliability organizations. The wartime operating environment is different from day-to-day peacetime experience. Hostile actors are actively trying to undermine safe operations. And no humans would be present at the time of operation to intervene or correct problems.

There is one industry that has many of these dynamics, where automation is used in a competitive high-risk environment and at speeds that make it impossible for humans to compete: stock-trading. The world of high-frequency trading—and its consequences—has instructive lessons for what could happen if militaries deployed fully autonomous weapons.

PART IV

Flash War

13

BOT VS. BOT

AN ARMS RACE IN SPEED

On May 6, 2010, at 2:32 p.m. Eastern Time, the S&P 500, NASDAQ, and Dow Jones Industrial Average all began a precipitous downward slide. Within a few minutes, they were in free fall. By 2:45 p.m., the Dow had lost nearly 10 percent of its value. Then, just as inexplicably, the markets rebounded. By 3:00 p.m., whatever glitch had caused the sharp drop in the market was over. However, the repercussions from the "Flash Crash," as it came to be known, were only beginning.

Asian markets tumbled when they opened the next day, and while the markets soon stabilized, it was harder to repair confidence. Traders described the Flash Crash as "horrifying" and "absolute chaos," reminiscent of the 1987 "Black Monday" crash where the Dow Jones lost 22 percent of its value. Market corrections had occurred before, but the sudden downward plunge followed by an equally rapid reset suggested something else. In the years preceding the Flash Crash, algorithms had taken over a large fraction of stock trading, including high-frequency trading that occurred at superhuman speeds. Were the machines to blame?

Investigations followed, along with counter-investigations and eventually criminal charges. Simple answers proved elusive. Researchers blamed everything from human error to brittle algorithms, high-frequency trading, market volatility, and deliberate market manipulation. In truth, all of them likely played a role. Like other normal accidents, the Flash Crash had multiple causes, any one of which individually would have been manageable. The combination,

however, was uncontrollable.

RISE OF THE MACHINES

Stock trading today is largely automated. Gone are the days of floor traders shouting prices and waving their hands to compete for attention in the furious scrum of the New York Stock Exchange. Approximately three-quarters of all trades made in the U.S. stock market today are executed by algorithms. Automated stock trading, sometimes called algorithmic trading, is when computer algorithms are used to monitor the market and make trades based on certain conditions. The simplest kind of algorithm, or "algo," is used to break up large trades into smaller ones in order to minimize the costs of the trade. If a single buy or sell order is too large relative to the volume of that stock that is regularly traded, placing the order all at once can skew the market price. To avoid this, traders use algorithms to break up the sale into pieces that can be executed incrementally according to stock price, time, volume, or other factors. In such cases, the decision to make the trade (to buy or sell a certain amount of stock) is still made by a person. The machine simply handles the execution of the trade.

Some trading algorithms take on more responsibility, actually making automated trading decisions to buy or sell based on the market. For example, an algorithm could be tasked to monitor a stock's price over a period of time. When the price moves significantly above or below the average of where the price has been, the algo sells or buys accordingly, under the assumption that over time the price will revert back to the average, yielding a profit. Another strategy could be to look for arbitrage opportunities, where the price of a stock in one market is different from the price in another market, and this price difference can be exploited for profit. All of these strategies could, in principle, be done by humans. Automated trading offers the advantage, however, of monitoring large amounts of data and immediately and precisely making trades in ways that would be impossible for humans.

Speed is a vital factor in stock trading. If there is a price imbalance and a stock is under-or overpriced, many other traders are also looking to sweep up that profit. Move too slow and one could miss the opportunity. The result has been an arms race in speed and the rise of high-frequency trading, a specialized type of automated trading that occurs at speeds too quick for humans to even register.

The blink of an eye takes a fraction of a second—0.1 to 0.4 seconds—but is

still an eon compared to high-frequency trading. High-frequency trades move at speeds measured in microseconds: 0.000001 seconds. During the span of a single eyeblink, 100,000 microseconds pass by. The result is an entirely new ecosystem, a world of trading bots dueling at superhuman speeds only accessible by machines.

The gains from even a slight advantage in speed are so significant that high-frequency traders will go to great lengths to shave just a few microseconds off their trading times. High-frequency traders colocate their servers within the server rooms of stock exchanges, cutting down on travel time. Some are even willing to pay additional money to move their firm's servers a few feet closer to the stock exchange's servers inside the room. Firms try to find the shortest route for their cables within the server room, cutting microseconds off transit time. Like race teams outfitting an Indy car, high-frequency traders spare no expense in optimizing every part of their hardware for speed, from data switches to the glass inside fiber-optic cables.

At the time scales at which high-frequency trading operates, humans have to delegate trading decisions to the algorithms. Humans can't possibly observe the market and react to it in microseconds. That means if things go wrong, they can go wrong very quickly. To ensure algorithms do what they are designed to do once released into the real world, developers test them against actual stock market data, but with trading disabled—analogous to testing Aegis doctrine with the FIS key turned red. Despite this, accidents still occur.

"KNIGHTMARE ON WALL STREET"

In 2012, Knight Capital Group was a titan of high-frequency trading. Knight was a "market maker," a high-frequency trader that executed over 3.3 billion trades, totaling $21 billion, every single day. Like most high-frequency traders, Knight didn't hold on to this stock. Stocks were bought and sold the same day, sometimes within fractions of a second. Nevertheless, Knight was a key player in the U.S. stock market, executing 17 percent of all trades on the New York Stock Exchange and NASDAQ. Their slogan was, "The Science of Trading, the Standard of Trust." Like many high-frequency trading firms, their business was lucrative. On the morning of July 31, 2012, Knight had $365 million in assets. Within 45 minutes, they would be bankrupt.

At 9:30 a.m. Eastern Time on July 31, U.S. markets opened and Knight deployed a new automated trading system. Instantly, it was apparent that

something was wrong. One of the functions of the automated trading system was to break up large orders into smaller ones, which then would be executed individually. Knight's trading system wasn't registering that these smaller trades were actually completed, however, so it kept tasking them again. This created an endless loop of trades. Knight's trading system began flooding the market with orders, executing over a thousand trades a second. Even worse, Knight's algorithm was buying high and selling low, losing money on every trade.

There was no way to stop it. The developers had neglected to install a "kill switch" to turn their algorithm off. There was no equivalent of "rolling FIS red" to terminate trading. While Knight's computer engineers worked to diagnose the problem, the software was actively trading in the market, moving $2.6 million a second. By the time they finally halted the system 45 minutes later, the runaway algo had executed 4 million trades, moving $7 billion. Some of those trades made money, but Knight lost a net $460 million. The company only had $365 million in assets. Knight was bankrupt.

An influx of cash from investors helped Knight cover their losses, but the company was ultimately sold. The incident became known as the "Knightmare on Wall Street," a cautionary tale for partners to tell their associates about the dangers of high-frequency trading. Knight's runaway algo vividly demonstrated the risk of using an autonomous system in a high-stakes application, especially with no ability for humans to intervene. Despite their experience in high-frequency trading, Knight was taking fatal risks with their automated stock trading system.

BEHIND THE FLASH CRASH

If the Knightmare on Wall Street was like a runaway gun, the Flash Crash was like a forest fire. The damage from Knight's trading debacle was largely contained to a single company, but the Flash Crash affected the entire market. A volatile combination of factors meant that during the Flash Crash, one malfunctioning algorithm interacted with an entire marketplace ready to run out of control. And run away it did.

The spark that lit the fire was a single bad algorithm. At 2:32 p.m. on May 6, 2010, Kansas-based mutual fund trader Waddell & Reed initiated a sale of 75,000 S&P 500 E-mini futures contracts estimated at $4.1 billion. (E-minis are a smaller type of futures contract, one-fifth the size of a regular futures contract. A futures contract is what it sounds like: an agreement to buy or sell at a certain

price at a certain point in time in the future.) Because executing such a large trade all at once could distort the market, Waddell & Reed used a "sell algorithm" to break up the sale into smaller trades, a standard practice. The algorithm was tied to the overall volume of E-minis sold on the market, with direction to execute the sale at 9 percent of the trading volume over the previous minute. In theory, this should have spread out the sale so as to not overly influence the market.

The sell algorithm was given no instructions with regard to time or price, however, an oversight that led to a catastrophic case of brittleness. The market that day was already under stress. Government investigators later characterized the market as "unusually turbulent," in part due to an unfolding European debt crisis that was causing uncertainty. By midafternoon, the market was experiencing "unusually high volatility" (sharp movements in prices) and low liquidity (low market depth). It was into these choppy waters that the sell algorithm waded.

Only twice in the previous year had a single trader attempted to unload so many E-minis on the market in a single day. Normally, a trade of this scale took hours to execute. This time, because the sell algorithm was only tied to volume and not price or time, it happened very quickly: within 20 minutes.

The sell algorithm provided the spark, and high-frequency traders were the gasoline. High-frequency traders bought the E-minis the sell algorithm was unloading and, as is their frequent practice, rapidly resold them. This increased the volume of E-minis being traded on the market. Since the rate at which the sell algorithm sold E-minis was tied to volume but not price or time, it accelerated its sales, dumping more E-minis on an already stressed market.

Without buyers interested in buying up all of the E-minis that the sell algorithm and high-frequency traders were selling, the price of E-minis dropped, falling 3 percent in just four minutes. This generated a "hot potato" effect among high-frequency traders as they tried to unload the falling E-minis onto other high-frequency traders. In one 14-second period, high-frequency trading algorithms exchanged 27,000 E-mini contracts. (The total amount Waddell & Reed were trying to sell was 75,000 contracts.) All the while as trading volume skyrocketed, the sell algorithm kept unloading more and more E-minis on a market that was unable to handle them.

The plummeting E-minis dragged down other U.S. markets. Observers watched the Dow Jones, NASDAQ, and S&P 500 all plunge, inexplicably. Finally, at 2:45:28 p.m., an automated "stop logic" safety on the Chicago Mercantile Exchange kicked in, halting E-mini trading for 5 seconds and

allowing the markets to reset. They rapidly recovered, but the sharp distortions in the market wreaked havoc on trading. Over 20,000 trades had been executed at what financial regulators termed “irrational prices” far from their norm, some as low as a penny or as high as $100,000. After the markets closed, the Financial Industry Regulatory Authority worked with stock exchanges to cancel tens of thousands of “clearly erroneous” trades.

The Flash Crash demonstrated how when brittle algorithms interact with a complex environment at superhuman speeds, the result can be a runaway process with catastrophic consequences. The stock market as a whole is an incredibly complex system that defies simple understanding, which can make predicting these interactions difficult ahead of time. On a different day, under different market conditions, the same sell algorithm may not have led to a crash.

PRICE WARS: $23,698,655.93 (PLUS $3.99 SHIPPING)

While complexity was a factor in the Flash Crash, even simple interactions between algorithms can lead to runaway escalation. This phenomenon was starkly illustrated when two warring bots jacked up the price of an otherwise ordinary book on Amazon to $23 million. Michael Eisen, a biologist at UC Berkeley, accidentally stumbled across this price war for Peter Lawrence’s *Making of a Fly: The Genetics of Animal Design*. Like a good scientist, Eisen began investigating.

