Cyber and Artificial Intelligence—Technological Trends and National Challenges

Liran Antebi and Gil Baram

Autonomous systems based on artificial intelligence are playing an increasingly meaningful role in everyday life in a variety of fields, including industry, medicine, the economy, and security. Because they are computerized, these systems are exposed to coding errors, which may lead to incorrect decision making and the execution of unwanted actions. In addition, they are vulnerable to cyberattacks that may harm or completely suspend their activity. This article examines the risks posed to autonomous systems as a component of the arms race among the powers and discusses policy steps to contend with these threats at the national level.

Keywords: Artificial intelligence, cyber warfare, national security, arms race

Introduction

"Artificial intelligence is the future, not only for Russia, but for all humankind . . . It comes with colossal opportunities, but also threats that are difficult to predict. Whoever becomes the leader in this sphere will become the ruler of the world." These were the words of Russia's President Vladimir Putin in

Liran Antebi is a research fellow at INSS, where she directs the research field on advanced technology. She lectures at Ben Gurion University and advises in the field of advanced technologies. Gil Baram is the head of research at the Yuval Ne'eman Workshop for Science, Technology and Security and a research fellow in the Blavatnik Interdisciplinary Cyber Research Center at Tel Aviv University.

a September 2017 lecture.[1] And indeed it seems that autonomous systems based on artificial intelligence (AI) are becoming increasingly ubiquitous in a variety of fields, including industry, medicine, the economy, and security. As computerized systems, they are vulnerable to coding errors, which may lead to incorrect decision making and the execution of unwanted actions. Additionally, they are vulnerable to cyberattacks that may harm or completely suspend their activity. At the same time, systems with some autonomous abilities are increasingly being used; these systems do require some human involvement in decision making for their operation, but both their calculation and recommendation processes are autonomous and generally not explainable.

This article examines the risks to autonomous systems and ways to contend with them at the national level. The first part surveys the uses of AI in the security realm. It describes the arms race taking place in this field, its influence on the international arena, and the incentives for carrying out a cyberattack on these systems. The second part describes potential cyberattacks on AI-based systems—the attacks and manipulations that are unique to cyber systems—and reviews possible uses of AI for both defensive and offensive purposes in cyber warfare. Finally, the article suggests policy steps aimed at reducing the risks that are increasing as the use of autonomous systems expands and human dependence upon them grows.

Artificial Intelligence and Autonomous Systems—Development and Key Uses

AI is a subdivision of computer science research that has existed since the 1950s. One of the simplest and most widespread definitions of AI is "the science of making machines do things that would require intelligence if done by men."[2] Over the past decade, significant advances have been made in the field of AI, partly due to advances in computer science research, development of advanced hardware and software in the fields of computing and communications, and the development of cloud computing and big data. Within this framework, subsets such as machine learning and deep neural networks also have evolved, enabling various advanced applications

[1] "Whoever Leads in AI Will Rule the World': Putin to Russian Children on Knowledge Day," *RT World News,* September 1, 2017, https://www.rt.com/news/401731-ai-rule-world-putin.

[2] Edward Geist and Andrew Lohn, *How Might Artificial Intelligence Affect the Risk of Nuclear War?* (Santa Monica: RAND Corporation, 2018), 9.

in different fields. These include image analysis applications, which are used in the medical world to help analyze various tests; speech recognition applications, which enable the operation of "smart assistants," such as Siri and Alexa; and many predictive algorithms, which offer people online products or services similar to those they have previously purchased or in which they have shown interest.

The Defense Advanced Research Projects Agency (DARPA) at the US Department of Defense defines AI as "programmed ability to process information."[3] Despite this definition, it is important to clarify that not all computing systems use AI. AI algorithms are designed to make decisions and typically do so using real-time data. These are not passive machines capable only of mechanical or predetermined responses—to which we have become accustomed in the age of automation—such as automatic doors. Rather, they are machines capable of integrating information from different sources, including sensors, digital data and even remote inputs, analyzing this information immediately and acting in accordance with conclusions derived from this data. This allows the processing of data at levels of sophistication and speed that did not previously exist.[4]

The most common uses of AI today are in the subset known as machine learning. This subset uses statistical algorithms to imitate human cognitive tasks, by inferring rules about these tasks based on analysis of large quantities of data on a given subject. In practice, the algorithm "trained" on existing data, and through this process creates a statistical model of its own, which will later be able to carry out the same task using new data it had not previously encountered.[5]

The use of AI technology is increasing, and many countries, companies, and security agencies now rely on these systems for various purposes. Civilian uses of AI include services such as navigation apps, algorithms offering targeted goods or services, banking and financial commerce, maintenance

[3] John Launchbury, "A DARPA Perspective on Artificial Intelligence," *TechnicaCuriosa*, 2017, https://machinelearning.technicacuriosa.com/2017/03/19/a-darpa-perspective-on-artificial-intelligence.

