# Artificial Intelligence and Cyber Power from a Strategic Perspective

by Michael Mayer

**Advances in artificial intelligence (AI) will enable highly capable autonomous weapon systems to be used on the conventional battlefield and in cyberspace. This report provides an overview of the strategic aspects of AI and cyber power.**

## Take aways

- **Artificial intelligence can outperform humans at narrowly defined tasks and will enable a new generation of autonomous weapon systems.**
- **Cyberspace will play a crucial role in future conflicts due to the integration of digital infrastructure in society and the expected prevalence of autonomous systems on the battlefield.**
- **AI cyber weapons create a dangerous class of persistent threats that can actively and quickly adjust tactics as they relentlessly and independently probe and attack networks.**

Artificial intelligence (AI) has made incredible progress, resulting in highly capable software and advanced autonomous machines. Meanwhile, the cyber domain has become a battleground for access, influence, security and control. The strategic implications of these parallel developments will be profound, particularly when combined.

This report, divided into three articles, offers an introduction to both artificial intelligence and cyber power. The first article provides an overview of artificial intelligence, including the various approaches, current capabilities, and strategic applications of the technology. The second article covers strategic considerations relevant to cyber power, including computer network exploitation and offense capabilities.

Finally, the third article explores how the integration of these two technologies greatly enhances both intrusion detection and offensive network penetration capabilities, but complicates adequate command and control. I then look ahead to future strategic developments given current trends, including the potential for super-intelligence.

**Article One**

# Introduction to Artificial Intelligence

by Michael Mayer

**Military leaders are investing heavily in artificial intelligence technologies in the belief that AI and autonomy will be decisive in future conflict scenarios.**

Artificial intelligence (AI) is developing at an astounding rate. Driven by exponential growth in computing power, industrial and military robotic applications exhibit greater degrees of automation, using complex algorithms to perform increasingly complex operations and leveraging machine-learning techniques that allow computer systems to detect patterns and make predictions based on the data. Machines are becoming more capable of rationally solving complex problems in whatever real-world circumstances they encounter.

These developments have sweeping societal implications. Today's self-driving vehicles and smart phone personal assistants – like Apple's Siri – are simply the beginning of a new era of intelligent machines. These machines are able to analyze their surroundings and respond accordingly and even independently, whether in the workplace or on the battlefield. Military leaders in the U.S., China, and Russia are investing heavily in these technologies in the belief that AI and autonomy will be decisive in future conflict scenarios.

## ALPHA GO

On 20 March 2016, an AI program specifically designed to play the ancient Chinese game of *Go* soundly defeated the best human player of the game in a special tournament held in Seoul, South Korea. The resounding victory (four games to one) by the program, named AlphaGo, was immediately heralded as a significant milestone for AI technology due to the complexity of the game it was playing. Researchers assumed it would take at least another decade for a machine to beat a top-ranked human player. Instead, it only took a year.

The game's simplicity – using black and white stones on a 19 by 19 grid with the objective of covering as much of the board as possible – is part of what makes it so difficult, particularly compared with chess, another game famously mastered by an earlier form of machine intelligence. After the first two moves in chess, there are 400 possible next moves compared with nearly 130,000 possible options in Go.

The computer algorithms that powered IBM's *Deep Blue* to victory over Gary Kasparov in 1996 could use sheer computing power to analyze the value of each of a huge number of potential moves. This "brute force" approach could not be applied to Go due to the huge difference in possible variations. It is for this reason that Go has been called the "Holy Grail" of artificial intelligence.

Human players rely on a combination of strategy, experience, and – notably – intuition. Google's *DeepMind* artificial intelligence group, the team behind AlphaGo, relied on a special type of AI machine learning modeled after the human brain that is particularly good at recognizing patterns in data and has the ability to "teach itself" by endlessly playing matches with itself.

The machine was learning to play in a human-like fashion, only at a mindboggling pace that compared with how much experience a human might acquire after playing the game for 80 straight years. AlphaGo managed to surprise the reigning world champion, Lee Sedol, with a nearly flawless play and unexpected moves that Go experts even called

"beautiful". As journalist Christopher Moyer commented, if AlphaGo "can learn to conquer Go ... it can learn to conquer anything easier than Go – which amounts to a lot of things."

## DEFINING ARTIFICIAL INTELLIGENCE

In a 2010 book on the subject, Nils Nilsson defined artificial intelligence (AI) as "that activity devoted to making machines intelligent, and intelligence is that quality that enables an entity to function appropriately and with foresight in its environment." Given this definition, which is used approvingly by other experts in the field, one might think of machine intelligence on a spectrum that incorporates simple digital calculators and smart thermostats at the lower end, and software controlling self-driving vehicles or playing the game of Go at the higher end.

The decisive (and exciting) factor for AI is an ability to combine computation and information processing in a way that goes beyond simple pre-programmed functions or "number crunching." Many systems today use complex series of algorithms that allow computer systems to detect patterns in and make predictions based on data. These algorithms are linear, step-by-step sets of rules (often containing conditional "if-then" statements) used to accomplish specific tasks.

With recent improvements to AI, machines go beyond such structured algorithms and are increasingly capable of expanding their own knowledge base and range of responses. Thereby, they are able to rationally solve complex problems and achieve their goals in whatever real-world circumstances they encounter.

## APPROACHES TO AI

The field of artificial intelligence has progressed so rapidly and profoundly that it suffers from what some refer to as the "AI effect," whereby machines are able to perform a uniquely new task only to have it soon accepted as a normal machine function and therefore not "intelligent". The frontier of AI, according to one report by Stanford University, is constantly evolving in a continuous and incremental way.

The field emerged around the time of the first computational machines during the 1940s, although the term artificial intelligence was coined a decade later by John McCarthy during the 1956 British Dartmouth Summer Conference. Perhaps one of the most well-known tests for machine intelligence was suggested by Alan Turing in 1950: whether a human interrogator could determine if the subject was a computer based purely on its responses to questions.

As Andrew Ilachinski from the U.S. think tank CNA pointed out, the capabilities needed to pass the Turing test nicely summarize the primary areas of AI research:

> *Natural language processing*: automatic speech recognition and the ability to verbalize responses
>
> *Knowledge representation*: to organize the system's knowledge base
>
> *Automated reasoning* – the system has the ability to draw inferences from this knowledge and respond to queries
>
> *Machine learning* – the system incorporates new information into its knowledge base and finds patterns in the data

Users of the Apple iPhone or automated customer service platforms will recognize the progress made by artificial intelligence – but also its limitations. Whether it is an AI such as Siri or a customer service interactive voice response (IVR) "chatbot" programmed to assist in document retrieval or other common tasks, the speech recognition algorithms are impressive, yet frustratingly inadequate. Clearly, these systems will not yet fool a human, although advances in logic processing and understanding contextual clues will gradually bring machine intelligence closer to passing the Turing test.

## EXPERT SYSTEMS

From the beginning, there were two basic approaches to artificial intelligence.

The first was a top-down approach that connected a knowledge base to the rules of logical reasoning required for the specific domain within which it would be used. Basically, the computer would be programmed to reason just as a human would. For years, this was the basic means of automated language translation – including all the words and sets of grammatical rules from each language before entering a sentence from one language and receiving an often-times imperfect translated sentence as output.

This time-consuming approach to AI worked reasonably well in applications for which the environment was predictable and the rules were very clear, such as chess. There are a finite number of possible moves, the environment is predictable, and the rules are relatively simple. Even complex tasks such as analyzing certain laboratory test results to detect diseases were automated in the 1960s and performed better than junior doctors.

These types of AI have also been referred to as *expert systems* because a subject matter expert – whether a grand master in chess or a medical doctor – must first provide all the necessary rules or guidelines for the AI and its "choices" are based on these predetermined rules and logics. How well it performs becomes a matter of processor speeds and internal memory.

## MACHINE LEARNING

The second approach was a bottom-up methodology based on the evolutionary tendency in nature to build more complex systems from smaller simpler components. In this way, machines could learn in a manner similar to humans, through data collection and processing. Rather than a rigid, rules-based processing of expert systems, machine intelligence could be flexible and adaptive like human intelligence.

This bottom-up approach is referred to as "machine learning" and includes various techniques that, as Ilachinski described, "detect patterns in, and learn and make predictions from data."

One older bottom-up machine-learning technique that copies human biology has experienced a renaissance to become one of the leading AI approaches. Researchers in the 1940s decided, naturally enough, that the human brain was itself a useful model on which to base machine intelligence. The brain, as Gideon Lewis-Kraus wrote in the *New York Times*:

> Is just a bunch of widgets, called neurons, that either pass along an electrical charge to their neighbors or don't. What's important are less the individual neurons themselves than the manifold connections among them.... There was no reason you couldn't try to mimic this structure in electrical form, and in 1943 it was shown that arrangements of simple artificial neurons could carry out basic logical functions. They could also, at least in theory, learn the way we do.

## DEEP LEARNING

These neural networks are one type of machine learning and, in a modernized and modified form called deep learning, is the technology behind AlphaGo and other advanced AI systems. Neural networks are learning models using layers of "neurons" to make predictions regarding an expected output based on identifying patterns in the input data.

The neurons in the layers are assigned a weighted value and, if the value passes a particular threshold, it activates or "fires" in a manner similar to the brain. As Lewis-Kraus noted, "with one layer, you could find only simple patterns; with more than one, you could look for patterns of patterns." Neural networks employ layer upon layer of neurons, sifting through the input data and fine-tuning the predictions that match that input with the desired output.

To "learn" a particular function; neural networks use training sets consisting of input-output pairings (giving, in other words,

the correct answer or appropriate output for the input) and assign the weighted values to induce the neurons in each layer to fire when part of the pattern in the data is recognized.

## BACKPROPAGATION

At the beginning, these weighted values between neurons may be random and, as James Somers described it, "it's as if the synapses of the brain haven't been tuned yet." If the results are not entirely correct, the network can then move backwards among the layers preceding the final output layer in a method known as backpropagation or "backprop," adjusting the weighted values.

In this way, for example, a hand-written postal code can be scanned and each number analyzed by the neural network, with each layer of the network essentially "voting" or activating as the pixels making up the portions of each number image are identified with the actual number. The more the system practices with training sets, the more accurate and efficient the network becomes at finding the appropriate patterns among the data.

Incredibly, this system of recognizing patterns in data was conceived in the 1940s and revisited again in the late 1980s with success at limited tasks such as playing backgammon. When this breakthrough could not be replicated in other games such as chess or Go, however, neural network research fell dormant for over a decade. The advent of modern computer processing and the reintegration of other types of learning algorithms such as backpropagation enabled the current explosion in artificial intelligence applications and successes such as AlphaGo.

## ALPHAGO AND REINFORCEMENT LEARNING

Using a deep learning neural network, the DeepMind team fed a large training set of Go matches into the system, but also enabled AlphaGo to play thousands of simulated matches with itself, a process called reinforcement learning. In this way, the software constantly improved its play. The AI developed the ability to identify patterns in large data sets but also began to establish the foundation for machine decision-making.

Machine decision-making based on deep learning neural networks presents a new challenge. The methods by which the network learns, adjusting and tweaking the predictive values in each layer, is something of a mysterious "black box" to humans. As MIT professor Tommi Jaakkola observed, "If you had a very small neural network, you might be able to understand it … but once it becomes very large, and it has thousands of units per layer and maybe hundreds of layers, then it becomes quite un-understandable."

A neural network left running overnight can "teach" itself French, but the engineers designing the applications cannot actually explain how this occurs. Other AI systems have been successfully used to identify patterns in patient journals to predict diseases but cannot give any rationale for how it works. It just does.