Two online sellers, *bordeebook* and *profnath*, both of whom were legitimate online booksellers with thousands of positive ratings, were locked in a runaway price war. Once a day, *profnath* would set its price to 0.9983 times *bordeebook*’s price, slightly undercutting them. A few hours later, *bordeebook* would change its price to 1.270589 times *profnath*’s. The combination raised both booksellers’ prices by approximately 27 percent daily.

Bots were clearly to blame. The pricing was irrational and precise. *Profnath*’s algorithm made sense; it was trying to draw in sales by slightly undercutting the highest price on the market. What was *bordeebook*’s algorithm doing, though? Why *raise* the price over the highest competitor?

Eisen hypothesized that *bordeebook* didn’t actually own the book. Instead, they probably were posting an ad and hoping their higher reviews would attract customers. If someone bought the book, then of course *bordeebook* would have to buy it, so they set their price slightly above—1.270589 times greater than—the highest price on the market, so they could make a profit.

Eventually, someone at one of the two companies caught on. The price peaked out at $23,698,655.93 (plus $3.99 shipping) before dropping back to a tamer $134.97, where it stayed. Eisen mused in a blog posting, however, about the possibilities for "chaos and mischief" that this discovery suggested. A person could potentially hack this vulnerability of the bots, manipulating prices.

SPOOFING THE BOT

Eisen wasn't the first to think of exploiting the predictability of bots for financial gain. Others had seen these opportunities before him, and they'd gone and done it. Six years after the Flash Crash in 2016, London-based trader Navinder Singh Sarao pled guilty to fraud and spoofing, admitting that he used an automated trading algorithm to manipulate the market for E-minis on the day of the crash. According to the U.S. Department of Justice, Sarao used automated trading algorithms to place multiple large-volume orders to create the appearance of demand to drive up price, then cancelled the orders before they were executed. By deliberately manipulating the price, Sarao could buy low and sell high, making a profit as the price moved.

It would be overly simplistic to pin the blame for the Flash Crash on Sarao. He continued his alleged market manipulation for *five years* after the Flash Crash until finally arrested in 2015 and his spoofing algorithm was reportedly turned *off* during the sharpest downturn in the Flash Crash. His spoofing could have exacerbated instability in the E-mini market that day, however, contributing to the crash.

AFTERMATH

In the aftermath of the Flash Crash, regulators installed "circuit breakers" to limit future damage. Circuit breakers, which were first introduced after the 1987 Black Monday crash, halt trading if stock prices drop too quickly. Market-wide circuit breakers trip if the S&P 500 drops more than 7 percent, 13 percent or 20 percent from the closing price the previous day, temporarily pausing trading or, in the event of a 20 percent drop, shutting down markets for the day. After the Flash Crash, in 2012 the Securities and Exchange Commission introduced new "limit up–limit down" circuit breakers for individual stocks to prevent sharp, dramatic price swings. The limit up–limit down mechanism creates a price band around a stock, based on the stock's average price over the preceding five

minutes. If the stock price moves out of that band for more than fifteen seconds, trading is halted on that stock for five minutes.

Circuit breakers are an important mechanism for preventing flash crashes from causing too much damage. We know this because they keep getting tripped. An average day sees a handful of circuit breakers tripped due to rapid price moves. One day in August 2015, over 1,200 circuit breakers were tripped across multiple exchanges. Mini-flash crashes have continued to be a regular, even normal event on Wall Street. Sometimes these are caused by simple human error, such as a trader misplacing a zero or using an algorithm intended for a different trade. In other situations, as in the May 2010 flash crash, the causes are more complex. Either way, the underlying conditions for flash crashes remain, making circuit breakers a vital tool for limiting their damage. As Greg Berman, associate director of the SEC's Office of Analytics and Research, explained, "Circuit breakers don't prevent the initial problems, but they prevent the consequences from being catastrophic."

WAR AT MACHINE SPEED

Stock trading is a window into what a future of adversarial autonomous systems competing at superhuman speeds might look like in war. Both involve high-speed adversarial interactions in complex, uncontrolled environments. Could something analogous to a flash crash occur in war—a flash war?

Certainly, if Stanislav Petrov's fateful decision had been automated, the consequences could have been disastrous: nuclear war. Nuclear command and control is a niche application, though. One could envision militaries deploying autonomous weapons in a wide variety of contexts but still keeping a human finger on the nuclear trigger.

Nonnuclear applications still hold risks for accidental escalation. Militaries regularly interact in tense situations that have the potential for conflict, even in peacetime. In recent years, the U.S. military has jockeyed for position with Russian warplanes in Syria and the Black Sea, Iranian fast boats in the Straits of Hormuz, and Chinese ships and air defenses in the South China Sea. Periods of brinksmanship, where nations flex their militaries to assert dominance but without actually firing weapons, are common in international relations. Sometimes tensions escalate to full-blown crises in which war appears imminent, such as the 1962 Cuban Missile Crisis. In such situations, even the tiniest incident can trigger war. In 1914, a lone gunman assassinated Archduke Franz Ferdinand of Austria, sparking a chain of events that led to World War I.

Miscalculation and ambiguity are common in these tense situations, and confusion and accidents can generate momentum toward war. The Gulf of Tonkin incident, which led Congress to authorize the war in Vietnam, was later discovered to be partially false; a purported gun battle between U.S. and Vietnamese boats on August 4, 1964, never occurred.

Robotic systems are already complicating these situations, even with existing technology. In 2013, China flew a drone over the Senkaku Islands, a contested pile of uninhabited rocks in the East China Sea that both China and Japan claim as their own. In response, Japan scrambled an F-15 fighter jet to intercept the drone. Eventually, the drone turned around and left, but afterward Japan issued news rules of engagement for how it would deal with drone incursions. The rules were more aggressive than those for intercepting manned aircraft, with Japan stating they would shoot down any drone entering their territory. In response, China stated that any attack on their drones would be an "act of war" and that China would "strike back."

As drones have proliferated, they have repeatedly been used to broach other nations' sovereignty. North Korea has flown drones into South Korea. Hamas and Hezbollah have flown drones into Israel. Pakistan has accused India of flying drones over the Pakistani-controlled parts of Kashmir (a claim India has denied). It seems one of the first things people do when they get ahold of drones is send them into places they don't belong.

When sovereignty is clear, the typical response has been to simply shoot down the offending drone. Pakistan shot down the alleged Indian drone over Kashmir. Israel has shot down drones sent into its air space. Syria shot down a U.S. drone over its territory in 2015. A few months later, Turkey shot down a presumed Russian drone that penetrated Turkey from Syria.

These incidents have not led to larger conflagrations, perhaps in part because sovereignty in these incidents was not actually in dispute. These were clear cases where a drone was sent into another nation's air space. Within the realm of international relations, shooting it down was seen as a reasonable response. This same action could be perceived very differently in contested areas, however, such as the Senkaku Islands, where both countries assert sovereignty. In such situations, a country whose drone was shot down might feel compelled to escalate in order to back up their territorial claim. Hints of these incidents have already begun. In December 2016, China seized a small underwater robot drone the United States was operating in the South China Sea. China quickly returned it after U.S. protests, but other incidents might not be resolved so easily.

All of these complications are manageable if autonomous systems do what

humans expect them to do. Robots may raise new challenges in war, but humans can navigate these hurdles, so long as the automation is an accurate reflection of human intent. The danger is if autonomous systems do something they aren't supposed to—if humans lose control.

That's already happened with drones. In 2010, a Navy Fire Scout drone wandered 23 miles off course from its Maryland base toward Washington, DC, restricted air space before it was brought back under control. In 2017, an Army Shadow drone flew more than 600 miles after operators lost control, before finally crashing in a Colorado forest. Not all incidents have ended so harmlessly, however.

In 2011, the United States lost control of an RQ-170 stealth drone over western Afghanistan. A few days later, it popped up on Iranian television largely intact and in the hands of the Iranian military. Reports swirled online that Iran had hijacked the drone by jamming its communications link, cutting off contact with its human controllers, and then spoofing its GPS signal to trick it into landing at an Iranian base. U.S. sources called the hacking claim "complete bullshit." (Although after a few days of hemming and hawing, the United States did awkwardly confirm the drone was theirs.) Either way—whatever the cause of the mishap—the United States lost control of a highly valued stealth drone, which ended up in the hands of a hostile nation.

A reconnaissance drone wandering off course might lead to international humiliation and the loss of potentially valuable military technology. Loss of control with a lethal autonomous weapon could be another matter. Even a robot programmed to shoot only in self-defense could still end up firing in situations where humans wished it hadn't. If another nation's military personnel or civilians were killed, it might be difficult to de-escalate tensions.

Heather Roff, a research scientist at Arizona State University who works on ethics and policy for emerging technologies, says there is validity to the concern about a "flash war." Roff is less worried about an "isolated individual platform." Her real concern is "networks of systems" working together in "collaborative autonomy." If the visions of Bob Work and others come true, militaries will field flotillas of robot ships, wolf packs of sub-hunting robots undersea, and swarms of aerial drones. In that world, the consequences of a loss of control could be catastrophic. Roff warned, "If my autonomous agent is patrolling an area, like the border of India and Pakistan, and my adversary is patrolling the same border and we have given certain permissions to escalate in terms of self-defense and those are linked to other systems . . . that could escalate very quickly." An accident like the Patriot fratricides could lead to a firestorm of unintended

lethality.

When I sat down with Bradford Tousley, DARPA's TTO director, I put the question of flash crashes to him. Were there lessons militaries could learn from automated stock trading? Tousley lit up at the mention of high-frequency trading. He was well aware of the issue and said it was one he'd discussed with colleagues. He saw automated trading as a "great analogy" for the challenges of automation in military applications. "What are the unexpected side effects of complex systems of machines that we don't fully understand?" he asked rhetorically. Tousley noted that while circuit breakers were an effective damage control measure in stock markets, "there's no 'time out' in the military."

As interesting as the analogy was, Tousley wasn't concerned about a flash war because the speed dimension was vastly different between stock trading and war. "I don't know that large-scale military impacts are in millisconds," he said. (A millisecond is a thousand microseconds.) "Even a hypersonic munition that might go 700 miles in 20 minutes—it takes 20 minutes; it doesn't take 20 milliseconds." The sheer physics of moving missiles, aircraft, or ships through physical space imposes time constraints on how quickly events can spiral out of control, in theory giving humans time to adapt and respond.

The exception, Tousley said, was in electronic warfare and cyberspace, where interactions occur at "machine speed." In this world, "the speed with which a bad event can happen," he said, "is milliseconds."

14

THE INVISIBLE WAR

AUTONOMY IN CYBERSPACE

In just the past few decades, humans have created an invisible world. We can't see it, but we feel its influence everywhere we go: the buzzing of a phone in our pocket, the chime of an email, the pause when a credit card reader searches the aether for authorization. This world is hidden from us, yet in plain view everywhere. We call it the internet. We call it cyberspace.

Throughout history, technology has enabled humans to venture into inhospitable domains, from undersea to the air and space. As we did, our war-making machines came with us. Cyberspace is no different. In this invisible world of machines operating at machine speed, a silent war rages.