[4] Darrell M. West and John R. Allen, "How Artificial Intelligence Is Transforming the World," *Brookings,* 2018, https://www.brookings.edu/research/how-artificial-intelligence-is-transforming-the-world.

[5] Kelley M. Sayler and Daniel S. Hoadley, *Artificial Intelligence and National Security* (Congressional Research Service, 2019), 2.

and logistics systems, and more. As already mentioned, AI-based systems are used in the following national security fields:

1. **Intelligence**: AI has many uses in the field of intelligence. Today, machine learning and other algorithms are commonly used for image and text analysis. Algorithms are also used for language translation, video and audio analysis, and more. One of the best-known projects in this field is the Maven project, which was a collaboration between Google and the US Department of Defense; it used AI to analyze UAV's photography.[6] China and other countries are also working on creating systems for optimal categorizing of intelligence content and merging information from different systems in order to produce civil and military intelligence, using AI capabilities.[7]
2. **Logistics**: There are AI applications for military logistical use just as there are for civilian use. The US military has been using such systems since the 1990s; one AI system was used to plan and optimize troop movements during the first Gulf War, which enabled savings and a return of thirty years of investments in AI research.[8] Among the most innovative systems are those that assist in system maintenance in ways that were not previously possible, such as by reporting in advance about future wear-and-tear of parts and making it possible to replace them on an individual basis rather than on the basis of generalized statistical information as in the past. This system makes it possible to save significantly while increasing safety.[9]
3. **Command and control**: Command and control systems will make increasing use of AI, including as advisory frameworks that assist in decision making, while being subject to and in cooperation with human operators.

[6] Samuel Gibbs, "Google's AI Is Being Used by US Military Drone Programme," *The Guardian*, May 7, 2018, https://www.theguardian.com/technology/2018/mar/07/google-ai-us-department-of-defense-military-drone-project-maven-tensorflow. It should be noted that this project generated opposition among Google employees, due to fears that the knowledge it created would not only be used for analyzing intelligence material but also for creating autonomous weapons systems that would be able to attack without human involvement.

[7] For more on this topic, see Stephen Chen, "Inside the AI Revolution that's Reshaping Chinese Society," *South China Morning Post*, June 29, 2017, https://www.scmp.com/news/china/society/article/2100427/chinas-ai-revolution-and-how-its-rivalling-us.

[8] Nurit Cohen-Inger and Gal Kaminka, "And the Forecast: The IDF on the Way to an Intelligent Military—A Road Map for Adopting Artificial Intelligence Technologies in the IDF," *Bein Haktavim* 18 (2018): 95 [Hebrew].

[9] Sayler and Hoadley, *Artificial Intelligence and National Security*, 9.

4. **Autonomous vehicles**: Autonomous driving is commonly associated with driverless vehicles, archetypes now seen on the roads in various places around the world. This was, in fact, one of the central issues that DARPA dealt with over the past decade, enabling significant progress in this field.[10] In civilian life, the primary use of autonomous vehicles is on the ground. In the past several decades, unmanned vehicles used for security purposes and with a variety of autonomous capacities have been developed for air, sea, and land. These vehicles play a significant role on the battlefield, and they can enhance or replace human presence in dangerous areas; however, most of these vehicles rely primarily on human operation and intervention, despite their autonomous capacities.
5. **Autonomous military systems**: This is one of the most widespread fields of AI. Many countries, led by the United States, Israel, the United Kingdom, and France, have identified the security potential of unmanned systems over the past few decades and took steps to purchase and develop independently their own autonomous military systems. Autonomous military systems include a sub-group of autonomous weapons systems that can search for, identify, and attack targets independently without human input.[11] these are a game-changer systems, because they can cause fatal damage, without human involvement. These systems are subject of widespread public debate, and in the United Nations there is already a discussion regarding possible limitations on their use; nevertheless, they are being developed at an accelerated pace today, to the point that some fear that we are on the brink of a new arms race in this field,[12] or even at its peak. Although this field is still in its infancy, multiple countries have already acquired battlefield experience with these systems. These include air defense systems, such as the American Patriot system and Israel's Iron Dome. These systems are capable of being highly autonomous and even can operate completely autonomously; however, due to decisions of the countries that operate them, these systems are still dependent upon

[10] "The Grand Challenge for Autonomous Vehicles," *DARPA*, 2019, https://www.darpa.mil/about-us/timeline/-grand-challenge-for-autonomous-vehicles.