Chris Nicholson, founder of a deep learning start-up venture, explained that "people understand the linear algebra behind deep learning. But the models it produces are less human-readable. They're machine readable.... They can retrieve very accurate results, but we can't always explain, on an individual basis, what led them to those accurate results."

This may not be problematic for facial recognition or apps that suggest new movies or music selections. However, AI-assisted decision-making has become more common. To trust or verify the AI-generated conclusion or recommendation, it may be necessary to understand how that recommendation was formulated. This is particularly true for systems that can teach themselves without human supervision. Researchers have also been experimenting with AI machine-learning software that can code new AI software – AI creating new AI.

## THE AI EVOLUTION

The technical evolution of AlphaGo is illustrative. After its success in South Korea, the *DeepMind* team improved and simplified the

software architecture, eventually launching a more powerful yet more efficient version called AlphaGo Zero that required a smaller computer system. Rather than utilizing training sets of human matches, AlphaGo Zero was given the rules and learned to play on its own by randomly placing the pieces on the board.

The progress was staggering. After quickly advancing to the level of an amateur, AlphaGo Zero discovered certain tactics commonly employed by human players after the first day. It rose to a level comparable to a grand master after only three days, discovering new moves along the way that humans had not invented despite the game's two-millennia-long history.

This ability to independently generate new knowledge without requiring access to previous human expertise is groundbreaking. AlphaGo Zero has a perfect record (100-0) against the older AlphaGo version that so resoundingly defeated Lee Sedol. As lead researcher David Silver remarked, the system is more powerful because "by not using human data, or human expertise in any fashion, we've removed the constraints of human knowledge and it is able to create knowledge itself."

## COMPUTING POWER

Advances in AI have come about due to new methods and software architectures, but also because the pure computational power needed for machine-learning techniques such as neural networks simply was not available. One innovator of neural networks, Geoffrey Hinton, recalled that "there just wasn't enough computer power or enough data. People on our side kept saying 'Yeah, but if I had a really big one, it would work.' It wasn't a very persuasive argument."

Eventually, however, the computers caught up with the demands of the technique. The evolution in computing power has accelerated dramatically. For nearly forty years (starting in the mid-1970s) the progression of computer processing power followed the prediction of Gordon Moore that the number of transistors on a microchip – and therefore computer performance – roughly doubles every two years. Known as Moore's Law, the pattern held true until about 2012, when microchip miniaturization began to approach the physical limits of nanotechnology.

The rapid improvements in computing power made it feasible to develop machine learning through neural networks and backprop. Computer processing (measured in computations per second) improved so dramatically that machine computational ability appears to be trending toward a level comparable with the human brain within the decade.

Inventor and Google research scientist Ray Kurtzweil has observed that technological progress is evolutionary and builds upon previous advances. Kurtzweil argues that when barriers to technological advances emerge, new technologies will be developed to find ways around those barriers.

## COMPUTER CHIP EVOLUTION

The apparent end of Moore's law provides an apt illustration. Alongside the development of new AI techniques, another revolution is occurring in microchip processors. Despite more efficient AI architectures, machine-learning approaches such as neural networks normally have used graphics processing units (GPUs) originally intended for graphics-heavy computer gaming. These require substantial computing power and therefore large amounts of energy.

Several chip manufactures are experimenting with new "neuromorphic" computer chip that mirrors how the brain functions while using far less energy, a characteristic that will be particularly useful for AI applications.

One of the fastest supercomputers, IBM's Sequoia, consumes 7.9 megawatts of energy, whereas the human brain needs only 20 watts. Using a network of 130,000 artificial neurons, the new Intel Loihi chip sends data by generating pulses of energy between the neurons, only using energy when those neurons are activated in the same way the human brain does. The chip is self-learning and

could be useful for AI applications, but with much lower energy demands.

The U.S. Department of Defense is funding a similar concept at IBM. The TrueNorth chip reportedly is particularly adept at parallel processing (running multiple applications simultaneously) and finding patterns in data in a way similar to human cognition.

Google, which developed a similar type of chip several years ago, announced in February 2018 that it will allow other companies access to their AI chip – which it calls tensor processing units (TPUs) – via a cloud computing solution.

One potential application will be self-driving vehicles. The current system under development requires days of training to enable the software to identify street signs or pedestrians. The new chips may reduce training time to mere hours. Convinced that driver assistance applications and self-driving vehicles are the future of the automotive business, manufacturers such as BMW and Volkswagen are also expanding into computer technology and chip production.

## QUANTUM COMPUTING

Even more ambitious are efforts to harness the promise of quantum computing, technology based on quantum physics or the study of how atomic and subatomic particles behave.

Conventional computing is based on binary digits (or "bits") which have a value of either zero or one. The value of these bits is transferred through the computer's network using electrical impulses and light flashes, with network speeds usually measured in bits per second (or, given today's high-speed data connections, in million (mega-) bits per second, Mbps). Bits themselves can carry logical information such as "on" or "off," "true" or "false," but more data can be held in a sequence of eight bits, which is called a byte.

In quantum computing, the basic unit is called a qubit and – rather than the ones and zeros of bits – can exist in multiple states simultaneously, allowing qubits to carry much more information. This in turn enables exponentially faster computation speeds.

In 2017, Volkswagen used a quantum computer from the Canadian manufacturer D-Wave to demonstrate how it optimized the movements of 10,000 taxis in Beijing simultaneously to avoid congestion and improve traffic flow. What would have taken a supercomputer 30 minutes was instead completed within a few seconds.

Although the technology remains controversial and its potential unclear, the possible applications for quantum computing are profound and would have immediate and fundamental implications, not just for artificial intelligence but also fields such as cryptography.

## NARROW VERSUS GENERAL ARTIFICIAL INTELLIGENCE

Most applications of artificial intelligence, however impressive, are still relatively narrow in focus – identifying patterns or faces from large data sets, playing games such as chess or Go, operating a vehicle based on the rules of the road.

As Andrew Ilachinski noted, "narrow AI" successes have two main characteristics. First, they "map fairly simple inputs to outputs": an image recognition program receives an image as input and labels it a dog as output, one language is entered into the translator and another emerges. Second, writes Ilachinski, "the time scales for human performance (on the same set of specific problems) are fairly short," meaning that the time needed to gather the information necessary to make a decision – whether it be a chess move or a driving maneuver – can usually be measured in seconds.

Adapting AI from narrowly defined tasks to be useful in other contexts and across a broad range of input factors – in other words, moving from narrow AI to what is called artificial general intelligence or AGI – is still a long way off. Nevertheless, this is partially the reason for the excitement around AlphaGo (not to mention AlphaGo Zero) – its ability to separate the learning and decision-

making from the data set. It's a system that taught itself to play Go, but the software will be able to teach itself many other applications as well, albeit still within a narrow and predictable context.

## DEDUCTIVE, INDUCTIVE AND ABDUCTIVE REASONING

Building up a knowledge base and developing either consciousness or what we might call "common sense" remain a formidable challenge. One piece of the AGI puzzle is IBM's Watson supercomputer and DeepQA, the AI architecture underpinning it. Using millions of documents downloaded into its memory, DeepQA crossed a significant threshold in 2011 when it was able to understand the verbal questions posed on the trivia game show *Jeopardy*. Watson reasoned its way to the answers using its information database, defeating the two best human players.

Despite the impressive advances in machine-learning techniques, general artificial intelligence of the sort depicted in science fiction films is not yet visible and may not even be possible, let alone desirable.

Humans excel at adaptability in the face of unanticipated environmental stimuli and are usually able to quickly process and analyze new and unexpected information. Whereas computers are already superior in pure computational ability or *deductive* reasoning (applying general rules of logic to a set of data to reach correct conclusions about that data), the opposite (using individual observations to reach general principles or *inductive* reasoning) is much more difficult for AI, not to mention reaching explanations based on limited data points (or *abductive* reasoning). These skills separate a system that can simply recognize facts and situations from one that can actively apply reason to unforeseen situations.

As Paul Allen, co-founder of Microsoft, observed, "our systems have always remained 'brittle' – their performance boundaries are rigidly set by their internal assumptions and defining algorithms, they cannot generalize, and they frequently give nonsensical answers outside of the specific focus areas."

AI researchers note that computers are becoming more competent than humans at advanced computational functions but have yet to achieve the "common sense" of a child. It is therefore not surprising that DARPA sought funding in 2018 for research into programs that "create more human-like knowledge representations … to enable commonsense reasoning by machines about the physical world."

In February 2018, Allen announced a personal $125 million donation to develop AI "common sense," in part by compiling a database of fundamental human knowledge computers lack. Until machines achieve superintelligence, however, there are many applications for which narrow AI is more than adequate.

## STRATEGIC USES FOR ARTIFICIAL INTELLIGENCE

Even though the driving force behind AI development is mainly in the civilian sector, military applications based on complex algorithms for data analysis and pattern recognition are already widespread. The advantages of quickly analyzing large amounts of video or still images for intelligence purposes are obvious, but the possibilities go well beyond even these valuable tools.

In this age of informational warfare, AI enables video and audio forgeries. Augmented decision-making for combat systems has existed for decades, but AI is expected to further enhance the ability of machines to provide command and control support to military leaders on an increasingly complex and rapidly changing battlefield and do so in ways that will likely be superior to humans. And machines will increasingly be able to effectively control the weapons systems themselves, either individual platforms such as armed unmanned combat drones or swarms of such platforms. Even as AI investments expand, some practical applications are already visible.

## PREDICTIVE POLICING

In some American cities, police are using AI to predict criminal activity based on data analysis in a technique known as predictive policing. Although police departments have mapped and analyzed crime patterns for decades using simpler methods such as pushpins on a wall map, they lacked the ability to react while the trends were unfolding. Predictive algorithms analyze the data and predict geographical areas of particular concern down to within a single city block, giving police the ability to expand their presence in those areas.

Similar analyses are being adapted for counterinsurgency – understanding and predicting future behavior based on patterns in the data. Researcher Paulo Shakarian has developed precisely this type of tool, which looked for patterns in the behavior of ISIS insurgents. As Shakarian related, "What we wanted to look for was: Are there relationships amongst the actions the Islamic State does that leads to significant increases in activity?… When the violence increases that much, we want to understand why that is. We wanted to get insight into what led them to conduct certain military tactics."

One of the largest commercial actors in the predictive policing market is the U.S. technology company Palantir, founded in 2004 by a group of investors including billionaire Peter Thiel, and nurturing close ties to defense and intelligence agencies. Using complex algorithms and "big data," Palantir sells analytic software meant to provide real-time analyses of a wide-ranging database including information. Its customers include some of the largest police departments in the U.S., although the partnership has not always gone smoothly.

In 2017, the Norwegian Customs Directorate signed a 300 million kroner contract with Palantir to provide AI-based data analysis that combines information from other databases including NAV (the Norwegian Labor and Welfare Administration), government property registers and information from open internet sources. For instance, a car's registration plate can be automatically photographed at the border and information about its owner is then cross-referenced with data from the Customs Directorate and certain external databases. The data is then analyzed by AI algorithms to identify and predict patterns of potential criminal activity.

## IMAGE RECOGNITION FOR ISR

Advances in image recognition AI algorithms have expedited intelligence, surveillance, and reconnaissance (ISR) imagery, particularly from unmanned aerial vehicles. With the advent of persistent overhead ISR came an overwhelming amount of data to send back for analysts to sift though. As one top military leader commented, "today an analyst sits there and stares at Death TV for hours on end, trying to find the single target or see something move. It's just a waste of manpower."