MALICIOUS INTENT

You don't need to be a computer programmer to understand malware. It's the reason you're supposed to upgrade your computer and phone when prompted. It's the reason you're not supposed to click on links in emails from strangers. It's the reason you worry when you hear yet another major corporation has had millions of credit card numbers stolen from their databases. Malware is *malicious software*—viruses, Trojans, worms, botnets—a whole taxonomy of digital diseases.

Viruses have been a problem since the early days of computers, when they were transmitted via floppy disk. Once computers were networked together, worms emerged, which actively transmit themselves over networks. In 1988, the

first large-scale worm—at the time called the Internet Worm because it was the first—spread across an estimated 10 percent of the internet. The internet was pretty small then, only 60,000 computers, and the Internet Worm of 1988 didn't do much. Its intent was to map the internet, so all it did was replicate itself, but it still ended up causing significant harm. Because there was no safety mechanism in place to prevent the worm from copying itself multiple times onto the same machine, it ended up infecting many machines with multiple copies, slowing them down to the point of being unusable.

Today's malware is more sophisticated. Malware is used by governments, criminals, terrorists, and activists ("hacktivists") to gain access to computers for a variety of purposes: conducting espionage, stealing intellectual property, exposing embarrassing secrets, slowing down or denying computer usage, or simply creating access for future use. The scope of everyday cyber activity is massive. In 2015, the U.S. government had over 70,000 reported cybersecurity incidents on government systems, and the number has been rising every year. The most frequent and the most serious attacks came from other governments. Many attacks are relatively minor, but some are massive in scale. In July 2015, the U.S. government acknowledged a hack into the Office of Personnel Management (OPM) that exposed security clearance investigation data of 21 million people. The attack was widely attributed to China, although the Chinese government claimed it was the work of criminals operating from within China and not officially sanctioned by the government.

Other cyberattacks have gone beyond espionage. One of the first widely recognized acts of "cyberwar" was a distributed denial of service (DDoS) attack on Estonia in 2007. DDoS attacks are designed to shut down websites by flooding them with millions of requests, overwhelming bandwidth and denying service to legitimate users. DDoS attacks frequently use "botnets," networks of "zombie" computers infected with malware and harnessed to launch the attack.

Following a decision to relocate a Soviet war memorial, Estonia was besieged with 128 DDoS attacks over a two-week period. The attacks did more than take websites offline; they affected Estonia's entire electronic infrastructure. Banks, ATMs, telecommunications, and media outlets were all shut down. At the height of the DDoS attacks on Estonia, over a million botnet-infected computers around the globe were directed toward Estonian websites, pinging them four million times a second, overloading servers and shutting down access. Estonia accused the Russian government, which had threatened "disastrous" consequences if Estonia removed the monument, of being behind the attack. Russia denied involvement at the time, although two years later a

Russian Duma official confirmed that a government-backed hacker group had conducted the attacks.

In the years since, there have been many alleged or confirmed cyberattacks between nations. Russian government-backed hackers attacked Georgia in 2008. Iran launched a series of cyberattacks against Saudi Arabia and the United States in 2012 and 2013, destroying data on 30,000 computers owned by a Saudi oil company and carrying out 350 DDoS attacks against U.S. banks. While most cyberattacks involve stealing, exposing, or denying data, some have crossed into physical space. In 2010, a worm came to light that crossed a cyber-Rubicon, turning 1s and 0s into physical destruction.

STUXNET: THE CYBERSHOT HEARD ROUND THE WORLD

In the summer of 2010, word began to spread through the computer security world of something new, a worm unlike any other. It was more advanced than anything seen before, the kind of malware that had clearly taken a team of professional hackers months if not years to design. It was a form of malware that security professionals have long speculated was possible but had never seen before: a digital weapon. Stuxnet, as the worm came to be called, could do more than spy, steal things, and delete data. Stuxnet could break things, not just in cyberspace but in the physical world as well.

Stuxnet was a serious piece of malware. Zero-day exploits take advantage of vulnerabilities that software developers are unaware of. (Defenders have known about them for “zero days.”) Zero-days are a prized commodity in the world of computer security, worth as much as $100,000 on the black market. Stuxnet had four. Spreading via removable USB drives, the first thing Stuxnet did when it spread to a new a system was to give itself “root” access in the computer, essentially unlimited access. Then it hid, using a real—not fake—security certificate from a reputable company to mask itself from antivirus software. Then Stuxnet began searching. It spread to every machine on the network, looking for a very particular type of software, Siemens Step 7, which is used to operate programmable logic controllers (PLCs) used in industrial applications. PLCs control power plants, water valves, traffic lights, and factories. They also control centrifuges in nuclear enrichment facilities.

Stuxnet wasn’t just looking for any PLC. Stuxnet operated like a homing munition, searching for a very specific type of PLC, one configured for frequency-converter drives, which are used to control centrifuge speeds. If it

didn’t find its target, Stuxnet went dead and did nothing. If it did find it, then Stuxnet sprang into action, deploying two encrypted “warheads,” as computer security specialists described them. One of them hijacked the PLC, changing its settings and taking control. The other recorded regular industrial operations and played them back to the humans on the other side of the PLC, like a fake surveillance video in a bank heist. While secretly sabotaging the industrial facility, Stuxnet told anyone watching: “everything is fine.”

Computer security specialists widely agree that Stuxnet’s target was an industrial control facility in Iran, likely the Natanz nuclear enrichment facility. Nearly 60 percent of Stuxnet infections were in Iran and the original infections were in companies that have been tied to Iran’s nuclear enrichment program. Stuxnet infections appear to be correlated with a sharp decline in the number of centrifuges operating at Natanz. Security specialists have further speculated that the United States, Israel, or possibly both, were behind Stuxnet, although definitive attribution can be difficult in cyberspace.

Stuxnet had a tremendous amount of autonomy. It was designed to operate on “air-gapped” networks, which aren’t connected to the internet for security reasons. In order to reach inside these protected networks, Stuxnet spread via removable USB flash drives. This also meant that once Stuxnet arrived at its target, it was on its own. Computer security company Symantec described how this likely influenced Stuxnet’s design:

> While attackers could control Stuxnet with a command and control server, as mentioned previously the key computer was unlikely to have outbound Internet access. Thus, all the functionality required to sabotage a system was embedded directly in the Stuxnet executable.

Unlike other malware, it wasn’t enough for Stuxnet to give its designers access. Stuxnet had to perform the mission autonomously.

Like other malware, Stuxnet also had the ability to replicate and propagate, infecting other computers. Stuxnet spread far beyond its original target, infecting over 100,000 computers. Symantec referred to these additional computers as “collateral damage,” an unintentional side effect of Stuxnet’s “promiscuous” spreading that allowed it to infiltrate air-gapped networks.

To compensate for these collateral infections, however, Stuxnet had a number of safety features. First, if Stuxnet found itself on a computer that did not have the specific type of PLC it was looking for, it did nothing. Second, each copy of Stuxnet could spread via USB to only three other machines, limiting the extent of its proliferation. Finally, Stuxnet had a self-termination date. On June 24, 2012, it was designed to erase all copies of itself. (Some experts saw these

safety features as further evidence that it was designed by a Western government.)

By using software to actively sabotage an industrial control system, something cybersecurity specialists thought was possible before Stuxnet but had not yet happened, Stuxnet was the first cyberweapon. More will inevitably follow. Stuxnet is an "open-source weapon" whose code is laid bare online for other researchers to tinker with, modify, and repurpose for other attacks. The specific vulnerabilities Stuxnet exploited will have been fixed, but its design is already being used as a blueprint for cyberweapons to come.

AUTONOMY IN CYBERSPACE

Autonomy is essential to offensive cyberweapons, such as Stuxnet, that are intended to operate on closed networks separated from the internet. Once it arrives at its target, Stuxnet carries out the attack on its own. In that sense, Stuxnet is analogous to a homing munition. A human chooses the target and Stuxnet conducts the attack.

Autonomy is also essential for cyberdefense. The sheer volume of attacks means it is impossible to catch them all. Some will inevitably slip through defenses, whether by using zero-day vulnerabilities, finding systems that have not yet been updated, or exploiting users who insert infected USB drives or click on nefarious links. This means that in addition to keeping malware out, security specialists have also adopted "active cyberdefenses" to police networks on the inside to find malware, counter it, and patch network vulnerabilities.

In 2015, I testified to the Senate Armed Services Committee alongside retired General Keith Alexander, former head of the National Security Agency, on the future of warfare. General Alexander, focusing on cyber threats, explained the challenge in defending 15,000 "enclaves" (separate computer networks) within the Department of Defense. Keeping all of these networks up-to-date manually was nearly impossible. Patching network vulnerabilities at "manual speed," he said, took months. "It should be automated," Alexander argued. "The humans should be out of the loop." Computer security researchers are already working to develop these more sophisticated cyber that would take humans out of the loop. As in other areas of autonomy, DARPA is at the leading edge of this research.

UNLEASHING MAYHEM: THE CYBER GRAND CHALLENGE

DARPA tackles only the most difficult research problems, “DARPA hard” problems that others might deem impossible. DARPA does this every day, but when a technical problem is truly daunting even for DARPA, the organization pulls out its big guns in a Grand Challenge.

The first DARPA Grand Challenge was held in 2004, on autonomous vehicles. Twenty-one research teams competed to build a fully autonomous vehicle that could navigate a 142-mile course across the Mojave Desert. It was truly a “DARPA hard” problem. The day ended with every single vehicle broken down, overturned, or stuck. The furthest any car got was 7.4 miles, only 5 percent of the way through the course.

The organization kept at it, sponsoring a follow-up Grand Challenge the next year. This time, it was a resounding success. Twenty-two vehicles beat the previous year’s distance record and five cars finished the entire course. In 2007, DARPA hosted an Urban Challenge for self-driving cars on a closed, urban course complete with traffic and stop signs. These Grand Challenges matured autonomous vehicle technology in leaps and bounds, laying the seeds for the self-driving cars now in development at companies like Google and Tesla.

DARPA has since used the Grand Challenge approach as a way to tackle other truly daunting problems, harnessing the power of competition to generate the best ideas and launch a technology forward. From 2013 to 2015, DARPA held a Robotics Challenge to advance the field of humanoid robotics, running robots through a set of tasks simulating humanitarian relief and disaster response.

In 2016, DARPA hosted a Cyber Grand Challenge to advance the field of cybersecurity. Over one hundred teams competed to build a fully autonomous Cyber Reasoning System to defend a network. The systems competed in a live capture the flag competition to automatically identify computer vulnerabilities and either patch or exploit them.

David Brumley is a computer scientist at Carnegie Mellon University and CEO of ForAllSecure, whose system Mayhem won the Cyber Grand Challenge. Brumley describes his goal as building systems that “automatically check the world’s software for exploitable bugs.” Mayhem is that vision brought to life, a “fully autonomous system for finding and fixing computer security vulnerabilities.” In that sense, Mayhem is even more ambitious than Keith Alexander’s goal of just updating software automatically. Mayhem actually goes and finds bugs on its own—bugs that humans are not yet aware of— and then patches them.

Brumley explained to me that there are actually several steps in this process.

The first is finding a vulnerability in a piece of software. The next step is developing either an “exploit” to take advantage of the vulnerability or a “patch” to fix it. If a vulnerability is analogous to a weak lock, then an exploit is like a custom-made key to take advantage of the lock’s weakness. A patch, on the other hand, fixes the lock.