[11] "Autonomous Weapon Systems—Q & A," *International Committee of the Red Cross*, November 12, 2014, https://www.icrc.org/en/document/autonomous-weapon-systems-challenge-human-control-over-use-force.

[12] Billy Perrigo, "A Global Arms Race for Killer Robots Is Transforming the Battlefield," *Time*, April 9, 2018, http://time.com/5230567/killer-robots.

human operators who are a part of their operation cycle.[13] In addition to air defense systems, there are also loitering munitions such as Harop. This is an airborne system that is capable of flying, hovering, locating, tracking, and attacking targets by means of homing via radar signal.[14] Current research indicates that fully autonomous vehicles will become technologically possible within twenty years, and it is highly likely that they will become more significant in the activity of modern militaries.[15]

6. **Cyber warfare**: This is one of the leading fields in the use of AI. In this field, "first-generation" AI is still in use,[16] while later-generation capabilities are being developed. Algorithms assist in preventing cyberattacks, or in locating attacks on various computerized systems. At the same time, the cyberattackers use AI capabilities in various ways as we further discuss.

The Artificial Intelligence Arms Race

In recent years, many countries have identified the potential impact of AI on their economies and on national security.[17] The United States is considered the leader in this field and is working to formulate a comprehensive strategy on the matter. Its national defense strategy includes a commitment by the US Department of Defense to invest in military implementation of autonomous technologies, AI, and machine learning, while also using groundbreaking

[13] Human Rights Watch, *Losing Humanity: The Case Against Killer Robots* (November 2012), 11–12, http://www.hrw.org/sites/default/files/reports/arms1112ForUpload_0_0.pdf.

[14] Dan Gettinger and Arthur Holland Michel, *Loitering Munitions* (Center for the Study of the Drone at Bard College, 2017), http://dronecenter.bard.edu/files/2017/02/CSD-Loitering-Munitions.pdf.

[15] Yoav Zacks and Liran Antebi, eds., *The Use of Unmanned Military Vehicles in 2033: National Policy Recommendations Based on Technology Forecasting—Expert Assessments*, Memorandum no. 145 (Tel Aviv: INSS, December 2014) [Hebrew]; Paul Scharre, *Robotics on the Battlefield Part I: Range, Persistence and Daring* (Washington DC: Center for a New American Security, May 2014), https://s3.amazonaws.com/files.cnas.org/documents/CNAS_RoboticsOnTheBattlefield_Scharre.pdf?mtime=20160906081925; Launchbury, "A DARPA Perspective on Artificial Intelligence."

[16] Launchbury, "A DARPA Perspective on Artificial Intelligence."

[17] At the time of this writing, a type of arms race is taking place among the powers in developing advanced AI capabilities. In parallel, some are claiming that the discussion should be about combining competition and collaboration, and not about the "arms race," which has a negative connotation, and instead call upon the United States and China to commence a dialogue that would lead to collaboration in developing AI. See Tim Hwang and Alex Pascal, "Artificial Intelligence Isn't an Arms Race," *Foreign Policy,* December 11, 2019, https://foreignpolicy.com/2019/12/11/artificial-intelligence-ai-not-arms-race-china-united-states/.

commercial technologies, with the aim of maintaining the US military's competitive advantage in this field.[18]

In early 2019, the White House updated the national research and development strategy for AI technologies that the Obama administration had published in 2016.[19] The updated strategy calls for developing effective methods for human-AI collaboration and ensuring that AI systems are well protected. The United States invests large sums in this area and is working to lay out a broad strategy for promoting and defending AI technologies on a national level, via collaboration between the government, the private sector, academia, the public, and international partnerships.[20]

In July 2019, the Joint Artificial Intelligence Center (JAIC) called private companies to submit ideas and proposals for AI technologies for cyber defense, which would include automatically correcting weaknesses in military network-security collecting cyber intelligence about those active on the dark web, and more.[21] Despite all its efforts, the United States' main challenge is the increased competition with China, given its aspirations to become the leader in AI within the next decade.