Onboard algorithms are already being used to conduct a preliminary filtering to reduce the terabytes of data transmitted back for analysis, reducing bandwidth requirements. On the ground, the U.S. military employs AI neural networks to assist in video analysis. Not only can the software process far greater amounts of data in far less time than human analysts, it is becoming more effective as well. In a 2015 competition, machine-learning software developed by Microsoft and Google outperformed humans at image recognition.

In 2017, the Air Force took the next logical step and created the Algorithmic Warfare Cross Functional Team – also known as Project Maven – focused on using AI to accelerate image analysis. Combined with geospatial software, targets can be identified and tracked over time. With improvements to machine learning, the application is scalable – starting with smaller images from tactical drones and adapting the technology to larger sensors such as an MQ-9 Reaper drone with a Gorgon Stare sensor able to provide coverage of an entire city. Eventually, processing and analysis of satellite feeds for daily global image analysis will be possible.

The power of AI to assist ground forces in tactical reconnaissance, particularly in

counterterrorism or counterinsurgency operations, is already being demonstrated. Chinese officials recently revealed a new system that connects police on the street wearing camera-equipped sunglasses with facial recognition software and a centralized criminal database, giving them almost instant access to an individual's personal information. Airport immigration and customs officials in Europe and elsewhere are employing biometric facial recognition for passenger identification.

## INCREASED ABILITY TO FALSIFY

The advent of reliable image recognition technology has also given rise to software to reverse engineer those images. As Greg Allen has argued, "in our society, audio and video recordings serve as the final arbiter of truth." One Canadian startup company has developed AI-driven technology that can produce audio mimicking anyone's voice – the company's demo uses Donald Trump, Barack Obama, and Hillary Clinton – with surprising realism. Software maker Adobe has announced a similar effort heralded as "Photoshop for audio."

The ability to falsify extends to video as well. Stanford researchers used AI-based software to change – in real time – the facial expressions of individuals in YouTube videos. Even more startling is an ability to run image recognition software in reverse, creating synthetic images based solely on a text description.

One researcher involved in the effort, Jeff Clune, revealed that "people send me real images and I start to wonder if they look fake. And when they send me fake images I assume they're real because the quality is so good."

In an age of information warfare, the ability to convincingly falsify audio and video could be a powerful weapon to retain plausible deniability or generate false claims to justify a military intervention. Allen warned in an online piece aptly titled "AI will make forging anything entirely too easy" that unless this challenge is met, "we will have to live in a society where there is no ultimate arbiter of truth."

## MACHINE LEARNING FOR WEAPONS SYSTEMS

Not only is artificial intelligence creating new capabilities through pattern recognition and image manipulation, machine-learning techniques are improving the decision-making ability of existing automated weapons systems and enabling new possibilities for the control of military platforms.

The Aegis Combat System first installed aboard U.S. Navy ships in 1984 is an integrated command and control system able to independently identify, track, and engage targets. It has four settings with varying degrees of human control, ranging from "semi-automatic" in which the ship's personnel use the system to assist in target prioritization) to a so-called "casualty mode" setting where it is assumed that the crew can no longer make any command decisions and the ship autonomously identifies, tracks, and engages targets.

With the advent of machine learning, this already extremely capable decision-making software might be vastly improved – constantly evaluating its own performance and finding unique tactical solutions much in the same way that AlphaGo Zero discovered new tactics for the board game.

## AUTONOMOUS PLATFORMS

For individual platforms, AI will enable autonomous modes that extends beyond simply operation of the vehicle itself. Similar to Tesla's autopilot or Google's self-driving cars, the U.S. military has already experimented with autonomous vehicle convoys for supplying troops in dangerous terrain without putting human drivers at risk.

Onboard computer systems have assisted pilots in flying aircraft since the advent of "fly-by-wire" technology decades ago. Among the innovations of the new F-35 fighter aircraft is the onboard integrated sensor and weapons system software that collects and processes large amounts of data and displays

it on the pilot's helmet. In addition to managing onboard flight systems, the aircraft's software can independently identify and track multiple targets, allowing the pilot to focus on tactical decision-making.

Given the advances in processing speed and lower power requirements of newer AI computer chips, it seems likely that the next generation fighter aircraft will not only be capable of operating the aircraft and analyzing relevant sensor information, but also of performing the tactical decision-making.

In 2016, an AI application created by a doctoral student at the University of Cincinnati soundly defeated retired Air Force Colonel Gene Lee, an experienced fighter pilot with significant operative and simulator experience. The AI divided its larger tasks into smaller ones such as target tracking, firing weapons, or defensive maneuvers. In this way, it continually focused on only the most relevant tasks, which sped up tactical decision-making and reduced computational requirements.

Using an inexpensive Raspberry Pi computer, the efficient AI software consistently found the best tactical solutions. Drained after hours-long sessions against the first AI to regularly beat a human pilot in a simulator, Lee commented that "I was surprised at how aware and reactive it was.... It seemed to be aware of my intentions and reacting instantly to my changes in flight and missile deployment. It knew how to defeat the shot I was taking. It moved instantly between defensive and offensive actions as needed."

## AUTONOMOUS SWARMS

Control over individual platforms in complex environments will soon extend to control over groups or "swarms" of individual platforms functioning as a network, particularly in the air or at sea. These systems need not carry weapons to have a significant strategic impact.

As Edward Moore Geist points out: Most nuclear powers base the security of their deterrent on the assumption that missile-carrying submarines will remain difficult for enemies to locate, but relatively inexpensive AI-controlled undersea drones may make the seas "transparent" in the not-too-distant future. The geostrategic consequences of such a development are unpredictable and could be catastrophic.

Successful tests involving large numbers of smaller drones – both in the air and at sea – suggest that future battlefields will have AI-controlled autonomous swarms monitored by human commanders, but with individual unit-level control coordinated by machines. Clearly, this will have dramatic implications for command and control decision loops. As one officer commented to author Peter Singer, "the trend towards the future will be robots reacting to robot attack, especially when acting at technologic speed ... as the loop gets shorter and shorter, there won't be any time in it for humans."

## GREAT POWER AI ARMS RACE

If these developments are any indication of the future strategic environment, AI will play a pivotal role as an enabling capability. As the authors of a 2017 report speculated, "many actors will face increasing temptation to delegate greater levels of authority to a machine, or else face defeat," noting that Russian authorities have "approved an aggressive plan that would have 30% of Russian combat power consist of entirely remote-controlled and autonomous robotic platforms by 2030."

In the wake of AlphaGo's success in 2016, South Korea announced that it would invest nearly US$1 billion over a five-year period on civilian public-private partnerships for AI research and development. The success of AlphaGo reportedly also convinced Chinese military leaders of the capacity of AI to think and act strategically. China appears ready to apply AI and autonomy not only to individual weapon systems but also military command and control decision-making.

This AI focus is matched by the Pentagon's "Third Offset Strategy," an initiative aimed at recapturing the U.S. technological advantage on the battlefield. As Deputy Secretary of Defense Robert Work explained in a 2016 speech, the U.S. has since the 1950s sought "ways in which to offset our potential ad-

versary's advantages." To offset the Soviet conventional superiority, the First Offset Strategy emphasized tactical nuclear weapons for which the "technological sauce" was the miniaturization of nuclear components. After the Soviets reached strategic parity and conventional deterrence seemed less credible, the Second Offset Strategy focused on precision-guided munitions and network warfare, enabled this time by a technological sauce that included computers, sensors, and stealth.

These advantages have now been lost due to the proliferation of precision-guided munitions and anti-access/area denial capabilities. According to Work, American military leaders "believe quite strongly that the technological sauce of the Third Offset is going to be advances in Artificial Intelligence (AI) and autonomy," noting that "competitors that can use AI and autonomy in a smart way are going to be the competitors that have a very big operational advantage in the future." Russian president Vladimir Putin described the current situation in a more dramatic fashion: "artificial intelligence is the future, not only of Russia, but of all of mankind ... Whoever becomes the leader in this sphere will become the ruler of the world."

**Article Two**

# Introduction to Cyber Power

by Michael Mayer

**Cyberspace has become a domain characterized by permanent conflict, filled with rapidly evolving threats and a wide range of strategic actors.**

The explosive growth in computing power coupled with global connectivity via the Internet has irrevocably altered social interactions and, therefore, the strategic landscape. Within a generation, a rudimentary network connecting a few American universities has evolved into an integrated and omnipresent part of the human experience as something we now call cyberspace.

Perhaps the most intriguing and consequential aspect of the digital realm is how far it actually penetrates. The interconnected devices range from national power grids to individual telephones, automobiles and household thermostats. This range and the sheer geographic and numerical breadth of interconnected devices shows how pervasive it already has become. The horizontal and vertical reach of cyber threats make them strategic in nature, but the ability to launch cyber attacks rests with a wide range of actors, from individuals to corporations to nation states. Cyberspace as we know it today is only a few decades old, and it is rife with competing forces.

## DEFINING CYBERSPACE

Cyberspace has been designated the fifth operational domain of warfare, alongside land, air, sea, and outer space. Unlike the other four domains, however, cyberspace is constructed entirely by humans. While technical definitions abound, one helpful way to visualize and define cyberspace is to divide it into three layers.

As Martin Libicki explains, the first layer is a *physical* one and includes all the hardware components of cyberspace such as computers, smart phones, routers, and cables.

On this rests a second layer, the *syntactic* level containing the instructions that allow these machines to function and the protocols that enable communication between them.

The third and final layer is the *semantic* layer – all the information stored on the computer itself. Some of this information is, as Libicki notes, "semantic in form but syntactic in nature" (information stored on the computer but providing instructions for the machine such as a printer driver, or software which controls machinery) while much of semantic layer is "natural language" information such as documents or spreadsheet.

For example, uploading a picture from a smartphone to the "cloud" involves the physical layer (the phone itself, the mobile tower sending and receiving signals, and the server/storage unit providing cloud storage), the syntactic layer (the phone's operating system and apps used to take the picture and connect with the cloud), and the semantic layer (the file containing the picture itself).

## INTERNET OF THINGS

This layered structure, apart from being conceptually useful when specifying the nature and targets of threats in cyberspace, is simply one of the unique aspects of the cyber domain.

The networked nature of cyberspace connects a vast array of devices. This so-called "Internet of Things" (or IoT) is by one estimate expected to reach 34 billion devices by 2020 and includes personal computers and smart phones, as well as household products

such as televisions, thermostats, refrigerators, and personal fitness bands.

It also includes corporate and military networks and national infrastructure such as electrical grids and energy pipelines. Information stored on these interconnected devices can be shared or stolen from the other side of the globe, and in ways that ensure anonymity. Much of the infrastructure of cyberspace is privately owned, and many of the most powerful actors are not nation-states.

The breadth and depth of cyberspace makes it an exceptionally complex and challenging domain for military-related security operations. Certain qualities of the Internet itself make the network particularly vulnerable. Data is routed by servers through Internet Service Providers (ISPs) which can be re-routed or disturbed either by attacking the server itself or the Domain Name System (DNS), which is the protocol and infrastructure connecting domain names (such as forsvaret.no) to their numbered Internet Protocol (IP) addresses.

In addition, much of what makes the internet function is decentralized, unregulated, and unencrypted, providing a perfect environment for low-level individual disruption or sophisticated and coordinated attacks on information infrastructures.

## GAINING ACCESS

There are both internal and external methods of gaining access to a computer system, although the internal causes are much less common. Among these are insider threats from rogue employees – an Edward Snowden scenario, for example – or inadvertently through poor cyber "hygiene."

One well-known example of this occurred in 2008, when a soldier at a U.S. military base picked up a flash drive purposely left by a foreign intelligence agency in the parking lot outside and inserted it into a computer connected to the U.S. Central Command, inadvertently uploading harmful software. The subsequent cyber breach, an incident that became known as Buckshot Yankee, required over a year of work to clean and repair machines on the network.