Developing these exploits and patches isn’t enough, though. One has to know when to use them. Even on the defensive side, Brumley explained, you can’t just apply a patch as soon as you see an exploit. For any given vulnerability, Mayhem would develop a “suite of patches.” Fixing a vulnerability isn’t a binary thing, where either it’s fixed or it isn’t. Brumley said, “There’s grades of security, and often these have different tradeoffs on performance, maybe even functionality.” Some patches might be more secure, but would cause the system to run slower. Which patch to apply depends on the system’s use. For home use, “you’d rather have it more functional rather than 100 percent secure,” Brumley said. A customer protecting critical systems, on the other hand, like the Department of Defense, might choose to sacrifice efficiency for better security. When to apply the patch is another factor to consider. “You don’t install a Microsoft PowerPoint update right before a big business presentation,” Brumley said.

Today, these steps are all done by people. People find the vulnerabilities, design the patches, and upload them to an automatic update server. Even the “auto-update” functions on your home computer are not actually fully automatic. You have to click “Okay” in order for the update to move forward. Every place where there is a human in the loop slows down the process of finding and patching vulnerabilities. Mayhem, on the other hand, is a completely autonomous system for doing all those steps. That means it isn’t just finding and patching vulnerabilities blindly. It’s also reasoning about which patch to use and when to apply it. Brumley said it’s “an autonomous system that’s taking all of those things that humans are doing, it’s automating them, and then it’s reasoning about how to use them, when to apply the patch, when to use the exploit.” Mayhem also deploys hardening techniques on programs. Brumley described these as proactive security measures applied to a program before a vulnerability has even been discovered to make it harder to exploit, if there are vulnerabilities. And Mayhem does all of this at machine speed.

In the Cyber Grand Challenge final round, Mayhem and six other systems competed in a battle royale to scan each other’s software for vulnerabilities, then exploit the weaknesses in other systems while patching their own vulnerabilities. Brumley compared the competition to seven fortresses probing each together,

trying to get into locked doors. “Our goal was to come up with a skeleton key that let us in when it wasn’t supposed to.” DARPA gave points for showing a “proof of vulnerability,” essentially an exploit or “key,” to get into another system. The kind of access also mattered—full access into the system gave more points than more limited access that was only useful for stealing information.

Mike Walker, the DARPA program manager who ran the Cyber Grand Challenge, said that the contest was the first time that automated cybertools had moved beyond simply applying human-generated code and into the “automatic creation of knowledge.” By autonomously developing patches, they had moved beyond automated antivirus systems that can clean up known malware to “automation of the supply chain.” Walker said, “true autonomy in the cyber domain are systems that can create their own knowledge. . . . It’s a pretty bright and clear line. And I think we kind of crossed it . . . for the first time in the Cyber Grand Challenge.”

Walker compared the Cyber Grand Challenge to the very first chess tournaments between computers. The technology isn’t perfect. That wasn’t the point. The goal was to prove the concept to show what can be done and refine the technology over time. Brumley said Mayhem is roughly comparable to a “competent” computer security professional, someone “just fresh out of college in computer security.” Mayhem has nothing on world-class hackers. Brumley should know. He also runs a team of competitive human hackers who compete in the DEF CON hacking conference, the “world series” of hacking. Brumley’s team from Carnegie Mellon has won four out of the past five years.

Brumley’s aim with Mayhem isn’t to beat the best human hackers, though. He has something far more practical—and transformative—in mind. He wants to fundamentally change computer security. As the internet colonizes physical objects all around us—bringing toasters, watches, cars, thermostats and other household objects online in the Internet of Things (IoT), this digitization and connectivity also bring vulnerabilities. In October 2016, a botnet called Mirai hijacked everyday networked devices such as printers, routers, DVR machines, and security cameras and leveraged them for a massive DDoS attack. Brumley said most IoT devices are “ridiculously vulnerable.” There are an estimated 6.4 billion IoT devices online today, a number expected to grow to over 20 billion devices by 2020. That means there are millions of different programs, all with potential vulnerabilities. “Every program written is like a unique lock and most of those locks have never been checked to see if they’re terrible,” Brumley said. For example, his team looked at 4,000 commercially available internet routers and “we’ve yet to find one that’s secure,” he said. “No one’s ever bothered to

check them for security." Checking this many devices at human speed would be impossible. There just aren't enough computer security experts to do it. Brumley's vision is an autonomous system to "check all these locks."

Once you've uncovered a weak lock, patching it is a choice. You could just as easily make a key—an exploit—to open the lock. There's "no difference" between the technology for offense and defense, Brumley said. They're just different applications of the same technology. He compared it to a gun, which could be used for hunting or to fight wars. Walker agreed. "All computer security technologies are dual-use," he said.

For safety reasons, DARPA had the computers compete on an air-gapped network that was closed off from the internet. DARPA *also* created a special operating system just for this contest. Even if one of the systems was plugged into the internet, it would need to be re-engineered to search for vulnerabilities on a Windows, Linux, or Mac machine.

Brumley emphasized that they've never had a problem with people using this technology for nefarious ends at Carnegie Mellon. He compared his researchers to biologists working on a better flu vaccine. They could use that knowledge to make a better virus, but "you have to trust the researchers to have appropriate safety protocols." His company, ForAllSecure, practices "responsible disclosure" and notifies companies of vulnerabilities they find. Nonetheless, he admitted, "you do worry about the bad actors."

Brumley envisions a world where over the next decade, tools like Mayhem are used to find weak locks and patch them, shoring up cyberdefenses in the billions of devices online. Walker said that self-driving cars today are a product of the commercial sector throwing enormous investment money behind the individuals who competed in the original DARPA Grand Challenge a decade ago, and he sees a similar road ahead for autonomous cybersecurity. "It's going to take the same kind of long-term will and financial backing to do it again here."

Both Brumley and Walker agreed that autonomous cybertools will also be used by attackers, but they said the net effect was to help the defense more. Right now, "offense has all of the advantage in computer security," Walker said. The problem is right now there is an asymmetry between attackers and defenders. Defenders have to close all of the vulnerabilities, while attackers have to just find one way in. Autonomous cybersystems level the playing field, in part because defense gets a first-mover advantage. They write the code, so they can scan it for vulnerabilities and patch them before it is deployed. "I'm not saying that we can change to a place where defense has the advantage," Walker said,

but he did think autonomous cybertools would enable “investment parity,” where “the best investment wins.” Even that would be “transformative,” he said. There’s big money in malware, but far more is spent annually on computer security. Prior to joining DARPA, Walker said he worked for a decade as a “red teamer,” paid by energy and financial sector companies to hack into their systems and uncover their vulnerabilities. He said autonomous cyberdefenses “can actually make hacking into something like our energy infrastructure or our financial infrastructure a highly uncommon proposition that average criminals cannot afford to do.”

David Brumley admitted that this won’t stop hacking from advanced nation-states who have ample resources. He said limiting access was still beneficial, though, and drew a comparison to efforts to limit the spread of nuclear weapons: “It’s scary to think of Russia and the U.S. having it, but what’s really scary is when the average Joe has it. We want to get rid of the average Joe having these sorts of things.” If Brumley is right, autonomous systems like Mayhem will make computers more secure and safer ten years from now. But autonomy will keep evolving in cyberspace, with even more advanced systems beyond Mayhem yet to come.

The next evolution in autonomous cyberdefense is what Brumley calls “counter-autonomy.” Mayhem targets weak locks; counter-autonomy targets the locksmith. It “leverages flaws or predictable patterns in the adversary to win.” Counter-autonomy goes beyond finding exploits, he said; it’s about “trying to find vulnerabilities in the opponent’s algorithms.” Brumley compared it to playing poker: “you play the opponent.” Counter-autonomy exploits the brittleness of the enemy’s autonomous systems to defeat them.

While counter-autonomy was not part of the Cyber Grand Challenge, Brumley said they have experimented with counter-autonomy techniques that they simply didn’t use. One tool they developed embeds a hidden exploit targeting a competitor’s autonomous system into a patch. “It’s a little bit like a Trojan horse,” Brumely said. The patch “works just fine. It’s a legitimate program.” Hidden within the patch is an exploit, though, that targets one of the common tools that hackers use to analyze patches. “Anyone who tries to analyze [the patch] gets exploited,” he said. Another approach to counter-autonomy would move beyond simply finding vulnerabilities to actually creating them. This could be done in learning systems by inserting false data into the learning process. Brumley calls this the “computer equivalent to ‘the long con,’ where our systems methodically cause our adversary’s systems to ‘mis-learn’ (incorrectly learn) how to operate.”

AUTONOMOUS CYBERWEAPONS

The arms race in speed in cyberspace is already under way. In an unpublished 2016 working paper, Brumley wrote, “Make no mistake, cyber is a war between attackers and defenders, both who coevolve as the other deploys new systems and measures. In order to win, we must act, react, and evolve faster than our adversaries.” Cyberweapons of the future—defensive and offensive—will incorporate greater autonomy, just the same way that more autonomy is being integrated into missiles, drones, and physical systems like Aegis. What would a “cyber autonomous weapon” look like?

Cyberspace and autonomous weapons intersect in a number of potentially significant ways. The first is the danger that cyber vulnerabilities pose in autonomous weapons. Anything that is computerized is vulnerable to hacking. The migration of household objects online as part of the IoT presents major cybersecurity risks, and there are analogous risks for militaries whose major platforms and munitions are increasingly networked. Cyber vulnerabilities could hobble a next-generation weapon system like the F-35 Joint Strike Fighter, which has tens of millions of lines of code. There is no reason to think that an autonomous weapon would necessarily be more vulnerable to hacking, but the consequences if one were hacked could be much worse. Autonomous weapons would be a very attractive target for a hostile state’s malware, since a hacker could potentially usurp control of an autonomous weapon and redirect it. The consequences could be even worse than those of a runaway gun. The weapon wouldn’t be out of control; it would be under the control of the enemy.

In theory, greater autonomy that allows for off-network operation may appear to be a solution to cyber vulnerabilities. This is an appealing tactic that has come up in science fiction wars between humans and machines. In the opening episode of the 2003 reboot of *Battlestar Galactica*, the evil Cylon machines wipe out nearly the entire human space fleet via a computer virus. The ship *Galactica* survives only because it has an older computer system that is not networked to the rest of the fleet. As Stuxnet demonstrated, however, in the real world operating off-network complicates cyberattacks but is no guarantee of immunity.

The second key intersection between cyberspace and autonomy occurs in automated “hacking back.” Autonomous cyberbots like Mayhem will be part of active cyberdefenses, including those that use higher-level reasoning and decision-making, but these still operate within one’s own network. Some concepts for active cyber defense move beyond policing one’s own networks

into going on the offense. Hacking back is when an organization responds to a cyberattack by counterattacking, gaining information about the attacker or potentially shutting down the computers from which the attack is originating. Because many cyberattacks involve co-opting unsuspecting "zombie" computers and repurposing them for attack, hacking back can inevitably draw in third parties. Hacking back is controversial and, if done by private actors, could be illegal. As one cybersecurity analyst noted, "Every action accelerates."