As a serious competitor in AI, China has already proven it can make rapid progress on advanced technological projects, such as by becoming a major manufacturer and exporter of unmanned aerial vehicles (UAVs) within a decade. The total sum China has invested in AI research and development is unknown to the public, but it is estimated at billions of dollars at a minimum. Some estimates say that planned future investments will reach $150 billion.[22] This investment is partly due to China's prominent advantages in this field, which is the almost total lack of distinction or boundaries between civilian and military uses, given that its civilian life is also subject to strict government supervision.

[18] Department of Defense, *Summary of 2018 National Defense Strategy of The United States of America* (Washington DC, 2018), 5.

[19] Aaron Boyd**,** "White House Updates National Artificial Intelligence Strategy," *Defense One*, June 22, 2019, http://bit.ly/2ZYY2U4.

[20] White House and Office and Science and Technology Policy, "Artificial Intelligence for the American People," *The White House*, 2019, https://www.whitehouse.gov/ai/executive-order-ai.

[21] "DoD's JAIC to Call for Private Sector Cyber Tech Pitches," *MeriTalk*, July 8, 2019, https://www.meritalk.com/articles/dods-jaic-to-call-for-private-sector-cyber-tech-pitches/.

[22] "DoD's JAIC to Call for Private Sector Cyber Tech Pitches."

Another prominent Chinese advantage is due to its nonadherence to Western norms of democracy, individual rights, and privacy. It has thus collected and coded information about its citizens for many years. This process has rendered China an enormous mine of big data, leading companies and entities from around the world to work with it in order to get access to this information. China notably also collects information about citizens of other countries by perpetrating cyberattacks and theft of vast information reserves as well as through Chinese-made systems and applications used by citizens of other countries. Legislation also allows Chinese government agencies to insert "backdoors" at the assembly line of all Chinese manufacturers. This same legislation obligates Chinese tech manufacturers to give the government their technologies' source code.[23]

Some have assessed that China will become the most dominant country in the field of AI in the future. In November 2017, Eric Schmidt, then the chairman of Google, stated that China would equal the United States in its AI capabilities by 2020 and would surpass it by 2025.[24] Current assessments support Schmidt's prediction. In terms of research, Chinese researchers are expected to publish an equal number of academic papers on AI to that of their American peers, indicating the growing significance of the subject in China.[25]

In addition to China and the United States, Russia is also administering AI programs, and in 2019, it formulated a national AI strategy.[26] Russia, however, lags behind both the United States and China: In addition to low investments in this field relative to its principal competitors, it also suffers from problems in its tech ecosystem.[27] Due to these conditions, analysts believe that Russia will only emerge as the leader in certain narrow subfields of AI and not in the field as a whole.[28]

[23] The authors wish to thank Dr. Harel Minshari, the director of cyber studies at the Holon Institute of Technology, for his helpful comments on this issue.

[24] Sam Shead, "Eric Schmidt on AI: 'Trust Me, These Chinese People Are Good,'" *Business Insider*, November 1, 2017, https://www.businessinsider.com/eric-schmidt-on-artificial-intelligence-china-2017-11.

[25] Tom Simontie, "China is Catching up to the US in AI Research—FAST," *Wired*, March 13, 2019, https://www.wired.com/story/china-catching-up-us-in-ai-research/.

[26] Samuel Bendett, "Putin Orders Up a National AI Strategy," *Defense One*, 2019, https://www.defenseone.com/technology/2019/01/putin-orders-national-ai-strategy/154555/.

[27] Bendett, "Putin Orders Up a National AI Strategy."

[28] Andrew P. Hunter and others, *Artificial Intelligence and National Security: The Importance of an AI Ecosystem* (Washington, DC: CSIS, 2018), 48, https://www.csis.org/analysis/artificial-intelligence-and-national-security-importance-ai-ecosystem.