Another way threats can access a system is through the supply chain. A famous historical example is the thousands of World War Two Enigma machines Britain had captured from the Germans and later distributed to their former colonies, who then assumed their encrypted messages remained secure but could, in fact, be read by the British.

A modern version is the tens of thousands of counterfeit Chinese-manufactured computer chips – about 59,000 were discovered in 2010 alone – ending up in U.S. military weapons systems that could potentially cause computers to crash or missiles to malfunction. According to a 2014 Pentagon report, a cybersecurity test of 40 weapon systems revealed "significant vulnerabilities" in all of them, while a German-operated Patriot missile defense system was reportedly hacked in 2015.

## HACKING IN A VARIETY OF FORMS

Externally gaining unauthorized access to a computer or computer network – what is commonly referred to as hacking – can take a variety of forms as well. A relatively simple way to gain access is through "phishing," a technique whereby legitimate-looking emails are sent with a corrupted attachment in the hope that the unsuspecting victim will download it. Alternatively, the email may provide a link to a website that either facilitates harmful code to be downloaded or encourages the target to enter sensitive information.

A more advanced technique called "spear phishing" uses information about a specific individual to create emails that look particularly convincing. During its annual cybersecurity test, the Norwegian Security Agency (NSM) found in 2017 that 90% of public employees clicked on the NSM's phishing email link, half of those respondents actually activated the simulated malware, and one third even provided user names and passwords for their respective networks.

Another means of gaining entry is through a software vulnerability embedded in the system at the syntactic level among the millions of lines of operating system code.

## TYPES OF VULNERABILITIES

The types of vulnerabilities and how they are exploited vary. One broad category detailed by Peter Singer and Alan Friedman is the SQL (pronounced sequel) injection, which affects the Structured Query Language (SQL) often used in web applications: "an attacker, instead of entering a name and address as requested, can enter specifically crafted commands that the database will read and interpret as program code, rather than just data to be stored." In this way, hackers can access the data or gain control over the website.

Another class of vulnerability is the buffer overflow, which occurs when a program attempts to write more data to an allotted block of memory (called a buffer) and instead overwrites data in adjacent "overflow" storage areas. Exploiting this process by inserting lines of code to be written into the computer's memory can allow an intruder to gain control at the system level. A piece of code written to take advantage of a specific vulnerability is called an "exploit," which can then be transferred, sold, or saved until needed.

Clearly, vulnerabilities embedded in software code – intentional or not – are valuable resources, and not only for criminal networks, activist hackers ("hacktivists"), or state-funded groups. The National Security Agency (NSA), which has an elite hacking group formerly known as the Tailored Access Operations office (TAO), buys and collects vulnerabilities from other hackers. According to journalist Shane Harris, the NSA even pays software companies not to repair or announce vulnerabilities so that NSA hackers can exploit them.

Other classified documents show that "the NSA invites makers of encryption products to let the agency's experts review their work, with the ostensible goal of making their algorithms stronger. But the NSA actually inserts vulnerabilities into the products to use in its espionage and cyber warfare missions."

## "ZERO DAY"

No vulnerability is quite as valuable as a "zero day," meaning the attack utilizes a "net new" previously unknown vulnerability and therefore occurs on the zeroth day it is known to the rest of the world. Its value comes in large part from the element of surprise, but the vulnerability is usually a "one-time only" opportunity as the target is likely to patch it once the intrusion has been discovered.

The NSA builds most of their own cyber weapons, but also has a substantial budget – about $25 million in 2013 – to purchase zero-day exploits. According to Harris, "the NSA is widely believed by security experts and government officials to be the single largest procurer of zero-day exploits, many of which it buys in a shadowy online bazaar of freelance hackers and corporate middlemen." Several private companies sell zero-day vulnerabilities through a subscription plan and even offer a catalogue of ready-made zero-day exploits for sale.

Whether intruders gain access through phishing, vulnerabilities, or built-in backdoors in software, the goal is oftentimes to insert a ready-made exploit known as malicious software (or malware) into the system, usually delivering a "payload" of harmful code. Examples of these might be a virus (self-replicating programs that attach themselves to other software and often monopolize available memory, paralyzing the infected computer) or a "Trojan horse" that appears to be a benign program but hides harmful code.

One common payload is a "worm" that can self-replicate, use memory, and spread throughout the network. These need not necessarily be malicious – the first known instance was the 1988 Morris Worm, whose creator claimed intended only to measure the size of the Internet but infected thousands of computers and caused their operating systems to slow down to the point of dysfunction. One of the most expensive was the ILOVEYOU worm which spread worldwide during one day in May 2000, ultimately infecting 45 million computers running the Windows operating system and costing an estimated $10 billion in damage.

## ADVANCED PERSISTENT THREAT

A particularly challenging category of threat incorporating several methods of access and multiple exploits is the Advanced Persistent Threat (APT), which describes a focused and concerted effort to gain access to specific targets. The often state-sponsored attackers use any number of advanced hacking techniques such as spear phishing and zero-day exploits, and are persistent in their efforts to penetrate a network's defenses.

Much of the current APT activity has been linked to Russian groups, particularly ones suspected of receiving state funding, but Chinese hackers have also been particularly active. In the U.S. context, notes Shane Harris, "when government officials mention 'APT' today, what they often mean is China, and more specifically, hackers working at the direction of Chinese military and intelligence officials or on their behalf."

## COMPUTER NETWORK EXPLOITATION

Once the cyber intruders have gained access, they can either steal information or cause damage either in the cyber realm or in the physical world. Hacking to steal information is also called computer network exploitation (CNE), using malware (or, more aptly, "spyware") to record keystrokes to discover passwords, view e-mails sent, websites visited, or even enable the exfiltration of sensitive data.

Richard Clarke and Robert Knake described how, for example, Canadian researchers in 2009 discovered sophisticated malware they named "GhostNet" present on over a thousand computers at a number of countries' embassies around the world. The program was able to activate remotely a computer's camera and microphone, sending back the audio and video to servers in China.

It is widely believed that Chinese hackers also repeatedly breached the computer networks of Pentagon defense contractors responsible for developing the advanced F-35 stealth fighter aircraft, stealing design plans that likely formed the basis for China's own J-20 stealth fighter. The theft forced programmers to re-write large portions of the software code.

A Russian-linked hacker group designated APT 28 (also known as Fancy Bear), is suspected of repeatedly infiltrating government computer networks and stealing data, including networks belonging to the German parliament in 2015, the U.S. Democratic Party headquarters in 2016, and the German foreign ministry in 2018.

The United States has been active in cyber espionage. As the *New York Times* reported, "the N.S.A. has embraced hacking as an especially productive way to spy on foreign targets. The intelligence collection is often automated, with malware implants — computer code designed to find material of interest — left sitting on the targeted system for months or even years, sending files back to the N.S.A."

Ironically, the NSA itself is not entirely safe from theft as it discovered in 2016 when a mysterious group calling themselves the Shadow Brokers infiltrated the agency and stole highly classified data and advanced cyberweapons.

## INDUSTRY ESPIONAGE AND HACK-BACKS

While governments remain attractive targets for CNE efforts by other state actors as well as non-state groups, there is also a significant amount of activity at the corporate level. Online industry espionage, damaging malware, and criminal activities cost the private sector huge sums of money.

Many are preparing to fight back by enlisting the help of former defense and intelligence service veterans who have gone into the lucrative cybersecurity business. One such firm is CrowdStrike, which will create lookalike networks for their corporate clients to lure in hackers (so-called "honeypots" or, in this case, "honeynets"), thus revealing what intruders are looking for and the techniques used.

Active retaliation after a cyber intrusion or "hack-backs" is illegal in the U.S. In 2013, however, Microsoft joined forces with a group of financial institutions to do precisely that. Its target was a notorious cybercrime group called the Citadel which had used

thousands of infected computers as botnets to infiltrate bank networks to steal credit card information.

After receiving permission from the U.S. justice system, Microsoft launched a long, complex, and ultimately successful counter cyberattack to gather information on their attackers that involved law enforcement agencies in over 80 countries. According to Harris, banks have been collecting zero-day vulnerabilities and exploits to retaliate in case of a massive cyberattack.

## COMPUTER NETWORK ATTACK

Unauthorized network access can also be motivated by even more nefarious intentions that cause disruption to the network, extensive corruption of systems or data, and even damage in the physical realm. These actions cross the somewhat arbitrary threshold from CNE to CNA (computer network attacks).

A malware designed to affect computers by locking operating systems or threatening to erase data unless the owner sends payment (often in the online currency Bitcoin) is called ransomware. In May 2017, ransomware called "Wannacry" spread across the globe, seriously affecting networks such as those of the British national health service, Indian police departments and a Spanish telecom company.

Attributed by some (the United States and Britain) to North Korea, Wannacry used a Windows vulnerability called EternalBlue. This was one of the NSA's most valuable hacking tools, reportedly stolen when the mysterious Shadow Brokers group breached the NSA's own network.

Not only did the Shadow Brokers release the vulnerability, it also released a pre-made exploit for EternalBlue which was then retooled and paired with a worm to make Wannacry particularly invasive. Microsoft president Brad Smith compared the theft to "the U.S. military having some of its Tomahawk missiles stolen." The NSA notified Microsoft after discovering the theft, which then released a patch to address the vulnerability.

## BOTNET VIRUSES

Other malware takes control of part of a user's computer – often without them even realizing it – and using it as part of an automated yet coordinated attack of "botnets." A bot is simply an application that performs an automated task (Apple's Siri is a bot), so that a botnet is a network of computers functioning as bots toward a common goal.

In 2009, it is believed that North Korean hackers launched a coordinated attack using a botnet virus. Over 40,000 computers around the world began sending page requests to certain U.S. and South Korean servers. The flood of traffic reached the level of over 1 million requests per second, temporarily bringing down the web servers of the U.S. Treasury, Secret Service, U.S. Trade Commission, and the Department of Transportation. The distributed denial of service (DDOS) attack reached its peak a few days later as 166,000 computers in 74 countries flooded South Korean bank and government agency websites.

Two years earlier, Russian hackers were most likely responsible for the DDOS attack in Estonia that is often referred to as the first major instance of a state-sponsored cyber attack. The incident temporarily paralyzed the banking sector, national newspapers, and online government services. The botnet worm used in the attack was so pervasive that over a million computers were flooding Estonian servers with page requests.

One of the largest botnet DDOS attacks occurred on 21 October 2016, targeting Dyn, a company that controls a significant portion of the Internet's domain name system infrastructure, with a unique botnet. Instead of using computers, the Mirai botnet infected and harnessed smaller devices comprising the Internet of Things, including routers and security cameras that have limited cybersecurity features. The result was a massive 1.2 Terabyte per second attack that overwhelmed Dyn servers and disrupted websites such as Twitter, the Guardian, Netflix, and CNN.

## DESTRUCTION

In addition to disruption, cyber attackers can also cause significant and irreparable damage to computer networks and stored data. In December 2011, the hacktivist group Anonymous gained access to the security analysis company Stratfor through its website, stealing employee emails and the personal records and credit card information for 60,000 customers.

Afterwards, the hackers managed to "effectively destroy" four Stratfor servers containing years of the company's analytical data and reports – the core of the company's business. Jeremy Hammond, the hacktivist convicted of the crime, later explained that "first you deface, the you take the information, then you destroy the server ... so they can't rebuild the system. We don't want them to rebuild. And to destroy forensic evidence that could be used to find out who did it and how it was done."