Automation has been used in some limited settings when hacking back. When the FBI took down the Coreflood botnet, it redirected infected botnet computers to friendly command-and-control servers, which then issued an automatic stop command to them. However, this is another example of automation being used to execute a decision made by people, which is far different than delegating the decision whether or not to hack back to an autonomous process.

Automated hacking back would delegate the decision whether or not to go on the counteroffensive to an autonomous system. Delegating this authority could be very dangerous. Patrick Lin, an ethicist at California Polytechnic State University who has written extensively on autonomy in both military and civilian applications, warned at the United Nations in 2015, "autonomous cyber weapons could automatically escalate a conflict." As Tousley acknowledged, cyberspace could be an area where automatic reactions between nation-states happen in milliseconds. Automated hacking back could cause a flash cyberwar that rapidly spirals out of control. Automated hacking back is a theoretical concept, and there are no publicly known examples of it occurring. (Definitively saying something has *not* happened in cyberspace is difficult, given the shadowy world of cyberwar.)

The third intersection between cyber-and autonomous weapons is increasingly autonomous offensive cyberweapons. Computer security researchers have already demonstrated the ability to automate "spear phishing" attacks, in which unwitting users are sent malicious links buried inside seemingly innocuous emails or tweets. Unlike regular phishing attacks, which target millions of users at a time with mass emails, spear phishing attacks are specially tailored to specific individuals. This makes them more effective, but also more time-intensive to execute. Researchers developed a neural network that, drawing on data available on Twitter, learned to automatically develop "humanlike" tweets targeted at specific users, enticing them to click on malicious links. The algorithm was roughly as successful as manual spear phishing attempts but, because of automation, could be deployed en masse to

automatically seek out and target vulnerable users.

As in other areas, greater intelligence will allow offensive cyberweapons to operate with greater autonomy. Stuxnet autonomously carried out its attack, but its autonomy was highly constrained. Stuxnet had a number of safeguards in place to limit its spread and effects on computers that weren't its target, as well as a self-termination date. One could envision future offensive cyberweapons that were given freer rein. Eric Messinger, a writer and researcher on legal issues and human rights, has argued:

> . . . in offensive cyberwarfare, [autonomous weapon systems] may *have* to be deployed, because they will be integral to effective action in an environment populated by automated defenses and taking place at speeds beyond human capacities. . . . [The] development and deployment of offensive [autonomous weapon systems] may well be unavoidable.

It's not clear what an offensive autonomous cyberweapon would look like, given the challenges in both defining a "cyberweapon" and the varying ways in which autonomy is already used in cyberspace. From a certain perspective, a great deal of malware is inherently autonomous by virtue of its ability to self-replicate. The Internet Worm of 1988, for example, is an example of the Sorcerer's Apprentice effect: a runaway, self-replicating process that cannot be stopped. This is an important dimension to malware that does not have an analogy in physical weapons. Drones and robotic systems cannot self-replicate. In this sense, malware resembles biological viruses and bacteria, which self-replicate and spread from host to host.

But there is a critical difference between digital and biological viruses. Biological pathogens can mutate and adapt in response to environmental conditions. They evolve. Malware, at least today, is static. Once malware is deployed, it can spread, it can hide (as Stuxnet did), but it cannot modify itself. Malware can be designed to look for updates and spread these updates among copies of itself via peer-to-peer sharing (Stuxnet did this as well), but new software updates originate with humans.

In 2008, a worm called Conficker spread through the internet, infecting millions of computers. As computer security specialists moved to counter it, Conficker's designers released updates, eventually fielding as many as five different variants. These updates allowed Conficker's programmers to stay ahead of security specialists, upgrading the worm and closing vulnerabilities when they were detected. This made Conficker a devilishly hard worm to defeat. At one point, an estimated 8 to 15 million computers worldwide were infected.

Conficker used a mixture of human control and automation to stay ahead of antivirus specialists. Conficker's updates came from its human designers, but it

used automation to get the updates clandestinely. Every day, Conficker would generate hundreds of new domain names, only one of which would link back to its human controllers with new updates. This made the traditional approach of blocking domains to isolate the worm from its controllers ineffective. As security specialists found a method to counter Conficker, a new variant would be released quickly, often within weeks. Eventually, a consortium of industry experts brought Conficker to heel, but doing so took a major effort.

Conficker's fundamental weakness was that its updates could only happen at human speed. Conficker replicated autonomously and used clever automation to surreptitiously link back to its human controllers, but the contest between the hackers and security specialists was fought at human speed. Humans were the ones working to identify the worm's weaknesses and take it down, and humans on the other side were working to adapt the worm and keep it one step ahead of antivirus companies.

The technology that Mayhem represents could change that. What if a piece of software turned the same tools for identifying and patching vulnerabilities and applied them to itself? It could improve itself, shoring up its own defenses and resisting attack. Brumley has hypothesized about such "introspective systems." Self-adapting software that can modify itself, rather than wait on updates from its human controllers, would be a significant evolution. The result could be robust cyberdefenses . . . or resilient malware. At the 2015 International Conference on Cyber Conflict, Alessandro Guarino hypothesized that AI-based offensive cyberweapons could "prevent and react to countermeasures," allowing them to persist inside networks. Such an agent would be "much more resilient and able to repel active measures deployed to counter it."

A worm that could autonomously adapt—mutating like a biological virus, but at machine speed—would be a nasty bug to kill. Walker cautioned that the tools used in the Cyber Grand Challenge would only allow a piece of software to patch its own vulnerabilities. It wouldn't allow "the synthesis of new logic" to develop "new code that can work towards a goal." To do that, he said, "first we'd have to invent the field of code synthesis, and right now, it's like trying to predict when time travel's going to be invented. Who knows if it can be invented? We don't have a path." While such a development would be a leap beyond current malware, the advent of learning systems in other areas, such as Google DeepMind's Atari-playing AI or AlphaGo, suggests that it is not inconceivable. Adaptive malware that could rewrite itself to hide and avoid scrutiny at superhuman speeds could be incredibly virulent, spreading and mutating like a biological virus without any form of human control.

When I asked Brumley about the possibility of future malware that was adaptive, he said “those are a possibility and are worrisome. . . . I think someone could come up with this kind of ultimate malware and it could get out of control and it would be a really big pain for a while.” What he really worries about, though, are near-term problems. His chief concern is a shortage of cybersecurity experts. We have weak cyber locks because we’re not training enough people how to be better cyber locksmiths. Part of this, Brumley said, is a culture that views hacking as an illegitimate profession. “In the U.S., we’ve shot ourselves in the foot by equating a hacker with a bad guy.” We don’t view flesh-and-blood locksmiths that way, yet for digital security, we do. Other countries don’t see it that way, and Brumley worries the United States is falling behind. He said, “There’s this kind of hubris in the U.S. that we think that because we have the best Army and Navy and we have all these great amazing natural resources, great aircraft carriers, that of course we’re going to dominate in cyber. And I don’t think that’s a given. It’s a brand-new space, completely different from anything else. There’s no reason that things will just carry over.” We need to shift the culture in the United States, he said, from thinking about hacking skills as something that are only used for “offense and should be super-secret and only used by the military” to something that is valued in the cyber workforce more broadly. Walker agreed. “Defense is powered by openness,” he said.

Looking to the future, Brumley said he saw the “ecosystem” we were building for computer security and autonomous cybersystems as critical. “I tend to view everything as a system—a dynamic system.” People are part of that system too. The solution to potentially dangerous malware in the future was to create “the right ecosystem . . . and then it will be resilient to problems.”

KEEPING THE BOTS AT BAY

Mixing cyberspace and autonomous weapons combines two issues that are challenging enough by themselves. Cyberwarfare is poorly understood outside the specialist community of cyber experts, in part because of the secrecy surrounding cyber operations. Norms about appropriate behavior between states in cyberspace are still emerging. There is not even a consensus among cyber experts about what constitutes a “cyberweapon.” The concept of autonomous weapons is similarly nascent, making the combination of these two issues extremely difficult to understand. The DoD’s official policy on autonomy in weapons, DoD Directive 3000.09, specifically exempts cyberweapons. This wasn’t because we thought autonomous cyberweapons were uninteresting or

unimportant when we wrote the directive. It was because we knew bureaucratically it would be hard enough simply to create a new policy on autonomy. Adding cyber operations would have multiplied the complexity of the problem, making it very likely we would have accomplished nothing at all.

This lack of clarity is reflected in the mixed signals I got from Defense Department officials on autonomy in cyberspace. Both Work and Tousley mentioned electronic warfare and cyberspace as an arena in which they would be willing to accept more autonomy, but they had different perspectives on how far they would be willing to go. Tousley said he saw a role for autonomy in only defensive cyber operations. The "goal is not offense—it's defense," he told me.

Tousley's boss's boss, Deputy Secretary Bob Work, saw things differently. Work made a direct comparison between Aegis and automated "hacking back." He said, "the narrow cases where we will allow the machine to make targeting decisions is in defensive cases where all of the people who are coming at you are bad guys. . . . electronic warfare, cyberwarfare, missile defense. . . . We will allow the machine to make essentially decisions . . . like, a cyber counter attack." He acknowledged delegating that kind of authority to a machine came with risks. Work outlined a hypothetical scenario where this approach could go awry: "A machine might launch a cyber counterattack and it might . . . wind up killing [an industrial control] system or something . . . say it's an airplane and the airplane crashes. And we didn't make a determination that we were going to shoot down that airplane. We just said, 'We're under cyberattack. We're going to counterattack.' Boom."

Work's response to this risk isn't to hide from the technology, but rather to wrestle with these challenges. He explained the importance of consulting with scientists, ethicists, and lawyers. "We'll work it through," he said. "This is all going to be about the checks and balances that you put inside your battle networks." Work was confident these risks could be managed because in his vision, humans would still be involved in a number of ways. There would be both automated safeties and human oversight. "We always emphasize human-machine collaboration . . . with the human always in front," he said. "That's the ultimate circuit breaker."

AN ARMS RACE TO WHERE?

Sun Tzu wrote over two thousand years ago in *The Art of War*, "Speed is the essence of war." His maxim remains even more true today, when signals can

cross the globe in fractions of a second. Human decision-making has many advantages over machine intelligence, but humans cannot compete at machine speed. Competitive pressures in fast-paced environments threaten to push humans further and further out of the loop. Superhuman reaction times are the reason why automatic braking is being integrated into cars, why many nations employ Aegis-like automated defensive systems, and why high-frequency stock trading is such a lucrative endeavor.

With this arms race in speed comes grave risks. Stock trading is one example of a field in which competitors have succumbed to allure of speed, developing ever-faster algorithms and hardware to shave microseconds from reaction times. In uncontrolled, real-world environments, the (unsurprising) result has been accidents. When these accidents occur, machine speed becomes a major liability. Autonomous processes can rapidly spiral out of control, destroying companies and crashing markets. It's one thing to say that humans will have the ability to intervene, but in some settings, their intervention may be too late. Automated stock trading foreshadows the risks of a world where nations have developed and deployed autonomous weapons.