Israel, known as a worldwide tech leader, particularly in cyber and unmanned aerial vehicles, is one of many other countries competing in AI. Israel does not currently have a defined strategy for AI, although a commission appointed by the prime minister is carrying out comprehensive research on the issue, and its conclusions and recommendations will be used to formulate strategy and policy. An AI headquarters may also be established.[29] Israel has a significant advantage in its unique ecosystem, which includes close connections between the government, academia, industry, and the military, as well as the ability to respond rapidly to changes in the arena. Israel also has the advantage of significant knowledge transmission between the military and civilian companies in the industry, as a result of its unique model of mandatory military service and reserve duty. This model creates the opportunity for some workers to acquire and transmit knowledge between security agencies and the tech industry in an ongoing and productive manner.[30]

Many companies in Israel are working in AI. AI is the heart of some of the companies, and it is an enabling technology or a force multiplier for others. International companies, including Amazon, Intel, Microsoft, and Invidia have established R&D centers in Israel that focus on AI.[31] Israel also has developed leading AI companies, which develop both software and hardware.[32] In the security field, Israel also develops AI technology in the framework of the Ministry of Defense, the Directorate of Arms and Infrastructure Development, and various technological military units, as well as in its security industries.

The international AI arms race, the increasing presence of these systems, and our reliance on them in different areas necessitate discussing the threats posed to AI systems and ways of locating, identifying, and preventing or thwarting these threats.

[29] "The Science Committee: First Discussion on Government Readiness in the AI Field," *Knesset News,* The Knesset, June 4, 2018, https://m.knesset.gov.il/News/PressReleases/pages/press04.06.18ec.aspx [Hebrew].

[30] Dafna Gatz and others, *Artificial Intelligence, Data Science and Smart Robotics, First Report* (Shmuel Neeman Institute for National Security Research, 2018) [Hebrew].

[31] Amir Mizroch, "In Israel, A Stand Out Year for Artificial Intelligence Technologies," *Forbes*, March 11, 2019, https://www.forbes.com/sites/startupnationcentral/2019/03/11/in-israel-a-stand-out-year-for-artificial-intelligence-technologies/#13acbc7530a8.

[32] For example, see the Israeli companies Mellanox and Habana Labs, which were sold to international companies for billions of dollars. For more on this topic, see Sagi Cohen, "Exit Warning: Tech Giants Fight over the Future of Computerization," *TheMarker*, December 18, 2019, https://www.themarker.com/technation/.premium-1.8285726 [Hebrew].

Cyberattacks and AI System Manipulation

The increase in cyber threats over the past few years is a threat to AI systems. At the same time, it raises fears that AI technology will be exploited in order to carry out cyberattacks on a much wider scale than previously possible. The risk is even greater for security systems that are not completely disconnected from the network and given the increasing military use of AI technology.

Although AI technology is considered inseparable from the possibility of cyberattacks, it can also be an effective tool for more effective management of cyberattacks, such as by using deep learning techniques that are capable of tracking suspicious activity and classifying different viruses. At the same time AI systems are vulnerable to cyberattacks against them and are likely to be subject to different manipulations. Autonomous and AI systems are computerized systems and therefore are exposed, like other systems, to the kind of cyberattacks with which we are familiar on regular computerized systems. Due to their unique nature, however, autonomous and AI systems are also vulnerable to unique attacks for the following reasons:

1. The desire to allow them to function autonomously, without human involvement, due to considerations of efficiency, accuracy, and speed, may leave them vulnerable to cyberattacks. However, it can be surmised that these systems will inform their operator of any anomalies or attacks that they identify.
2. Some AI-based systems operate today in ways that we do not know how to explain or analyze retroactively. This is referred to as the “explainability challenge” or the “black box” challenge. This leaves an opening for attacks, which in some cases would be difficult to identify, because it is not clear whether it is an attack or the proper functioning of the system.
3. The processes of the training of AI systems, which are carried out using an enormous quantity of data, make it possible to introduce data that could deliberately “infect” the process and lead to incorrect or undesired results.

A few recent examples illustrate these threats. In April 2019, Col. Stoney Trent, the head of the Operations Department of the JAIC in the US Department of Defense, said that the problems with assessing cyber threats against AI technologies stem from lack of awareness among decision makers and from a dearth of tools and methods for examining the immunity of AI systems to hacking. According to Trent, one of the JAIC’s tasks is to encourage the development of these tools, which civilian and commercial developers do not perceive as worthwhile.[33]

According to former research director at the National Security Agency (NSA), Frederick Chang, the race to develop military technologies based on AI will significantly increase the scope of the attack surface, but governments are still not aware of most of the vulnerabilities of these systems. Chang warned that attackers may mislead a system's identification mechanism by using adversarial inputs, poison the data from which the system learns, or infiltrate it in order to understand how it operates and to thwart its functioning.[34]

In addition, the combination of cyber warfare and AI technology may lead to the development of new kinds of malware. For example, IBM researchers developed DeepLocker, an AI-based malware that aimed to understand how to combine multiple existing models of AI in order to create a new and more effective form of malware that has not yet been encountered. This malware disguises its aim until it reaches its target, which it identifies using voice or facial recognition. This kind of malware is considered especially effective because it may infect millions of systems without being discovered, unlike cyberattacks, which are sometimes widespread and use a noisy "spray and pray" approach.