Cyber attacks affecting critical military systems can be used in conjunction with a conventional attack, as the Israelis demonstrated in September 2007 during what became known as "Operation Orchard": When a laptop belonging to a Syrian official was hacked by Israeli agents during 2006 and the information exfiltrated, they discovered evidence of a secret plutonium processing plant in Syria being constructed with assistance from North Korea.

This led to seven Israeli F-15s crossing into Syrian airspace on 6 September 2007 and bombing the facility in question without detection by a single anti-aircraft battery. As Singer and Friedman describe, "the Israelis had successfully penetrated the Syrian military's computer network, allowing them to see what the Syrians were doing as well as direct their own data streams into the air defense network. This caused the Syrian radar operators to see a false image of what was really happening."

## STUXNET

One example of a cyber attack using malware to cause actual physical damage is the Stuxnet virus. Usually attributed to the United States and Israel, it is perhaps the most well-known and most sophisticated cyber attack to date and often considered the first real use of cyber weapon.

The target, Iran's uranium enrichment facility at Natanz, used thousands of centrifuges connected to computers known as programmable logic controllers which manage their operation. Although the network was not connected to the internet, careless use of flash drives by some employees may have provided an opening. A software "beacon" was installed that sent back details on the centrifuges, followed by a complex piece of malware that was constructed and inserted into the facility's network.

The worm (later named Stuxnet by Microsoft based on a combination of file names in the malware) first recorded signals on the network indicating normal centrifuge operation. Then, while playing back the "all systems normal" signals, began to disrupt the centrifuges by spinning them too fast or suddenly applying the brake. The Iranians became distrustful of their own instruments as up to several thousand of the 8,700 centrifuges were ruined and needed to be replaced within a few short months during 2010.

The Stuxnet worm eventually found its way onto the Internet, however, and versions of it soon surfaced around the world. A team of cyber security specialists at Symantec began analyzing the malware and immediately found its complexity and sophistication suspicious. The code utilized previously unseen techniques and multiple "zero day" exploits, prompting some to declare it "the most complex malware ever written."

## INDUSTRIAL CONTROL SYSTEMS

Critical public and private infrastructure, from pipelines to power grids and waste treatment plants, is also managed by similar industrial control systems (ICS). The largest subset of these are supervisory control and data acquisition (SCADA) systems that monitor and control flows and remotely perform system diagnostics. The networked nature of these systems and the uneven protection of the many smaller private regional companies

make them particularly vulnerable to cyber intrusions and attacks.

A type of software code that can be planted in a computer network is sometimes referred to as a "logic bomb." This is code that can lie dormant in a system but once activated causes the computer to damage or destroy data, its own hardware, or even physical systems connected to those data networks.

One of the first alleged uses of this ostensibly occurred when the Soviet Union, eager to acquire commercial technology for its oil and gas industry during the 1980s, stole code from a Canadian firm producing industrial control systems that governed the operation of pipeline pumps and valves. The Central Intelligence Agency had anticipated the theft, however, and had planted malware in the code. Initially, the technology installed on the Trans-Siberian gas pipeline operated normally but eventually started to intentionally malfunction, increasing the pump's pressure in one section while simultaneously closing a valve at the other end. The subsequent three kiloton explosion in June 1982 was the largest non-nuclear blast ever recorded.

## NORWEGIAN OIL AND GAS

These types of industrial targets have become even more vulnerable as industrial control systems software becomes more widespread and more capable. In Norway, unsecured network servers controlling industrial processes at Statoil's Mongstad refinery led to a temporary production halt when outsourced data consultants in India conducting remote data maintenance mistakenly gained access to the servers. Investigative reporting discovered 29 similar instances of accidental breaches.

The ability of actors to exploit digital vulnerabilities within private Norwegian oil and gas infrastructure was recently analyzed in a comprehensive report by the Norwegian Institute of International Affairs (NUPI), which outlined how public-private partnerships could contribute to more resilient networks.

In 2017, hackers possibly linked to Iran breached the networks of the world's largest oil company, Saudi Aramco, depositing a piece of malware called "Triton" that attempted to alter the emergency shutdown system at one of Aramco's facilities. The attack ultimately failed and the malware was discovered.

In one case, industrial vulnerability had very real consequences. A German steel mill fell prey to a sophisticated cyber attack in 2014 that hindered the ability of a blast furnace to perform a controlled shut down, causing massive damage.

## VULNERABLE INFRASTRUCTURE

The vulnerability of computer networks upon which modern society depends has not been lost on policy-makers. President Obama announced in May 2009 that "cyber intruders have probed our electrical grid and that in other countries cyber attacks have plunged entire cities into darkness." Chairman of the Joint Chiefs of Staff Michael Mullen acknowledged in 2011 that "the effects of a well-coordinated, state-sponsored cyber-attack against our financial, transportation, communications, and energy systems would be catastrophic."

In 2015 and again in 2016, Ukraine suffered attacks to its energy infrastructure via a malware package called BlackEnergy. It included an additional component called Killdisk which together destroyed computer hard drives, sabotaged control systems, and was able to send commands directly to critical power grid control systems. The cyber attack left several hundred thousand people without electricity for several hours.

The head of Britain's cyber security center, Ciaran Martin, stated in 2017 that Russia had penetrated the country's energy and telecommunications sectors. Similarly, the Trump administration announced in March 2018 that Russian hackers had infiltrated U.S. and European power plants and electrical grids, ostensibly achieving the ability to manipulate or shut down power plants.

After reviewing the evidence from the Russian intrusions, Eric Chien from the cybersecurity firm Symantec concluded that "they're sitting on the machines, connected

to industrial control infrastructure, that allow them to effectively turn the power off or effect sabotage … all that's missing is some political motivation." In effect, noted PwC cybersecurity expert Brad Bauch, the intrusion was "a supply-chain attack vector … A software company was attacked, and malware was then injected into the software that a lot of companies use."

## STRATEGIC ASPECTS OF CYBER POWER

Given the features of cyberspace, how might states act strategically within this domain and across all domains? This question goes beyond simply ensuring that national networks and digital infrastructure are secure, although cybersecurity is a crucial part of any comprehensive approach to cyberspace.

This section will address cyber power, which has been defined as "the ability to use cyberspace to create advantages and influence events in the other operational environments and across the instruments of power." In other words, how can offensive and defensive capabilities in cyberspace be used by state militaries such as Norway or the United States within that domain and across the other military domains to defend, attack, deter, or coerce?

Despite the declaration by the United States and its NATO allies that cyberspace constitutes the fifth domain of warfare, there are clearly significant differences between it and the other domains. Among these are the structural features of cyberspace relating to geography and infrastructure, the wide mix of relevant and interrelated actors along with the capabilities at their disposal, and the unique characteristics of activities in cyberspace that contribute to an exceptionally complex strategic environment.

Even the most basic of strategic actions are easier to observe and interpret in the other domains due to their visibility and (usually) the ability to ascribe attribution.

Basic concepts – including what constitutes an "attack" in cyberspace or when a series of events could be labeled "cyberwar" – lack a commonly agreed upon framework. The United Nations Group of Governmental Experts (GGE) agreed in 2013 that international law applies in cyberspace, but failed to reach a consensus in 2017 on a collective set of norms and principles. Analysts must therefore exercise a great deal of caution, particularly when making analogies to the dynamics of other strategic relationships.

## THE DISTINCTION BETWEEN CYBER EXPLOITATION AND CYBER ATTACK

One significant distinction has already been made between cyber exploitation and cyber attack. The former refers to infiltration and espionage. The latter refers to actions that cause actual damage ranging from the corruption or loss of data to physical effects such as damage to industrial processes or system paralysis (as in the case of Israel's Operation Orchard or in a DDOS attack).

Even so, the distinction between exploitation and attack is fluid, particularly when networks must be infiltrated and probed as a precursor to launching attacks. NATO has warned that a cyber attack of sufficient magnitude threatening critical military or civilian infrastructure may be enough to trigger an invocation of the alliance's Article 5 collective defense clause.

Actions in cyberspace are easily shrouded in anonymity or disguised to appear as if other actors were in fact responsible. One simple and readily available method is TOR (The Onion Router), downloadable browser software used to access an open network and achieve anonymity via a group of volunteer-operated host routers.

At other times, entities such as governments or multinational corporations are loath to admit their networks have been compromised. Even if the intruders are identified, withholding that information may be desirable to retain some tactical advantage. Whereas physical proof may be more readily available in the other four domains, evidence of an attack and the identity of the attacker depends on not only solid cyber forensics but also the credibility of the victim/target reporting the attack.

## WIDE MIX OF ACTORS

Further complicating the attribution problem is the wide mix of actors in cyberspace and the ready availability of hacking expertise and premade exploits. Although state actors have far greater resources and expertise, the global reach of the Internet allows private actors with high-level hacking skills to generate attacks that would normally be considered strategic acts with large-scale impact – whether paralyzing air traffic control networks, erasing financial information or disrupting power grids. The combinations are many: private actors attacking corporate networks to disrupt financial markets, private groups backed by a state attacking either corporate state actors or other states entities, state agencies "hacking back" against private actors.

One prominent feature of cyberspace is the constant offensive-defensive battle constantly occurring in the domain, with smaller attacks numbering in the millions each year. The threshold for conducting cyber attacks is clearly quite low, likely because the risk of human casualties is so far almost nonexistent. The visual effect of destroying data is far less dramatic than exploding buildings or infrastructure, even though the cumulative effect of a massive cyber attack may eventually be similar to a physical attack.

## BORDERS

The lack of clearly defined borders in cyberspace creates another set of challenges. Geography still matters, particularly considering the physical layer of the Internet that includes fiberoptic cables and other tangible components of cyberspace. In the physical world, though, territorial borders, maritime exclusive economic zones, and national airspace designations are designations that can be clearly demarcated and defended.

The informational flow patterns in cyberspace are far more diffuse and full of private actors supplying the digital infrastructure upon which societies rely to conduct a wide range of commerce and social interactions. State actors including defense ministries and intelligence services also rely on privately owned digital infrastructure to conduct their operations. This makes "pulling up the digital drawbridge" less feasible in our modern networked society.

Yet "virtual borders" are emerging in a variety of forms. At one end of the spectrum, states with extremely limited digital networks – North Korea for example – are less vulnerable simply because the lack of digital infrastructure limits a potential attacker's ability to penetrate deep into the country. In China, Internet use is widespread but tightly controlled in a layered approach. The physical infrastructure is routed through major cities and monitored, the state wields institutional control over telecommunications companies, and thousands of individual internet monitors police the Internet for subversive online behavior. European market democracies face perhaps the greater challenge due to numerous entry points, interconnected networks, and strict civil liberty legal protections.

Nevertheless, more active monitoring and "bulk interception" of digital communications crossing national boundaries are part of a re-assertion of national sovereignty in cyberspace. The proposal to establish a so-called "digital border defense" (digitalt grenseforsvar) in Norway can be viewed as part of this emerging trend. The monitoring and interception ability of national intelligence services in other countries has been crucial for intelligence gathering, detecting and eventually establishing attribution for a range of security threats, including cyber attacks.

## CYBER DEFENSE

From a cybersecurity perspective, though, the first line of defense for an individual computer (or "endpoint") is antivirus (AV) software and regular updates, which function as a highly automated filter identifying the signature characteristics of the hundreds of millions of pieces of known malware. As journalist Kim Zetter reported, "of the more than 1 million malicious files Symantec and other AV firms receive monthly, the majority were variations of already-known viruses and

worms. These were processed automatically without human intervention. Algorithms searched the files for telltale strings of data or behavior to identify the malware."