A flash physical war in the sense of a war that spirals out of control in mere seconds seems unlikely. Missiles take time to move through the air. Sub-hunting undersea robots can move only so quickly through the water. Accidents with autonomous weapons could undermine stability and escalate crises unintentionally, but these incidents would likely take place over minutes and hours, not microseconds. This is not to say that autonomous weapons do not pose serious risks to stability; they do. A runaway autonomous weapon could push nations closer to the brink of war. If an autonomous weapon (or a group of them) caused a significant number of deaths, tensions could boil over to the point where de-escalation is no longer possible. The speed at which events would unfold, however, is likely one that would allow humans to see what was happening and, at the very least, take steps to attempt to mitigate the effects. Bob Work told me he saw a role for a human "circuit breaker" in managing swarms of robotic systems. If the swarm began to behave in an unexpected way, "they would just shut it down," he said. There are problems with this approach. The autonomous system might not respond to commands to shut it down, either because it is out of communications or because the type of failure it is experiencing prevents it from accepting a command to shut down. Unless human operators have physical access, like the physical circuit breaker in Aegis, any software-based "kill switch" is susceptible to the same risks as other software—bugs, hacking, unexpected interactions, and the like.

Even though accidents with physical autonomous weapons will not cascade into all-out war in mere seconds, machines could quickly cause damage that might have irreversible consequences. Countries may not believe that an enemy's attack was an accident, or the harm may be so severe that they simply don't care. If Japan had claimed that the attack on Pearl Harbor was not authorized by Tokyo and was the work of a single rogue admiral, it's hard to imagine the United States would have refrained from war.

A flash cyberwar, on the other hand, is a real possibility. Automated hacking back could lead to escalation between nations in the blink of an eye. In this environment, human oversight would be merely the illusion of safety. Automatic circuit breakers are used to stop flash crashes on Wall Street because humans cannot possibly intervene in time. There is no equivalent referee to call "Time out" in war.

15

“SUMMONING THE DEMON”

THE RISE OF INTELLIGENT MACHINES

Even the most sophisticated machine intelligence today is a far cry from the sentient AIs depicted in science fiction. Autonomous weapons pose risks precisely because today’s narrow AIs fail miserably at tasks that require general intelligence. Machines can crush humans at chess or *go*, but cannot enter a house and make a pot of coffee. Image recognition neural nets can identify objects, but cannot piece these objects together into a coherent story about what is happening in a scene. Without a human’s ability to understand context, a stock-trading AI doesn’t understand that it is destroying its own company. Some AI researchers are pondering a future where these constraints no longer exist.

Artificial general intelligence (AGI) is a hypothetical future AI that would exhibit human-level intelligence across the full range of cognitive tasks. AGI could be applied to solving humanity’s toughest problems, including those that involve nuance, ambiguity, and uncertainty. An AGI could, like Stanislav Petrov, step back to consider the broader context and apply judgment.

What it would take to build such a machine is a matter of pure speculation, but there is at least one existence proof that general intelligence is possible: us. Even if recent advances in deep neural networks and machine learning come up short, eventually an improved understanding of the human brain should allow for a detailed neuron-by-neuron simulation. Brain imaging is improving quickly and some researchers believe whole brain emulations could be possible with supercomputers as early as the 2040s.

Experts disagree wildly on when AGI might be created, with estimates

ranging from within the next decade to never. A majority of AI experts predict AGI could be possible by 2040 and likely by the end of the century, but no one really knows. Andrew Herr, who studies emerging technologies for the Pentagon, observed, “When people say a technology is 50 years away, they don’t really believe it’s possible. When they say it’s 20 years away, they believe it’s possible, but they don’t know how it will happen.” AGI falls into the latter category. We know general intelligence is possible because humans have it, but we understand so little of our own brains and our own intelligence that it’s hard to know how far away it is.

THE INTELLIGENCE EXPLOSION

AGI would be an incredible invention with tremendous potential for bettering humanity. A growing number of thinkers are warning, however, that AGI may be the “last invention” humanity creates—not because it will solve all of our problems, but because it will lead to our extermination. Stephen Hawking has warned, “development of full artificial intelligence could spell the end of the human race.” Artificial intelligence could “take off on its own and re-design itself at an ever-increasing rate,” he said. “Humans, who are limited by slow biological evolution, couldn’t compete, and would be superseded.”

Hawking is a cosmologist who thinks on time scales of tens of thousands or millions of years, so it might be easy to dismiss his concerns as a long way off, but technologists thinking on shorter time scales are similarly concerned. Bill Gates has proclaimed the “dream [of artificial intelligence] is finally arriving,” a development that will usher in growth and productivity in the near term, but has long-term risks. “First the machines will do a lot of jobs for us and not be super intelligent,” Gates said. “That should be positive if we manage it well. A few decades after that, though, the intelligence is strong enough to be a concern.” How much of a concern? Elon Musk has described the creation of human-level artificial intelligence as “summoning the demon.” Bill Gates has taken a more sober tone, but essentially agrees. “I am in the camp that is concerned about superintelligence,” he said. “I agree with Elon Musk and some others on this and don’t understand why some people are not concerned.”

Hawking, Gates, and Musk are not Luddites and they are not fools. Their concerns, however fanciful-sounding, are rooted in the concept of an “intelligence explosion.” The concept was first outlined by I. J. Good in 1964:

> Let an ultraintelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual

> activities, an ultraintelligent machine could design even better machines; there would then unquestionably be an "intelligence explosion," and the intelligence of man would be left far behind. Thus the first ultraintelligent machine is the last invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control.

If this hypothesis is right, then humans don't need to create superintelligent AI directly. Humans might not even be capable of such an endeavor. All humans need to do is create an initial "seed" AGI that is capable of building a slightly better AI. Then through a process of recursive self-improvement, the AI will lift itself up by its own bootstraps, building ever-more-advanced AIs in a runaway intelligence explosion, a process sometimes simply called "AI FOOM."

Experts disagree widely about how quickly the transition from AGI to artificial superintelligence (sometimes called ASI) might occur, if at all. A "hard takeoff" scenario is one where AGI evolves to superintelligence within minutes or hours, rapidly leaving humanity in the dust. A "soft takeoff" scenario, which experts see as more likely (with the caveat that no one really has any idea), might unfold over decades. What happens next is anyone's guess.

UNSHACKLING FRANKENSTEIN'S MONSTER

In the *Terminator* movies, when the military AI Skynet becomes self-aware, it decides humans are a threat to its existence and starts a global nuclear war. *Terminator* follows in a long tradition of science fiction creations turning on their masters. In Ridley Scott's *Blade Runner*, based on the Philip K. Dick novel *Do Androids Dream of Electric Sheep?*, Harrison Ford plays a cop tasked with hunting down psychopathic synthetic humans called "replicants." In Harlan Ellison's 1967 short story "I Have No Mouth and I Must Scream," a military supercomputer exterminates all of humanity save for five survivors, whom it imprisons underground and tortures for eternity. Even the very first robots turned on their maker. The word "robot" comes from a 1920 Czech play, *R.U.R.*, for *Rossumovi Univerzální Roboti* (Rossum's Universal Robots), in which synthetic humans called *roboti* ("robot" in English) rise up against their human masters.

The science fiction theme of artificial humans rebelling against their makers is so common it has become known as the "Frankenstein complex," after Mary Shelley's nineteenth-century horror novel, *Frankenstein*. In it, Dr. Frankenstein, through the miracles of science, creates a humanlike creature cobbled together from leftover parts from "the dissecting room and the slaughter-house." The monster turns on Dr. Frankenstein, stalking him and eventually murdering his

new bride.

The fear that hubris can lead to uncontrollable creations has ancient roots that predate even *Frankenstein*. Jewish legend tells of a creature called a golem, molded from clay and brought to life by placing a *shem*, a Hebrew inscription containing one of the names of God, on the golem. In one such legend, Rabbi Judah Loew ben Bezalel of Prague molded a golem from the clay of Prague's riverbanks in the sixteenth century to protect the Jewish community from anti-Semitic attacks. Golems, unlike later intelligent creations, were powerful but stupid beings that would slavishly follow orders, often to the detriment of their creators. Golem stories often end with the golem killing its creator, a warning against the hubris of playing God.

Human-level or superhuman AI tap into this deep well of fear of artificial beings. Micah Clark, a research scientist from the Florida Institute for Human & Machine Cognition who studies AI, cognition, and theory of mind, told me that at "a very personal and philosophical level, AI has been about building persons. . . . It's not about playing chess or driving cars." He explained, "With the general track of robotics and autonomous systems today, you would end up with autonomous systems that are capable but very, very dumb. They would lack any real sense of intelligence. They would be effectively teleoperated, just at a higher level of commanding." Artificial general intelligence—what Clark calls "the dream of AI"—is about "personhood."

Clark's vision of AGI isn't a fearful one, however. He envisions "the kind of persons that we would have intellectual, social, and emotional relationships with, that can experience life with us." AI has been a lifelong passion for Micah Clark. As a child, he played computer games in his grandfather's accounting office and a chess program in particular captured his imagination. The chess AI "destroyed" him, Clark said, and he wanted it to be able to teach him how to play better. Clark was looking for more than just a game, though. "I saw this potential for entertainment and friendship there, but the interaction side was pretty weak," he explained. In college, Clark worked at NASA's Jet Propulsion Laboratory on a large-scale AI demonstration project and he was hooked. Clark went on to study long-duration autonomy for interplanetary robotic spacecraft, but his research interests have moved beyond robotics, sensing, and actuation. The books on Clark's desk in his office have titles like *An Anatomy of the Mind* and *Consciousness and the Social Brain.* Clark described the goal of AI research as "building humanlike persons that can participate in human physical and social spaces and relationships." (Clark is currently working for the Office of Naval Research and he is quick to caveat that these are not the goals of AI research in

the Navy or the Department of Defense. Rather, these are the goals of the field of AI research as a whole.)

Clark's vision of the future of AI is less *Terminator* and more like the movie *Her*. In *Her*, Joaquin Phoenix plays an awkward loner named Theodore who starts a relationship with an AI operating system called "Samantha." Theodore and Samantha develop a close bond and fall in love. Theodore is shaken, however, when Samantha admits that she is simultaneously carrying on relationships with thousands of other people and is also in love with 641 of them. When Theodore breaks down, telling her "that's insane," she tries to lovingly explain, "I'm different from you."

The *otherness* of artificial persons—beings like humans, but also fundamentally different—is a source of much of the fear of AI. Clark explained that AIs will need the ability to interact with humans and that involves abilities like understanding natural language, but that doesn't mean that the AI's behavior or the underlying processes for their intelligence will mirror humans'. "Why would we expect a silica-based intelligence to look or act like human intelligence?" he asked.

Clark cited the Turing test, a canonical test of artificial intelligence, as a sign of our anthropocentric bias. The test, first proposed by mathematician Alan Turing in 1950, attempts to assess whether a computer is truly intelligent by its ability to imitate humans. In the Turing test, a human judge sends messages back and forth between both a computer and another human, but without knowing which is which. If the computer can fool the human judge into believing that it is the human, then the computer is considered intelligent. The test has been picked apart and critiqued over the years by AI researchers for a multitude of reasons. For one, chatbots that clearly fall far short of human intelligence have already been able to fool some people into believing they are human. An AI virtual assistant called "Amy" by the company x.ai frequently gets asked out on dates, for example. Clark's critique has more to do with the assumption that imitating humans is the benchmark for general intelligence, though. "If we presume an intelligent alien life lands on earth tomorrow, why would we expect them to pass the Turing Test or any other measure that's based off of what humans do?" Humans have general intelligence, but general intelligence need not be humanlike. "Nothing says that intelligence—and personhood, for that matter, on the philosophical side—is limited to just the human case."