Given that autonomous systems are meant to function without human input or even with minimal human intervention, an effective manipulation or attack may not be discovered for a long time. In contrast to existing malware, a malware that incorporates AI will require significant expertise and advanced forensic tools in order to identify it. DeepLocker changed the rules of the game by hiding its activity in common applications, such as for videoconferencing. Its use of AI is almost impossible to detect or to reverse-engineer in order to discover its code. DeepLocker will only begin to function if it identifies its chosen target, and it will do so via use of the deep neural network model of AI. This model will only begin to work if it identifies a specific input or when it identifies its chosen target.[35]

[33] Theresa Hitchens, "Rush to Military AI Raises Cyber Threats," *Breaking Defense*, April 25, 2019, https://breakingdefense.com/2019/04/rush-to-military-ai-raises-cyber-threats.

[34] Hitchens, "Rush to Military AI Raises Cyber Threats."

[35] Marc P. Stoecklin, "DeepLocker: How AI Can Power a Stealthy New Breed of Malware," *SecurityIntelligence*, August 8, 2018, https://securityintelligence.com/deeplocker-how-ai-can-power-a-stealthy-new-breed-of-malware/.

Researchers estimate that AI systems will allow humans to carry out cyberattacks that were not possible prior to the use of these systems. They will also be able to identify new sources for attacks on AI systems by identifying new weak points. For example, a study published in 2017 showed that researchers used AI tools in order to decipher the passwords of LinkedIn users. In a sample set of 43 million user profiles, researchers successfully figured out 27 percent of the passwords.[36] Attacks on the systems responsible for autonomous tools is another possible scenario and would likely create widespread disruption and affect multiple tools.[37]

Three major attacks on AI systems can be demarcated: (1) Inputting false data into a system, so that it will generate false conclusions; (2) minor alterations to photographs or other inputs that the system processes, whether by inserting visible items or by changing the pixels of a photograph, so that the item will be classified incorrectly; (3) disrupting the assessment of information by internally damaging the system's sorting mechanism, instead of focusing on the particular data that was fed into the system.[38]

The following are several types of unique attacks against AI systems:

- **Adversarial attacks**: This is a technique for misleading AI systems' machine learning classifier by exploiting their vulnerability to the manipulation of the data they are fed and which they use to train themselves. In this way, the attackers create an input that appears to have a misleading classification and thus "interrupts" the information fed into the system in order to cause misclassification. The changes are almost invisible to the human eye. One study found that deep neural networks can easily be fooled by the input of false data.[39]

[36] Matthew Hutson, "Artificial Intelligence Just Made Guessing Your Password a Whole Lot Easier," *Science*, September 15, 2017, https://www.sciencemag.org/news/2017/09/artificial-intelligence-just-made-guessing-your-password-whole-lot-easier.

[37] Allan Dafoe, *AI Governance: A Research Agenda* (Future of Humanity Institute and University of Oxford, 2017), 5, https://www.fhi.ox.ac.uk/wp-content/uploads/GovAIAgenda.pdf; Miles Brundage and others, *The Malicious Use of Artificial Intelligence: Forecasting, Prevention and Mitigation* (Future of Humanity Institute and University of Oxford, 2018), 20, https://img1.wsimg.com/blobby/go/3d82daa4-97fe-4096-9c6b-376b92c619de/downloads/MaliciousUseofAI.pdf?ver=1553030594217.

[38] Jian hua Li, "Cyber Security Meets Artificial Intelligence: A Survey," *Frontiers of Information Technology and Electronic Engineering* 19, no. 12 (2018): 1462–1474, https://doi.org/10.1631/FITEE.1800573.

[39] Mesut Ozdag, "Adversarial Attacks and Defenses against Deep Neural Networks: A Survey," *Procedia Computer Science* 140 (2018): 152–161, https://doi.org/10.1016/j.procs.2018.10.315.