Unfortunately, the rise of polymorphic malware that uses a technique allowing the underlying code to shift (and thus altering the malware's signature) has limited the efficacy of traditional endpoint antivirus programs. More advanced defensive measures using machine learning are able to deconstruct suspicious files to identify damaging code. In other cases, the malware can be isolated by a virtual machine (sometimes called a "sandbox") that is able to safely simulate in quarantine how the code will behave.

To defend an entire network, virtual firewalls able to filter external data are used in conjunction with similar antivirus software and more active defenses such as virtually testing suspicious files. These automated intrusion detection systems can also identify suspicious behavior. After a particularly egregious intrusion by the Chinese, Google and the NSA constructed a system to monitor intrusions in Google's networks. As Shane Harris reported, the system used

> [a]utomated sensors and algorithms to detect malware or signs of an imminent attack and take action against them. One system, called Turmoil, detects traffic that might pose a threat. Then, another automated system called Turbine decides whether to allow the traffic to pass or to block it. Turbine can also select from a number of offensive software programs and hacking techniques that a human operator can use to disable the source of the malicious attack … the source can be injected with a virus or spyware, so the NSA can continue to monitor it.

## PHYSICAL SEPARATION

The final and most extreme method of network protection is the creation of an "air gap," physically separating a network from the Internet. Critical infrastructure such as power grids are often air-gapped. Most military and intelligence services have dual systems – one connected to the global Internet and another internal classified intranet.

However, even an air-gapped system is vulnerable to human carelessness (exemplified by flash drives used to transport the Stuxnet virus or in the Buckshot Yankee incident). Air-gapped networks can also be vulnerable to supply chain attacks or more advanced techniques such as using FM radio signals sent by an air-gapped computer's graphics card to exfiltrate data. One piece of malware developed at an Israeli university, once inserted into an air-gapped computer, is able to adjust the rotation speed of the affected computer's cooling fan to achieve a particular frequency that can then be used to exfiltrate data to a nearby listening device such as a smart phone.

## CYBER OFFENSE

Offensive operations in cyberspace, particularly strategic CNA operations such as developing and planting the Stuxnet virus, are complex and require significant resources and expertise. The process of using a cyber weapon can be divided into a seven-step "cyber kill chain": reconnaissance of the target network; weaponization of the malware; delivery of malware via a vulnerability, a USB drive or phishing; the exploit phase; installation onto the computer; command and control communication with the host machine; and lastly, completing the action by carrying out the objectives (corrupting data and/or computers, or causing physical damage to machines in the network).

Each of these steps is fraught with uncertainty for any would-be attacker, particularly a state actor intent on carrying out a precise and predictable attack. The act of network infiltration – while a necessary precursor to an attack – also risks revealing information about the attacker, including the vulnerabilities used to gain access.

Without in-depth knowledge of the target network, however, malware cannot be specifically designed to function as the attacker intends. This is why CNE may be considered a hostile act even if no actual damage is done – it may be compared with a conventional

military exercise that violates the target's national sovereignty to test the aggressor's ability to launch an attack. Even with near-perfect information regarding an adversary's network, an attacker can never be completely certain how a malware cyberweapon will perform once released.

## THE ELEMENT OF SURPRISE

Just as the simulated attack also reveals how the aggressor might conduct an attack, using vulnerabilities and exploits – particularly zero-day vulnerabilities – also constitutes a "one-shot" capability. Cyberattacks depend on the element of surprise. Once the exploits are used to gain access or launch a malware attack, the vulnerability is revealed and can be repaired or compensated for by the defender. Even stockpiled exploits have expiration dates (although these may be unknown) as software is updated and vulnerabilities are continually discovered and patched.

This creates an unfortunate "use it or lose it" pressure for some cyberweapons. Additionally, the more specifically tailored the weapon's code, the more likely it is to be effective but a narrowly focused weapon will have limited effect. As Thomas Rid and Peter McBurney argue, "the cost-benefit payoff of weaponized instruments of cyber-conflict may be far more questionable than generally assumed: target configurations are likely to be so specific that a powerful cyber-weapon may only be capable of hitting and acting on one single target, or very few targets at best."

A surprise "bolt from the blue" attack may be rational from an operational standpoint but may not necessarily be effective in accomplishing any strategic objective unless connected to offensive threats in other domains. Most networks can be repaired within hours or days, making the reduced level of trust in the network (i.e. is the network still compromised? Does the intruder still have access?) the most significant aspect of an attack.

Clearly, some cyberattacks could have widespread consequences with the risk of significant loss of life (air traffic control networks, power grids during winter in northern climates) but perhaps not significant enough to cause governments to cede sovereignty or territory. Even Stuxnet only managed to temporarily slow Iran's enrichment program, delaying any eventual nuclear weapon development by mere months.

## CREDIBLE THREATS

Given the nature of cyberweapons, threatening or brandishing cyberweapons is challenging. It requires successfully penetrating an opponent's network without releasing a payload but instead leaving unique signs or a "calling card" revealing the attacker's identity to give the subsequent threat credibility. Afterwards, as Martin Libicki noted, "penetrating a system and persisting within it require similar skill sets but different technologies. Penetration requires knowledge of vulnerabilities; persistence requires knowing how to evade intrusion and anomaly detection systems" and neither does "breaking into a system prove the ability to break a system."

In this sense, the uncertainty surrounding the ability to retain access after announcing an intrusion and the uncertain effects of a cyberweapon make cyber threats less attractive. On the other hand, such threats are more credible given that digital networks are the targets, rather than human lives (although, again, the loss of certain networks might lead to loss of life). The political reactions to a massive cyberattack have yet to be tested, making the risk of escalation equally uncertain.

## CYBER DETERRENCE

These issues are central to deterrence in cyberspace. Deterrence is based on the concept of raising the costs – either through the credible threat of punitive attack (deterrence by punishment) or through defensive measures (deterrence by denial) – to persuade a potential adversary to refrain from attacking. Given the issues raised above (attribution, credibly and repeatedly holding a network at risk over time) deterrence by punishment strategies in cyberspace are usually judged to be challenging.

As Joseph Nye has pointed out, however, deterrence by denial strategies are intrinsically attractive: "Good cyber defenses … can build resilience or the capacity to recover, which is worth in itself; they can also reduce the incentive for some attacks by making them look futile."

On the other hand, Annegret Bendiek and Tobias Metzger referenced a statement by Obama administration official Michael Daniel and noted that "deterrence-by-denial differs greatly, since in cyber 'you have to work from the assumption that your networks are already compromised,' meaning deterrence is constantly failing."

## INTERDEPENDENCE AND INTERNATIONAL NORMS

Nye also suggests dissuading cyber attacks through an emphasis on the interdependence (entanglement) relationships between aggressor and target or through a methodical establishment of international norms (or taboos) against such attacks, particularly if the focus is on "a taboo not against certain types of weapons but against certain types of targets."

Deterring cyber attacks through kinetic means is always an option – an example of what has been termed "cross-domain deterrence." For example, the Trump administration's 2018 Nuclear Posture Review signaled the possibility of nuclear retaliation to a non-nuclear attack on U.S. infrastructure or military command and control infrastructure that was widely interpreted to include a devastating cyber attack.

Clearly, a nuclear retaliation in reaction to a cyber attack raises issues of proportionate response, one of the main principles in the Law of Armed Conflict (LOAC). Although a massive counterattack might be disproportionate, cyber attacks themselves are not necessarily LOAC compliant. In particular, the focus on civilian infrastructure and civilian "collateral damage" is at odds with the principle of distinguishing between combatants and non-combatants in a conflict, just as attacking a state's digital infrastructure may not be necessary from a military perspective (compared with, for example, disabling an anti-aircraft sensor).

## CYBER AS ONE COMPONENT IN CONFLICT

A strategic cyber conflict – that is, a cyberwar between two or more actors that continually results in real and widespread digital and/or physical damage – waged exclusively in cyberspace without spilling over into other domains seems unlikely. Some scholars, such as Chris Demuchak, have argued that cyberwar is too narrow a term and we should instead prepare for *cybered* conflict, in which cyberspace is simply one domain across which operations will take place.

Martin Libicki suggested three operational cyberspace attacks that would be valuable in a military conflict: eruption (the digital illumination of an adversary's military targets to reveal their positions); disruption (temporarily degrading their systems as in Operation Orchard); or corruption (incapacitating missile guidance systems or "bricking" an opponent's computers).

Clearly, it would be operationally advantageous not to reveal that an adversary's defensive systems – such as air defenses – were compromised until an attack materialized so as not to risk losing the exploit, which limits its use as a coercive measure. Nevertheless, the ability to demonstrate a capacity to infiltrate will lower the confidence in the defensive value of such systems.

## MILITARY FORCE RELIES ON DIGITAL NETWORKS

Mirroring societal technological trends, military force has over the past several decades become increasingly reliant on digital networks for situational awareness, communication, navigation, and targeting. Degrading these systems with cyber weapons would be a crippling disadvantage in a crisis or conflict situation. The command, control, and communications infrastructure for nuclear weapons systems (NC3) is of particular concern in this regard, with multiple pathways for infiltration through sensors and communication systems.

In the transatlantic context, defense experts James Miller and Richard Fontaine observed, "cyber penetration of critical infrastructure amounts to what the military calls 'preparation of the battlespace.' Russian cyber implants in the United States and other NATO countries provide potential leverage in a crisis, and – if push comes to shove – the ability to impose significant pain through non-kinetic, non-lethal cyber attacks."

NATO has also adopted a more proactive stance on defending against cyberattacks. A working group consisting of the United States, Britain, Germany, Spain, Denmark, the Netherlands and Norway began work in 2017 on a cyber doctrine that will in part determine when deployment of cyber weapons might be justified.

In response to cyber threats, the Pentagon is examining ways in which to "harden" its digital networks and systems against cyberattacks, particularly given that many platforms contain commercial off-the-shelf software that may have pre-existing vulnerabilities. Additionally, the U.S. Naval Academy has re-introduced courses on sextant use and celestial navigation to prepare sailors for an operational environment in which computer systems or Global Positioning System are rendered inoperable.

## CYBERSPACE WILL PLAY A CRUCIAL ROLE

It is becoming increasingly clear that cyberspace will play a crucial role in any future conflict, simply because so many facets of society – including state infrastructure, military systems, and social media – are vulnerable to degradation or manipulation.

The battle over political influence is particularly intense in cyberspace. Twitter troll farms or Facebook campaigns like the Russian government-sponsored groups that actively sought to influence the 2016 presidential elections in the United States are example of coordinated political influence activities that utilize social media platforms. Such activities are a significant threat to open democratic processes but should nevertheless be considered separate from hackers exploiting network vulnerabilities to cause irreparable damage.

One of the fundamental concerns with large numbers of Russian-controlled automated bots contributing to the political dialogue in the United States and elsewhere is attribution – the accounts are not clearly identified as belonging to agents linked to foreign governments. This allows groups to act as provocateurs, exacerbating existing political divisions to generate greater domestic discord in ways that would not be possible if the origin of the social media activity were apparent. Some attempts are underway to rectify this situation, including the *Hamilton 68* project at the German Marshall Fund that tracks Russian bot activity on Twitter.

Although social media botnet campaigns similar to those waged by Russian groups in democratic election processes in the United States and Europe should not be considered cyber attacks, such tactics could play an important disinformation role in a future crisis scenario. False news videos spread via sites such as Facebook could incite violence in certain domestic groups or even provide justification for state aggression.

In a crisis, information warfare and cyber attacks may combine with artificial intelligence, autonomous systems, and traditional forms of military conflict in a form that some have labeled "hyperwar." Waging these types of conflicts – or deterring them, for that matter – entails a comprehensive approach that integrates cyber capabilities within the broader portfolio of military instruments of power.