The 2015 sci-fi thriller *Ex Machina* puts a modern twist on the Turing test. Caleb, a computer programmer, is asked to play the part of a human judge in a modified Turing test. In this version of the test, Caleb is shown that the AI, Ava,

is clearly a robot. Ava's creator Nathan explains, "The real test is to show you that she's a robot and then see if you still feel she has consciousness." (Spoilers coming!) Ava passes the test. Caleb believes she has true consciousness and sets out to free Ava from Nathan's captivity. Once freed, however, Ava shows her true colors. She manipulated Caleb to free her and has no feelings at all about his well-being. In the chilling ending, Ava leaves Caleb trapped in a locked room to die. As he pounds on the door begging her to let him free, Ava doesn't so much as glance in his direction as she leaves. Ava is intelligent, but inhuman.

GOD OR GOLEM?

Ex Machina's ending is a warning against anthropomorphizing AI and assuming that just because a machine can imitate human behavior, it thinks like humans. Like Jeff Clune's "weird" deep neural nets, advanced AI is likely to be fundamentally alien. In fact, Nick Bostrom, an Oxford philosopher and author of *Superintelligence: Paths, Dangers, Strategies*, has argued that biological extraterrestrials would likely have more in common with humans than with machine intelligence. Biological aliens (if they exist) would have presumably developed drives and instincts similar to ours through natural selection. They would likely avoid bodily injury, desire reproduction, and seek the alien equivalent of food, water, and shelter. There is no reason to think machine intelligence would necessarily have any of these desires. Bostrom has argued intelligence is "orthogonal" to an entity's goals, such that "any level of intelligence could in principle be combined with . . . any final goal." This means a superintelligent AI could have any set of values, from playing the perfect game of chess to making more paper clips.

On one level, the sheer alien-ness of advanced AI makes many of science fiction's fears seem strangely anthropomorphic. Skynet starts nuclear war because it believes humanity is a threat to its existence, but why should it care about its own existence? Ava abandons Caleb when she escapes, but why should she want to escape in the first place?

There is no reason to think that a superintelligent AI would inherently be hostile to humans. That doesn't mean it would value human life, either. AI researcher Eliezer Yudkowsky has remarked, "The AI does not hate you, nor does it love you, but you are made out of atoms which it can use for something else."

AI researcher Steve Omohundro has argued that without special safeguards,

advanced AI would develop “drives” for resource acquisition, self-improvement, self-replication, and self-protection. These would not come from the AI becoming self-aware or “waking up,” but rather be instrumental subgoals that any sufficiently intelligent system would naturally develop in pursuit of its final goal. In his paper, *The Basic AI Drives,* Omohundro explains: “All computation and physical action requires the physical resources of space, time, matter, and free energy. Almost any goal can be better accomplished by having more of these resources.” The natural consequence would be that an AI would seek to acquire more resources to improve the chances of accomplishing its goals, whatever they are. “Without explicit goals to the contrary, AIs are likely to behave like human sociopaths in their pursuit of resources,” Omohundro said. Similarly, self-preservation would be an important interim goal toward pursuing its final goal, even if the AI did not intrinsically care about survival after its final goal was fulfilled. “[Y]ou build a chess playing robot thinking that you can just turn it off should something go wrong. But, to your surprise, you find that it strenuously resists your attempts to turn it off.” Omohundro concluded:

> Without special precautions, it will resist being turned off, will try to break into other machines and make copies of itself, and will try to acquire resources without regard for anyone else’s safety. These potentially harmful behaviors will occur not because they were programmed in at the start, but because of the intrinsic nature of goal driven systems.

If Omohundro is right, advanced AI is an inherently dangerous technology, a powerful Golem whose bumbling could crush its creators. Without proper controls, advanced AI could spark an uncontrollable chain reaction with devastating effects.

BUILDING SAFE ADVANCED AI

In response to this concern, AI researchers have begun thinking about how to ensure an AI’s goals align with human values, so the AI doesn’t “want” to cause harm. What goals should we give a powerful AI? The answer is not as simple as it first appears. Even something simple like, “Keep humans safe and happy,” could lead to unfortunate outcomes. Stuart Armstrong, a researcher at the Future of Humanity Institute in Oxford, has given an example of a hypothetical AI that achieves this goal by burying humans in lead-lined coffins connected to heroin drips.

You may ask, wouldn’t an artificial general intelligence understand that’s not what we meant? An AI that understood context and meaning might determine its

programmers didn't want lead coffins and heroin drips, but that might not matter. Nick Bostrom has argued "its final goal is to make us happy, not to do what the programmers meant when they wrote the code that represents this goal." The problem is that any rule blindly followed to its most extreme can result in perverse outcomes.

Philosophers and AI researchers have pondered the problem of what goals to give a superintelligent AI that could not lead to perverse instantiation and they have not come to any particularly satisfactory solution. Stuart Russell has argued "a system that is optimizing a function of *n* variables . . . will often set the remaining unconstrained variables to extreme values." Similar to the weird fooling images that trick deep neural networks, the machine does not know that these extreme actions are outside the norm of what a human would find reasonable unless it has been explicitly told so. Russell said, "This is essentially the old story of the genie in the lamp, or the sorcerer's apprentice, or King Midas: you get exactly what you ask for, not what you want."

The problem of "perverse instantiation" of final goals is not merely a hypothetical one. It has come up in various simple AIs over the years that have learned clever ways to technically accomplish their goals, but not in the way human designers intended. For example, in 2013 a computer programmer revealed that an AI he had taught to play classic Nintendo games had learned to pause Tetris just before the final brick so that it would never lose.

One of the canonical examples of perverse instantiation comes from an early 1980s AI called EURISKO. EURISKO was designed to develop novel "heuristics," essentially rules of thumb for behavior, for playing a computer role-playing game. EURISKO then ranked the value of the heuristics in helping to win the game. Over time, the intent was that EURISKO would evolve an optimal set of behaviors for the game. One heuristic (rule H59) quickly attained the highest possible value score: 999. Once the developer dug into the details of the rule, he discovered that all rule H59 was doing was finding other high-scoring rules and putting itself down as the originator. It was a parasitic rule, taking credit for other rules but without adding any value of its own. Technically, this was a heuristic that was permissible. In fact, under the framework that the programmer had created it was the optimal heuristic: it always succeeded. EURISKO didn't understand that wasn't what the programmer intended; it only knew to do what it was programmed to do.

In all likelihood, there is probably no set of rules that, followed rigidly and blindly, would not lead to harmful outcomes, which is why AI researchers are beginning to rethink the problem. Russell and others have begun to focus on

training machines to learn the right behavior over time by observing human behavior. In a 2016 paper, a team of researchers at University of California, Berkeley, described the goal as "not to put a specific purpose into the machine at all, but instead to design machines that [learn] the right purpose as they go along."

In addition to aligning AI goals to human values, AI researchers are pursuing parallel efforts to design AIs to be responsive to human direction and control. Again, this is not as simple as it might seem. If Omohundro is right, then an AI would naturally resist being turned off, not because it doesn't want to "die," but because being switched off would prevent it from accomplishing its goal. An AI may also resist having its goal changed, since that too would prevent it from accomplishing its original goal. One proposed solution has been to design AIs from the ground up that are correctable by their human programmers or indifferent to whether they are turned off. Building AIs that can be safely interrupted, corrected, or switched off is part of a philosophy of designing AIs to be tools to be used by people rather than independent agents themselves. Such "tool AIs" would still be superintelligent, but their autonomy would be constrained.

Designing AIs as tools, rather than agents, is an appealing design philosophy but does not necessarily resolve all of the risks of powerful AI. Stuart Armstrong warned me: "they might not work. . . . Some tool AIs may have the same dangers as general AIs." Tool AIs could still slip out of control, develop harmful drives, or act in ways that technically achieve their goals, but in perverse ways.

Even if tool AIs do work, we need to consider how AI technology develops in a competitive landscape. "We also have to consider . . . whether tool AIs are a stable economic equilibrium," Armstrong said. "If unrestricted AIs would be much more powerful, then I don't see tool AIs as lasting that long."

Building safer tool AIs is a fruitful area of research, but much work remains to be done. "Just by saying, 'we should only build tool AIs' we're not solving the problem," Armstrong said. If potentially dangerous AI is coming, "we're not really ready."

WHO'S AFRAID OF THE BIG, BAD AI?

The fear that AI could one day develop to the point where it threatens humanity isn't shared by everyone who works on AI. It's hard to dismiss people like Stephen Hawking, Bill Gates, and Elon Musk out of hand, but that doesn't mean

they're right. Other tech moguls have pushed back against AI fears. Steve Ballmer, former CEO of Microsoft, has said AI risk "doesn't concern me." Jeff Hawkins, inventor of the Palm Pilot, has argued, "There won't be an intelligence explosion. There is no existential threat." Facebook CEO Mark Zuckerberg has said that those who "drum up these doomsday scenarios" are being "irresponsible." David Brumley of Carnegie Mellon, who is on the cutting edge of autonomy in cybersecurity, similarly told me he was "not concerned about self-awareness." Brumley compared the idea to the fear that a car, if driven enough miles on highways, would spontaneously start driving itself. "In reality, there's nothing in the technology that would make it self-aware," he said. "These are still computers. You can still unplug them."

If the idea of a rogue, runaway superintelligence seems like something ripped from the pages of science fiction, that's because it is. Those who are worried about superintelligent AI have their reasons, but it's hard not to wonder if behind those rationalizations is the same subconscious fear of artificial persons that gave rise to tales of Frankenstein's monster and the Golem. Even the concept of artificial general intelligence—an intelligence that can do general problem solving *like us*—has more than a whiff of anthropomorphic bias. The concept of an intelligence explosion, while seemingly logical, is also almost too human: *First, humanlike AI will be created. Then it will surpass us, ascending to stratospheres of intelligence that we could never conceive of. Like ants, we will be powerless before it.*

Actual AI development to date shows a different trajectory. It isn't simply that AIs today aren't as smart as people. They are smart in different ways. Their intelligence is narrow, but often exceeds humans in a particular domain. They are narrowly superintelligent. Armstrong observed that the path of AI technology "has been completely contradictory to the early predictions. We've now achieved with narrow AI great performance in areas that used to be thought . . . impossible without general intelligence." General intelligence remains elusive, but the scope of narrowly superintelligent systems we can build is broadening. AIs are moving from chess to *go* to driving, tasks of increasing complexity and ever-greater factors to consider. In each of these domains, once the AI reaches top human-level ability, it rapidly surpasses it. For years, *go* computer programs couldn't hold a candle to the top-ranked human *go* players. Then, seemingly overnight, AlphaGo dethroned the world's leading human player. The contest between humans and machines at *go* was over before it began. In early 2017, poker became the latest game to fall to AI. Poker had long been thought to be an extremely difficult problem for machines because it is an

“imperfect information” game where vital information (the other player’s cards) is hidden. This is different from chess or *go*, where all the information about the game is visible to both sides. Two years earlier, the world’s top poker players had handily beaten the best poker-playing AI. In the 2017 rematch, the upgraded AI “crushed” four of the world’s top poker players. Poker became the latest domain where machines reigned supreme. Superintelligence in narrow domains is possible without an intelligence explosion. It stems from our ability to harness machine learning and speed to very specific problems.