- **Data poisoning**: This is a technique in which the attacker inputs false data and systematically disrupts the data inputs used for training the system. To accomplish this, the attacker must have access to the data used to train the model. This data may be disrupted in order to benefit the attackers or to harm other groups—for example, in models which calculate insurance premiums or grant loans.[40]
- **Evasion attacks**: These are attacks in which the attacker manipulates a model's classification ability with the aim of evading detection. This type of attack intends to evade spam filters, malicious password detectors, network traffic monitors, and anomaly detectors.[41]
- **Model extraction**: These are attacks in which the attacker sends samples of data to the system model and analyzes its output in order to build the model on his own.[42]
- **Attacks on watermark tags**: Watermarking refer to when the attacker adds specific pixels to a picture in order to cause a model to react in a certain way.[43] An effective attack of this kind an intelligence or weapons system may be very problematic.

Systems based on machine learning sometimes contain sensitive information, such as facial recognition systems, and can be the target of a cyberattack in which attackers can take information about the people identified by the system. These attacks can be carried out in two different stages of the development process of the system: in the system training stage and in the stage of testing and drawing conclusions from system operation.

[40] Jacob Steinhardt, Pang Wei Koh, and Percy Liang, "Certified Defenses for Data Poisoning Attacks," *Advances in Neural Information Processing Systems*, no. i (December 2017): 3518–3530 ; Patrick Hall, "Proposals for Model Vulnerability and Security," *O'Reilly Media*, 2019, https://www.oreilly.com/ideas/proposals-for-model-vulnerability-and-security.

[41] Battista Biggio and others, "Evasion Attacks against Machine Learning at Test Time," *Lecture Notes in Computer Science*, part 3 (2013): 387–402, https://doi.org/10.1007/978-3-642-40994-3_25; Erwin Quiring and Konrad Rieck, "Adversarial Machine Learning against Digital Watermarking," *European Signal Processing Conference* (September 2018): 519–523, https://doi.org/10.23919/EUSIPCO.2018.8553343.

[42] Florian Tramèr and others, "Stealing Machine Learning Models via Prediction APIs," (Cornell University, October 2016), http://arxiv.org/abs/1609.02943.

[43] Romain Artru, Alexandre Gouaillard, and Touraj Ebrahimi, "Digital Watermarking of Video Streams: Review of the State-Of-The-Art," August 2019, https://arxiv.org/abs/1908.02039.

However, there are also various methods of defense in this field, which can be built for both stages.[44]

During the training stage, a data poisoning attack can be carried out, which would include inputting false information into the system and changing the data markings. Methods to protect data at this stage include filtering false data that enters the system (data sanitization), acting on potential attack scenarios in order to learn about likely actions by the opponent (adversarial training), refining methods of defense (defense distillation), combining methods (ensemble methods), and strengthening privacy by different means (differential privacy).

Different types of attacks can also be carried out during the testing stage. The first is evasion, which uses attack methods capable of evading detection by the system. A second method is impersonation, which allows hostile entities to imitate legitimate ones in order to enter the system and disrupt its data. A third method is inversion, which allows the theft of sensitive data from the system. The system can be defended from these kinds of attacks by using noise, such as a smokescreen protecting data, or through the random use of protective measures during training (differential privacy).[45]

As shown above, various attacks on autonomous systems and AI systems are possible, some of which are unique to these systems and differ from "regular" attacks on computerized systems. It is important to distinguish between the types of cyberattacks relevant to computer systems in general and those unique to the AI and autonomy systems, as we have sought to demonstrate in this article.

The risk from cyberattacks on AI systems is significantly greater, because it is usually impossible for any person to detect the problem within a short time frame, due to the system's characteristics. The "black box" or "explainability challenge" already mentioned is another. This challenge relates to the fact that in spite of the successful results of different system actions of machine learning or deep learning, it is currently impossible to

[44] Qiang Liu and others, "A Survey on Security Threats and Defensive Techniques of Machine Learning: A Data Driven View," *IEEE Access* 6 (2018): 12103–12117, https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8290925.