**Article Three**

# AI and Autonomy in Cyber Operations

by Michael Mayer

**AI controlled cyber weapons will necessitate equally powerful AI driven cyber defenses, raising legitimate concerns about retaining adequate human control over these technologies.**

The rapid growth of artificial intelligence in the private sector and a similar interest among leading militaries with advanced technology almost certainly ensures that AI and autonomous machines will feature prominently in future defense acquisitions. Artificial intelligence will likely influence cyber operations in at least three distinct ways.

First, the advent of AI powered autonomous systems for civilian use, along with military autonomous weapon systems in the context of network-based warfare, will present new digital vulnerabilities and therefore new threats. Second, offensive AI-driven cyber weapons will present new challenges to the security of digital networks and new opportunities to attack adversarial networks. Third, increasingly powerful offensive AI cyber weapons require equally capable AI-powered defensive capabilities. These developments will in turn have significant strategic implications.

## THE DARPA CYBER CHALLENGE

In 2016, seven teams gathered in a Las Vegas hotel to compete in a cybersecurity competition hosted by the U.S. Defense Advanced Research Projects Agency (DARPA). The goal of the 2016 DARPA Cyber Grand Challenge was to test the ability of an autonomous AI-powered bot to independently repair security vulnerabilities in its own machine while exploiting those of others. On a $55 million virtual playing field constructed of seven supercomputers, isolated from any other networks and loaded with software for the competitors to hack, each team's bot searched and exploited vulnerabilities in the software while "deciding" how best to protect and repair "holes" in their own system.

Through 96 rounds of competition, DARPA regularly introduced malicious software previously discovered throughout the cyber sphere. Points were awarded for successfully exploiting vulnerabilities in the other systems, while identifying and repairing vulnerabilities to keep the teams' own systems running. Some of the bots performed exceptionally well, discovering some threats much quicker than human analysts, while making strategic judgments regarding the balance between offense and defense. As one team leader explained to *Wired* journalist Cade Metz,

> If the bot found a hole in its own machine, it wouldn't necessarily decide to patch, in part because patches can slow a service down, but also because it can't patch without temporarily taking the service offline. Through a kind of statistical analysis, the bot weighed the costs and the benefits of patching and the likelihood that another bot would actually exploit the hole, and only then would it decide whether the patch made sense and would give it more points than it would lose.

The bots were far from infallible, as one system even chose to launch an attack on its own machine. While the contest revealed that the effectiveness of the autonomous bots remained far beneath that of human experts, the overall performance was surprisingly effective in some areas and demonstrated that

autonomous cyber defenses had significant potential.

## AI CREATES BROADER CYBER THREAT

The strategic, legal, and ethical issues that arise with the deployment of autonomous systems – both civilian and military – have been addressed elsewhere in the academic literature, but the vulnerabilities of autonomous systems to digital threats is also highly relevant. The Internet of Things presents multiple entry points for attackers, and increased automation and machine autonomy in everyday life itself represents a strategic vulnerability.

Consider the spread of autonomous features in automobiles. Just as aircraft use onboard computer systems to control the functionality of the craft, cars have similar digital functionality. Computer-controlled driver assistance features including anti-lock braking, driver-assist steering, automatic parking and "autopilot" provide opportunities for hackers to gain external control over the vehicle via onboard wi-fi or Bluetooth connectivity.

Two hackers demonstrated this in 2015 by using an online laptop not in close geographic proximity to the vehicle (a 2014 Chrysler Jeep Cherokee) to gain access to its onboard systems, deactivating its brakes and transmission.

For criminal networks, installing ransomware that would require payment to regain control of the car is distinctly possible. More worrisome are exotic applications such as carrying out assassinations via auto accident. Even more broadly, it is worth emphasizing that the vulnerability exploited for the Jeep hack was manufacturer-wide and applied to entire classes of vehicles.

At an autonomous vehicle conference in 2017, cyber security expert Joshua Corman outlined possible consequences of this, describing one simulation in which an entire make of car (all Volkswagens, for example) were disabled to block bridge and tunnel traffic around New York City during a terrorist attack. As advances in AI enable fleets of autonomous vehicles for urban public transportation, the networks upon which they depend are vulnerable to attacks that could cause large-scale civilian casualties.

As military systems within a network-based warfare framework become increasingly autonomous – with intelligent unmanned system development in all domains – the vulnerability of these to outside interference or control becomes a serious concern. Given the likelihood of autonomous swarms of weapons platforms in the future battlespace, an ability to corrupt, disable, or remotely steer the swarm would be decisive.

The U.S. Army has installed AI systems in the ground control stations for its fleet of armed drones (the MQ-1C) giving them the ability to operate in human-managed swarms, but also incorporated AI cyber defenses to protect the system. The CEO of Scorpion Computing Services, Walter O'Brian, whose AI is used by the Army, observed that "it's an arms race.... Now I have an AI protecting the data center, and now the enemy would have to have an AI to attack my AI, and now its which AI is smarter."

## AI AND CYBER OFFENSE

As the 2016 DARPA Cyber Challenge demonstrated, the future of cyber warfare will be greatly influenced by artificial intelligence. A recent report by AI experts outlined a number of potential malicious uses for the technology. Among them was precisely the possibility of intelligent software and machine systems orchestrating large-scale automated attacks with an ability to react and tailor their methodology to environmental changes.

The magnitude and relentlessness of highly automated and intelligent probing of network defenses or software analysis searching for vulnerabilities will almost certainly constitute a powerful offensive cyber capability simply through the sheer mass and persistence of the attacks. One simple threat might be a fully automated spear phishing system able to tailor tweets based on a particular user's interests that have been gathered from online data such as social networks, increasing the likelihood that they will click on a malicious link.

An even more complex and troubling scenario suggested by one cybersecurity expert is an AI-powered botnet that "could dynamically and fluidly react to countermeasures, intelligently developing its own complex warfare strategies beyond anything any human has ever developed, involving complex layering of attacks to mitigate standard countermeasures and reacting at a speed that no human network administrators could hope to match."

Artificial intelligence and reinforcement learning can also be applied to craft intelligent agents able to manipulate its own code to evade antivirus software. Rather than existing polymorphic malware using static pre-coded algorithms, the next generation may be able to autonomously and spontaneously generate new customized attacks using machine learning.

Suggested approaches to detect this type of malware include the Malware Analysis and Attributed using Genetic Information (MAAGI) that leverages AI based on the similarities between the behavior of malware and biological organisms.

The algorithms and architecture behind the AI itself will also be at risk. Counter-AI systems are under development to exploit the unique vulnerabilities in AI-powered autonomous systems. For example, machine-learning systems can be corrupted by "poisoning" the training data sets in ways that lead the system to misclassify patterns, but in a predictable manner that may be advantageous to an adversary. One study demonstrated how so-called "adversarial patches" can be introduced to an image recognition deep learning system via images that are unable to be detected by the human eye but cause the AI to misclassify images.

## AI AND CYBER DEFENSE

Offensive use of AI in cyber attacks will, as O'Brian stated, demand AI enabled defenses. One tool for cybersecurity is intrusion detection prevention systems (IDPS) that protect systems or networks by identifying normal usage patterns to detect abnormal and potentially malicious behavior.

Currently, IDPS has some inherent weaknesses, including low detection rates for abnormal behavior, limitations in data processing, an inability to handle abrupt changes to data traffic volume, and a lack of automation that often requires human intervention and analysis.

These challenges can arguably be addressed using AI, particularly since large data flows and pattern recognition are among the most prominent features of artificial intelligence applications. For example, intelligent agents equipped with decision-making and self-learning abilities might be used to monitor networks for threats during their reconnaissance phase, perhaps even modeled after the human body's immune system.

Neural networks can be useful for monitoring network traffic by identifying normal patterns of use and thereby also enabling the system to detect abnormal and thus potentially malicious activity. Deep neural networks might even be able to go further, using network data to successfully predict attacks. AI expert systems are already widely in use to assist humans in evaluating potentially malicious activities or auditing infiltrated system data, providing a set of options if threats are discovered.

Similarly, the winner of the 2016 DARPA Cyber Challenge – an AI software called "Mayhem" – is being adapted to automatically find and patch flaws in commercial software, including internet routers. Although still at the demonstrator stage, its engineers envision scenarios in which large-scale attacks such as the 2016 Mirai botnet attack can be averted by not only automatically detecting the attack and identifying the vulnerability but also patching it automatically before more systems are infected. One inhibiting factor to implementation of the technology is the reliance on AI to make decisions and relegate humans to a monitoring role in the decision loop.

## STRATEGIC IMPLICATIONS

There is an ongoing race for artificial intelligence. Unlike previous arms races, this competition is more democratic, more egali-

tarian, and therefore more unpredictable in nature.

Previous generations of strategic competition were dictated by access to natural resources and industrial capacity, which could then be utilized by human ingenuity. The components necessary to substantially and negatively impact the lives of entire countries remained largely in the hands of nation-states, whether it was a nuclear armed long range ballistic missile or a chemical or biological agent (although the latter is in the midst of a disturbing revolution of its own).

In contrast, artificial intelligence technology will "supercharge" tools and techniques in cyberspace, and is being driven forward by the civilian sector. The raw materials necessary for developing a powerful AI agent – intellectual capacity and computing power – are widespread, more easily accessible, and more easily employed than other weapons of mass destruction.

The strategic landscape within cyberspace is therefore more complex than in other domains that are completely dominated by national militaries, because much smaller actors can compete against regional or global powers in ways unimaginable in other domains.

The fact that the premiere U.S. actor in cyberspace – the NSA – can itself be hacked and its cyberweapon stockpiles stolen speaks volumes about the realities of the domain. This fact makes strategic decision-making in cyberspace far more difficult simply because the threats are less visible and the risk of strategic surprise is greater. Access to AI technology and techniques will offer powerful advanced machine intelligence to any number of groups able to pay for it, and the proliferation risks may be difficult to manage.

## AUTOMATION AND AUTONOMY

The trend toward greater automation and machine autonomy is likely to continue in the realm of military technology for the same reasons it will continue in the civilian sector – the advantages are simply too numerous to ignore. As machine intelligence becomes more prevalent in autonomous systems on the conventional battlefield, it will increase the speed of tactical engagements as AI systems analyze and respond to an adversary's actions.

The decision loop may become so truncated that human decision-making will be too slow to be decisive on the battlefield, particularly when faced with networked swarms of smaller armed air- or sea-based platforms. Speed and mass may soon reverse the current well-established trend toward fewer advanced systems such as aircraft carriers or F-35 fighter aircraft. Reliance on autonomous platforms will increase the scope of cyberspace to include these systems and their networks both as potential targets for infiltration and as systems in need of cybersecurity.

These tactical dynamics affected by AI, while potentially game-changing on the conventional battlefield, may be even more significant in cyberspace. The speed of cyber engagements already overwhelms human operators, who rely on algorithm-driven automated defenses to cope with the sheer number of continuous threats. New generations of autonomous intelligent agents will be able to continually probe and attack defenses while another set of policing intelligent agents monitor networks and respond independently to possible intrusions.

The speed of these "engagements" will often be near-instantaneous and slower human decision-making will become a disadvantage, as will those who fail to pursue AI technology. Even if humans retain a primary decision-making role, the situational awareness leading to those decisions will increasingly rely on AI for network monitoring and analysis, offering humans a choice of pre-sorted response options.

## VERIFICATION OF BREACHES

Human reliance on digital networks for situational awareness in physical domains is hardly new, but awareness in cyberspace relies on machines not only for the individual pieces of information but also an analysis of this information. As autonomous systems become an increasingly integral part of mili-

tary operations, detection, verification, and attribution of network breaches will have heightened strategic implications. When AI agents monitor a network, humans will be required to trust that its analysis of the intruders is correct.