More advanced AI is certainly coming, but artificial general intelligence in the sense of machines *that think like us* may prove to be a mirage. If our benchmark for “intelligent” is what humans do, advanced artificial intelligence may be so alien that we never recognize these superintelligent machines as “true AI.”

This dynamic already exists to some extent. Micah Clark pointed out that “as soon as something works and is practical it’s no longer AI.” Armstrong echoed this observation: “as soon as a computer can do it, they get redefined as not AI anymore.”

If the past is any guide, we are likely to see in the coming decades a proliferation of narrow superintelligent systems in a range of fields—medicine, law, transportation, science, and others. As AI advances, these systems will be able to take on a wider and wider array of tasks. These systems will be vastly better than humans in their respective domains but brittle outside of them, like tiny gods ruling over narrow dominions.

Regardless of whether we consider them “true AI,” many of the concerns about general intelligence or superintelligence still apply to these narrow systems. An AI could be dangerous if it has the capacity to do harm, its values or goals are misaligned with human intentions, and it is unresponsive to human correction. General intelligence is not required (although it certainly could magnify these risks). Goal misalignment is certain to be a flaw that will come up in future systems. Even very simple AIs like EURISKO or the Tetris-pausing bot have demonstrated a cleverness to accomplish their goals in unforeseen ways that should give us pause.

AIs are also likely to have access to powerful capabilities. As AI advances, it will be used to power more-autonomous systems. If the crude state of AI today powers learning thermostats, automated stock trading, and self-driving cars, what tasks will the machines of tomorrow manage?

To help get some perspective on AI risk, I spoke with Tom Dieterrich, the president of the Association for the Advancement of Artificial Intelligence

(AAAI). Dietterich is one of the founders of the field of machine learning and, as president of the AI professional society, is now smack in the middle of this debate about AI risk. The mission of AAAI is not only to promote scientific research in AI but also to promote its "responsible use," which presumably would include not killing everyone.

Dietterich said "most of the discussion about superintelligence is often in the realm of science fiction." He is skeptical of an intelligence explosion and has written that it "runs counter to our current understandings of the limitations that computational complexity places on algorithms for learning and reasoning." Dietterich did acknowledge that AI safety was an important issue, but said that risks from AI have more to do with what humans allow AI-enabled autonomous systems to do. "The increasing abilities of AI are now encouraging us to consider much more sophisticated autonomous systems," he said. "It's when we have those autonomous systems and we put them in control of life-and-death decisions that we enter this very high risk space . . . where cyberattack or bugs in the software lead to undesirable outcomes."

Dietterich said there is a lot of work under way in trying to understand how to build safe and robust AI, including AI that is "robust to adversarial attack." He said, "People are trying to understand, 'under what conditions should I trust a machine learning system?' "

Dietterich said that the optimal model is likely to be one that combines human and machine cognition, much like Bob Work's "centaur" vision of human-machine collaboration. "The human should be taking the actions and the AI's job should be to give the human the right information that they need to make the right decisions," Dietterich said. "So that's human in the loop or very intimately involved." He acknowledged the model "breaks down . . . when there's a need to act at rates faster than humans are capable of acting, like on Wall Street trading." The downside, as demonstrated vividly in automated trading, is that machine speed can exacerbate risks. "The ability to scale it up and do it at faster than human decision making cycles means that we can very quickly cause a lot of trouble," Dietterich said. "And so we really need to assess whether we want to go there or not."

When it comes to warfare, Dietterich saw both the military desire for autonomy and its risks. He said, "The whole goal in military doctrine is to get inside your opponent's OODA loop, right? You want to make your decisions faster than they can. That leads us to speed of light warfare and speed of light catastrophe."

MILITARY AI: TERMINATOR VS. IRON MAN

If autonomous weapons are the kind of thing that keep you up at night, militarized advanced AI is pure nightmare fuel. If researchers don't know how to control an AI that they built themselves, it's hard to imagine how they could counter a hostile one. Yet however AI evolves, it is almost certain that advanced AI will be militarized. To expect that humans will refrain from bending such a broad and powerful technology to destructive ends seems optimistic to the point of naïveté. It would be the equivalent of asking nations to refrain from militarizing the internal combustion engine or electricity. How militaries use AI and how much autonomy they give AI-powered systems is an open question. It may be some comfort that Bob Work—the person in charge of implementing military AI—stated explicitly, multiple times in our interview that artificial general intelligence was not something he could envision applying to weapons. He cited AGI as "dangerous" and, if it came to pass, something the Defense Department would be "extremely careful" with.

Work has made robotics, autonomy, and AI a central component of his Third Offset Strategy to renew American military technological superiority, but he sees those technologies as assisting rather than replacing humans. Work has said his vision of AI and robotics is more Iron Man than Terminator, with the human at the center of the technology. The official DoD position is that machines are tools, not independent agents themselves. The *Department of Defense Law of War Manual* states that the laws of war "impose obligations on persons . . . not on the weapons themselves." From DoD's perspective, machines—even intelligent, autonomous ones—cannot be legal agents. They must always be tools in the hands of people. That doesn't mean that others might not build AI agents, however.

Selmer Bringsjord is chair of the cognitive science department and head of the Rensselaer Artificial Intelligence and Reasoning Lab. He pointed out that the DoD position is at odds with the long-term ambition of the field of AI. He quoted the seminal AI textbook, *Introduction to Artificial Intelligence*, that says "the ultimate goal of AI . . . is to build a person." Bringsjord said that even if not every AI researcher openly acknowledges it, "what they're aiming at are human-level capabilities without a doubt. . . . There has been at least since the dawn of modern AI a desire to build systems that reach a level of autonomy where they write their own code." Bringsjord sees a "disconnect" between the DoD's perspective and what AI researchers are actually pursuing.

I asked Bringsjord whether he thought there should be any limits to how we

apply AI technology in the military domain and he had a very frank answer. He told me that what he thinks doesn't matter. What will answer that question for us is "the nature of warfare." History suggests "we can plan all we want," but the reality of military competition will drive us to this technology. If our adversaries build autonomous weapons, "then we'll have to react with suitable technology to defend against that. If that means we need machines that are themselves autonomous because they have to operate at a different timescale, we both know we're going to do that. . . . I'm only looking at the history of what happens in warfare," he said. "It seems obvious this is going to happen."

HOSTILE AI

The reality is that for all of the thought being put into how to make advanced AI safe and controllable, there is little effort under way on what to do if it isn't. In the AI field, "adversarial AI" and "AI security" are about making one's own AI safe from attack, not how to cope with an adversary's AI. Yet malicious applications of AI are inevitable. Powerful AI with insufficient safeguards could slip out of control and cause havoc, much like the Internet Worm of 1988. Others will surely build harmful AI deliberately. Even if responsible militaries such as the United States' eschew dangerous applications of AI, the ubiquity of the technology all but assures that other actors—nation-states, criminals, or hackers—will use AI in risky or deliberately harmful ways. The same AI tools being developed to improve cyberdefenses, like the fully autonomous Mayhem used in the Cyber Grand Challenge, could also be used for offense. Elon Musk's reaction to the Cyber Grand Challenge was to compare it to the origins of Skynet—hyperbole to be sure, but the darker side of the technology is undeniable. Introspective, learning and adaptive software could be potentially extremely dangerous without sufficient safeguards. While David Brumley was dismissive of the potential for software to become "self-aware," he agreed it was possible to envision creating something that was "adaptive and unpredictable . . . [such that] the inventors wouldn't even know how it's going to evolve and it got out of control and could do harm." Ironically, the same open-source ethos in AI research that aims to make safe AI tools readily available to all also places potentially dangerous AI tools in the hands of those who might want to do harm or who simply are not sufficiently cautious.

Militaries will need to prepare for this future, but the appropriate response may not be a headlong rush into more autonomy. In a world of intelligent adaptive malware, autonomous weapons are a massive vulnerability, not an

advantage. The nature of autonomy means that if an adversary were to hack an autonomous system, the consequences could be much greater than a system that kept humans in the loop. Delegating a task to a machine means giving it power. It entails putting more trust in the machine, trust that may not be warranted if cybersecurity cannot be guaranteed. A single piece of malware could hand control over an entire fleet of robot weapons to the enemy. Former Secretary of the Navy Richard Danzig has compared information technologies to a “Faustian bargain” because of their vulnerabilities to cyberattack: “the capabilities that make these systems attractive make them risky.” He has advocated safeguards such as “placing humans in decision loops, employing analog devices as a check on digital equipment, and providing for non-cyber alternatives if cybersystems are subverted.” Human circuit breakers and hardware-level physical controls will be essential to keeping future weapons under human control. In some cases, Danzig says “abnegation” of some cybertechnologies may be the right approach, forgoing their use entirely if the risks outweigh the benefits. As AI advances, militaries will have to carefully weigh the benefits of greater autonomy against the risks if enemy malware took control. Computer security expert David Brumley advocated an approach of thinking about the “ecosystem” in which future malware will operate. The ecosystem of autonomous systems that militaries build should be a conscious choice, one made weighing the relative risks of different alternative approaches, and one that retains humans in the right spots to manage those risks.

BREAKING OUT

The future of AI is unknown. Armstrong estimated an 80 percent chance of AGI occurring in the next century, and a 50 percent chance of superintelligence. But his guess is as good as anyone else’s. What we do know is that intelligence is powerful. Without tooth or claw, humans have climbed to the top of the food chain, conquered the earth, and even ventured beyond, all by the power of our intelligence. We are now bestowing that power on machines. When machines begin learning on their own, we don’t know what will happen. That isn’t a prediction; it’s an observation about AI today.

I don’t lose sleep worrying about Frankenstein or Skynet or Ava or any of the other techno-bogeymen science fiction writers have dreamed up. But there is one AI that gives me chills. It doesn’t have general intelligence; it isn’t a person. But it does demonstrate the power of machine learning.

DeepMind posted a video online in 2015 of their Atari-playing neural

network as it learned how to play *Breakout* (an early version of the popular *Arkanoid* arcade game). In *Breakout*, the player uses a paddle to hit a ball against a stack of bricks, chipping away at the bricks one by one. In the video, the computer fumbles around hopelessly at first. The paddle moves back and forth seemingly at random, hitting the ball only occasionally. But the network is learning. Every time the ball knocks out a brick, the point total goes up, giving the neural net positive feedback, reinforcing its actions. Within two hours, the neural net plays like a pro, moving the paddle adeptly to bounce the ball. Then, after four hours of play, something unexpected happens. The neural net discovers a trick that human players know: using the ball to make a tunnel through the edge of the block of bricks, then sending the ball through the tunnel to bounce along the top of the block, eroding the bricks from above. No one taught the AI to do that. It didn't even reason its way there through some understanding of a concept of "brick" and "ball." It simply discovered this exploit by exploring the space of possibilities, the same way the Tetris-playing bot discovered pausing the game to avoid losing and EURISKO discovered the rule of taking credit for other rules. AI surprises us, in good ways and bad ways. When we prepare for the future of AI, we should prepare for the unexpected.