[45] Liu and others, "A Survey on Security Threats and Defensive Techniques of Machine Learning."

explain the way that a system arrives at its result.[46] The lack of transparency makes it hard to verify system activities in general. This, together with the enormous quantity of data and the pace at which these systems process it, mean that human oversight of these systems is likely to be merely for show. Experts and military figures say that humans should be held accountable for the actions of AI systems, but this claim requires additional discussion, given that humans cannot locate and identify all risks and vulnerabilities present in these systems.[47]

Attention and effort should be invested in this issue from the research and development stage ahead, in order to try to build effective human oversight mechanisms. Efforts to solve the "explainability challenge" have been going on for some time,[48] but until an adequate technological solution is found for this matter, regulatory and legal mechanisms are needed for managing the lack of transparency. This is particularly important for strategic systems, or those whose outcome may harm human beings.

Conclusion and Recommendations

The use of AI systems, including systems based on machine learning and deep learning, is becoming increasingly common in many fields, including security. These systems work at a quick pace, often making human oversight difficult. In addition, enabling these systems to run autonomously in order to reduce the human involvement necessary is desired for a variety of reasons. It is important to understand the potential of cyberattacks on these systems and to develop means of thwarting them. The unique attacks to which these types of systems are vulnerable must also be understood, in order to develop effective oversight and defense mechanisms that will allow proper functioning of these systems and trust in them. In order to achieve all of this, action should be taken in a few policy directions.

[46] Richard Gall, "Machine Learning Explainability vs Interpretability: Two Concepts That Could Help Restore Trust in AI," *KDnuggets*, 2018, https://www.kdnuggets.com/2018/12/machine-learning-explainability-interpretability-ai.html.

[47] Connor McLemore and Charles Clark, "The Devil You Know: Trust in Military Applications of Artificial Intelligence," *War on the Rocks*, September 23, 2019, https://warontherocks.com/2019/09/the-devil-you-know-trust-in-military-applications-of-artificial-intelligence/.

[48] Zelros AI, "A Brief History of Machine Learning Models Explainability," *Medium*, September 24, 2018, https://medium.com/@Zelros/a-brief-history-of-machine-learning-models-explainability-f1c3301be9dc.

Israel, which is a power in the field of cyber and one of the world's leaders in AI,[49] could potentially become a leader in the field of cyber research and defense as it relates to AI and autonomous systems. Thus, several policy recommendations for Israel are listed below, which may be relevant for other countries as well:

1. Standards should be defined for the field of AI-based systems. In this framework, these systems should have embedded means of oversight or methods of verifying that they have not been attacked or manipulated. The standards defined must apply not only to security systems but also to critical civilian systems (and ideally also to non-critical systems, which assist in maintaining routine.) Likewise, the government should fund system defense in fields in which there is a market failure and no commercial incentive for solving a given issue.
2. The relevant agencies in the defense system must invest in research for ongoing mapping and locating of fields with high vulnerability potential and the risk of attacks that specifically target autonomous systems. Investment in developing specific solutions for this field is also an imperative.
3. Relevant state bodies must define procedures for supervising AI-based cyber operations, in order to avoid unwanted consequences of such operations. These procedures must be backed up by effective means of enforcement.[50]
4. Joint exercises should be conducted with allies, in which the defense capabilities of AI are tested. The exercises will expose weaknesses that should be corrected and give these systems information from which they can learn.
5. The international discourse on this issue should be expanded, in order to create collaborations between like-minded countries that are AI-power players and share common interests. The aim is to influence the entire international arena, given the difficulty rapidly affecting international institutions such as the United Nations. An international charter should also be formulated for this field. The importance of this issue will become clarified when cultural differences among different countries are taken

[49] Ori Berkovitz, "Investing 2 Billion Nis Per Year in Smart Cities, Agriculture and Academia: How Israel Plans to Become an AI Power," *Globes*, November 18, 2019, https://www.globes.co.il/news/article.aspx?did=1001307714 [Hebrew].

[50] Mariarosaria Taddeo and Luciano Floridi, "Regulate Artificial Intelligence to Avert Cyber Arms Race," *Nature*, 556, no. 7701 (April 16, 2018): 296–298, https://doi.org/10.1038/d41586-018-04602-6.

into account, as well as the potential impact on both the design of AI in other countries and on the definition of decision-making ethics in this field.

The implementation of these policy recommendations and any additional ones formulated as the problem becomes better understood can assist in preventing potentially harmful attacks. This potential rises with the increased use of AI-based systems and their responsibility for various critical functions. Appropriate actions ahead of time may help prevent destructive outcomes on the national and international levels.