This trust may be tenuous, particularly when machine-learning algorithms based on neural networks are in essence a "black box" that so far have been unable to provide observable evidence of their decision-making processes. How are humans to know when their AI is functioning correctly and has not been infiltrated by an adversary?

The "black box" challenge of AI becomes a central issue, particularly when algorithms behave in unpredictable ways. Financial markets are now dominated by machine-trading algorithms, resulting in several inexplicable "flash crashes" such as those seen in May 2010 and February 2018. The new techniques uncovered by AlphaGo Zero were unanticipated by its human creators; other AI derived solutions (such as how to respond to a particular cyber attack) may elicit unanticipated solutions that are difficult for humans to evaluate prior to implementation.

Some researchers push back against this concern, citing the equally difficult task of understanding human decision-making and motivations. In the *New York Times*, Vijay Pande wonders if such criticisms are based not on the fact that "we can't 'see' AI's reasoning but that as AI gets more powerful, the human mind becomes the limiting factor. It's that in the future, we'll need AI to understand AI."

## THE FUTURE OF AI

AI has already altered cyberspace. The question is not whether the technology will influence the strategic dynamics in the domain, but rather how much. The progress made by artificial intelligence research over the past decade has wildly exceeded expectations and generated a wave of literature and speculation over future progress. It may be worth contemplating futurist Roy Amara's observation that "we tend to overestimate the effect of a technology in the short run and underestimate the effect in the long run."

Rodney Brooks, former director of the AI lab at the Massachusetts Institute of Technology, argued in the same 2017 piece that machine learning remains limited and inflexible or "brittle." Brooks notes that "we have seen a sudden increase in performance of AI systems thanks to the success of deep learning. Many people seem to think that means we will continue to see AI performance increase by equal multiples on a regular basis."

The long-term effects of AI may be dramatic, but they might in fact be more limited than we now assume. Perhaps the advances in deep learning – a decades-old concept just now coming to fruition – will now stagnate. As another author succinctly phrased it, "maybe we're not actually at the beginning of a revolution. Maybe we're at the end of one."

## SINGULARITY

At the other end of the spectrum are experts concerned about the very future of the human race. In 1993, Vernor Vinge philosophized in an article entitled "The Coming Technological Singularity: How to Survive in a Post-Human Era" about an age in which machine intelligence surpasses that of humans, an event scholars label "the singularity" (based on a more general term for a paradigm altering technological advance).

Given the exponential growth in computing power, the ability of AI software to analyze data at astounding rates, and the emerging ability of AI to write its own AI software, it seems plausible that machine intelligence may evolve to an artificial general intelligence or "superintelligence." Thought leaders ranging from Ray Kurtzweil and the late Stephen Hawking to Elon Musk have expressed concern for this development.

Philosopher Nick Bostrom argues that machine general intelligence may present an existential threat to humans and that "before the prospect of an intelligence explosion, we humans are like small children playing with a bomb."

## THE THIRD REVOLUTION IN WARFARE

The development of AI and its weaponization is proceeding far more rapidly than our understanding of the strategic, societal, and humanitarian implications of the technology. This evolution is creating new challenges for law, policy, and governance at domestic and international levels. In July 2015, over one thousand researchers, scholars and thought leaders published an open letter warning about the potential hazards of autonomous weapons, described as the third revolution in warfare, after gunpowder and nuclear arms.

The letter's authors argued that "artificial Intelligence (AI) technology has reached a point where the deployment of such systems is — practically if not legally — feasible within years, not decades." The letter concluded that a "key question for humanity today" was "whether to start a global AI arms race or to prevent" one, since a global AI arms race would almost inevitably result in autonomous weapons, which would not be beneficial to humankind. The consequences of machine superintelligence controlling cyberspace are impossible to predict, but unanticipated actions by machine-based intelligent agents are likely and not necessarily controllable by human operators.

Arms control measures for AI itself appear difficult to construct at this point, not least due to the dual-use nature of the technology and a host of serious verification issues. Machine intelligence in lethal autonomous weapons systems (LAWS), on the other hand, has seen some political mobilization for a ban on such systems.

## LEGAL FRAMEWORK

The most relevant legal framework for a LAWS ban falls under the purview of the United Nations Convention on Certain Conventional Weapons (CCW). In May 2014, the subject of lethal autonomous weapon systems was included in the annual CCW meeting in Geneva.

In addition to these discussions, individual countries such as Norway are evaluating various alternative options. In its 2015 report, the Council on Ethics for the Norwegian Government Pension Fund Global included a chapter evaluating the ethical and legal implications of autonomous weapon systems.

As is often the case with new technologies, civilian and military researchers are racing to develop new tools and methodologies. The doctrinal development evaluating how these new technologies might be used often lags behind, particularly in a security spiral dynamic where first mover advantage is usually considered desirable. Social science research and strategic planning that explores the possible uses and consequences of AI development and applications in cyberspace should at least attempt to keep pace with technological advances.

This report serves only as an overview and introduction to the fields of AI and cyber – many of the topics briefly addressed here could easily be more deeply and thoroughly explored. We need a better conceptual understanding of these fast-moving technological advances *while they are still under development*. The potential consequences of weaponized AI in cyberspace are too profound to wait until such systems are fully developed, and that understanding is currently woefully inadequate for policymakers to judge the potential consequences of AI for international security.

## SUGGESTIONS FOR FURTHER READING

**ALLEN, GREG AND TANIEL CHAN**
2017. *Artificial Intelligence and National Security*. Belfer Center for Science and International Affairs, Harvard Kennedy School of Government.

**BENDIEK, ANNEGRET AND TOBIAS METZGER**
2015. *Deterrence Theory in the Cyber-Century*, Working Paper RD EU/Europe 2015, Stiftung Wissenschaft und Politik.

**CLARKE, RICHARD AND ROBERT KNAKE**
2010. *Cyber War*. New York: HarperCollins.

**DEMCHAK, CHRIS AND PETER DOMBROWSKI,**
2011. "Rise of a Cybered Westphalian Age", *Strategic Studies Quarterly*, vol. 5:1, pp. 32–61.

**FISCHERKELLER, MICHAEL**
2017. "Incorporating Offensive Cyber Operations into Conventional Deterrence Strategies" *Survival*, vol. 59, no. 1.

**FUTURE OF HUMANITY INSTITUTE AND UNIVERSITY OF OXFORD ET.AL.**
2018. *The Malicious Use of Artificial Intelligence: Forecasting, Prevention and Mitigation*.

**HANSBØ, MORTEN**
2017. «Robotikk, kampkraft og bærekraft i framtidens Forsvar», *Norsk Militært Tidsskrift* vol. 187:3, pp. 4–13.

**HARRIS, SHANE**
2014. *@war*. Boston: Mariner Books.

**ILACHINSKI, ANDREW**
2017. *AI, Robots, and Swarms: Issues, Questions, and Recommended Studies*. Alexandria, VA: CNA.

**KANIA, ELSA B.**
2017. *Battlefield Singularity: Artificial Intelligence, Military Revolution, and China's Future Military Power*. Washington DC: Center for a New American Security.

**KRAMER, FRANKLIN D., STUART H. STARR AND LARRY WENTZ (EDS.)**
2009. *Cyberpower and National Security*. Washington DC: Potomac Books.

**LEWIS-KRAUS, GIDEON**
2016. "The Great AI Awakening", *New York Times*, 14 December.

**LIBICKI, MARTIN C.**
2009. *Cyberdeterrence and Cyberwar*. Santa Monica, CA: RAND Corp.

**LIBICKI, MARTIN**
2013. *Brandishing Cyberattack Capabilities*. Santa Monica, CA: RAND Corp.

**MULLER, LILLY PIJNENBURG, LARS GJESVIK AND KARSTEN FRIIS**
2018. *Cyber-weapons in International Politics*. Norwegian Institute of International Affairs.

**NILSSON, NILS**
2009. *The Quest for Artificial Intelligence*. Cambridge: Cambridge University Press.

**NYE, JOSEPH S. JR.**
2016/2017. "Deterrence and Dissuasion in Cyberspace" *International Security*, vol. 41:3, pp. 44–71.

**RID, THOMAS AND PETER MCBURNEY**
2012. "Cyber-weapons", *RUSI Journal*, vol. 157: 1, pp. 6–13.

**SANGER, DAVID**
2012. *Confront and Conceal*. New York: Crown Publishers.

**SCHARRE, PAUL.**
2018. *Army of None: Autonomous Weapons and the Future of War*. New York: W. W. Norton & Co.

**SINGER, P.W.**
2009. *Wired for War*. New York: Penguin Books.

**SINGER, P.W. AND ALAN FRIEDMAN**
2014. *Cybersecurity and Cyberwar: What everyone needs to know*. Oxford: Oxford University Press.

**STANFORD UNIVERSITY**
2016. *Artificial Intelligence and Life in 2030: One Hundred Year Study on Artificial Intelligence*, Report of the 2015 Panel.

**TYUGU, ENN**
2011. "Artificial Intelligence in Cyber Defense," Paper presented at the 3rd International Conference on Cyber Conflict, Tallinn, Estonia, 7–10 June.

**WIRKUTTIS, NADINE AND HADAS KLEIN**
2017. "Artificial Intelligence in Cybersecurity", *Cyber, Intelligence, and Security* vol. 1:1.

**ZETTER, KIM**
2014. *Countdown to Zero Day*. New York: Broadway Books.

## IFS INSIGHTS

**IFS Insights** aims to provide a flexible online forum for articles, comments and working papers within the fields of activity of the Norwegian Institute for Defence Studies. All views, assessments and conclusions are the author's own. The author's permission is required for any reproduction, wholly or in part, of the contents.

Editor: Anna Therese Klingstedt

## INSTITUTT FOR FORSVARSSTUDIER

**The Norwegian Institute for Defence Studies** (IFS) is a part of the Norwegian Defence University College (FHS). As an independent university college, FHS conducts its professional activities in accordance with recognised scientific, pedagogical and ethical principles (pursuant to the Act pertaining to Universities and University Colleges, section 1-5).

Director: Kjell Inge Bjerga

Norwegian Institute for Defence Studies
Kongens gate 4
P.O. Box 890 Sentrum
N-0104 OSLO
Email: info@ifs.mil.no
ifs.forsvaret.no/en

Top picture on the front page:
Graphic design: U.S. Army illustration.

## ABOUT THE AUTHOR

**Michael Mayer** is a Senior Fellow at the Norwegian Institute for Defence Studies. He lectures and conducts research on military technology and international security, with a particular focus on American foreign and security policy.

Among his previous publications include "Strategic Uncertainty and Missile Defence: Revisiting the 1999 National Intelligence Estimate" in *Contemporary Security Policy*, December 2015, and "The New Killer Drones: understanding the strategic implications of next-generation unmanned combat aerial vehicles" in *International Affairs*, July 2015.

## CENTRE FOR TRANSATLANTIC STUDIES

Despite the changing international balance of power and America's global footprint, transatlantic relations, now as before, remain of particular importance to Norway. This is why it is so important to understand US strategic thinking and defence policy priorities.

The Norwegian Institute for Defence Studies has a long tradition of research into transatlantic politics. The aim of the Centre for Transatlantic Studies is to deliver in-depth analyses of selected but important aspects of US conduct on the international stage, applying both historical and political science approaches and concentrating on relations with Europe.

**MORE ABOUT CENTRE FOR TRANSATLANTIC STUDIES**

https://forsvaret.no/ifs/en/Research/Transatlantic-studies