Artificial Intelligence for Cybersecurity

Artificial intelligence (AI) promises to change cybersecurity in the coming years. It will likely enhance both cyber offense and defense, and contribute to shaping the cyber threat landscape. Governing these changes is challenging, particularly for state-related actors. It requires them to adopt adequate policy and normative frameworks.

By Matteo E. Bonfanti
and Kevin Kohler

Through machine learning, images can be generated based on the input of photos of real people.
*Guess: Which of these children is real, and which one is generated by AI? (*Solution on the last page)*

AI is an umbrella term John McCarthy, a computer scientist, coined in 1955 and defined as "the science and engineering of intelligent machines". Today, AI refers to an enabling system and a field of research. As such, AI is the scientific discipline devoted to making artificial systems able to perform tasks that are thought to require a certain degree of rationality or intelligence when performed by humans. There are different approaches to achieving such a goal. One of these is machine learning, whose core components are learning algorithms, data, and computational power for training algorithms. Most of the recent successes in AI come from a subset of machine learning: deep learning. It employs deep neural networks consisting of numerous layers of artificial neurons, each of which transforms the data it receives. Neural networks are inspired by the human brain. With increasing learning capacity and decision-making power, artificial systems can be expected to grow more autonomous over time.

AI is also described as an enabling technology because it can be deployed across many different domains, for civil (see CSS Analysis No. 260) and military purposes (see CSS Analysis No. 251), as well as to do good or harm. Unsurprisingly, it can also be applied to achieve cybersecurity-related goals, such as so-called AI for cybersecurity. The expression refers to technological solutions: Integrating machine learning approaches and capabilities to process large amounts of information and derive insights that can inform a course of action relevant for cyber-related purposes.

Security Concerns About AI

The AI research community has a bias for openness. Researchers believe knowledge should be free. They also look for public timestamps on research for professional prestige in a highly dynamic field. Thus, they do not often just share vague descriptions of their achievements but openly disseminate source codes, trained models, tutorials, and even datasets on the Internet. Access to the last main ingredient of AI, computing power, has also increased through on-demand cloud computing providers. Hence, AI research and developments are experiencing a

© 2020 Center for Security Studies (CSS), ETH Zurich

rapid diffusion, sometimes referred to as a democratization of AI.

To limit harm from the proliferation of evermore powerful multi-purpose tools to malicious actors, AI researchers are exploring new approaches, such as responsible AI licenses or the staged release of fully trained models. The latter is inspired by the responsible disclosure of zero-day vulnerabilities in cybersecurity and was first used in the release of OpenAI's language model GPT-2 in 2019. There is less openness when it comes to the applied side, as business datasets and the models trained thereon are viewed through an economic lens as intangible assets, the confidentiality of which must be protected against theft and espionage. In some countries, export restrictions are applied to particularly sensitive datasets, such as genomic information of the population, or algorithms. Furthermore, there are discussions about extending limitations to AI hardware and related tools.

AI still lacks robustness and often fails in distinctly non-human ways. For example, it might misclassify pictures based on accidental background correlations in the training dataset, unusual viewing angles, or due to manipulations of very low-level features that humans do not actively perceive. As such, adversarial actors can exploit a number of novel yet unresolved and often unknown vulnerabilities to impair the decision-making quality of AI-based systems. Exploits might consist of "data poisoning" attacks – injections into the training data that causes a learning algorithm to make mistakes – or "adversarial examples", digital inputs and real-life artefacts designed to be misclassified by machine-learning solutions. The latter are most effective if the parameters of the AI model are known, so-called white box attacks. However, they can also work without such knowledge, in "black box attacks".

AI and the Cyber Threat Landscape

The deployment of AI components for cyber-related purposes can impact the cyber threat landscape in three ways. Absent the adoption of any substantial preventive measure, AI can: expand existing cyber threats (quantity); alter the typical character of these threats (quality); and introduce new and unknown threats (quantity and quality).

AI could expand the set of actors who are capable of carrying out malicious cyber activities, the rate at which these actors can carry out the activities, and the set of plausible targets. This claim follows the efficiency, scalability, and adaptability of AI as well as the "democratization" of research and development in this field. In particular, the diffusion of AI components among traditional cyber threat actors – states, criminals, hacktivists, and terrorist groups – could increase the number of entities for whom carrying out attacks may become affordable. Given that AI applications are also scalable, actors who possess the resources to carry out attacks may gain the ability to do so at a higher rate. New targets to hit may become worthwhile for them.

AI can expand existing cyber threats, alter their character, and introduce new threats.

From a qualitative point of view, AI-powered cyberattacks could also feature in more effective, finely targeted, and sophisticated actions and attacks. Increased effectiveness derives from the attributes of efficiency, scalability, and adaptability of these solutions. Potential targets are more easily identified and scrutinized.

Finally, AI could enable a new variety of malicious activities that exploit the vulnerabilities these technologies introduced in the cyber systems that integrate them. In this regard, cybersecurity itself becomes relevant to AI research and development. To preserve their proper functioning, reliability, and integrity as well as to avoid nefarious effects, AI-integrated cyber systems require safeguards from cyber incidents or attacks. The adoption of cybersecurity practices as well as the promotion of broad cyber hygiene programs with specific requirements for AI research, development, and application is referred to as "cybersecurity for AI".

Defensive and Offensive Use

Many features of AI that make it appropriate for cyber defense applications also make it suitable for cyber offense. Therefore, in the next three to five years, one should expect organizations to adopt and implement AI-based cyber defense capabilities to safeguard their assets, such as networks, information, and people, from adversaries who might leverage both AI- and non-AI tools for offensive purposes. Similarly, there will be actors employing AI-powered cyber offense capabilities to compromise targets who might engage in AI- or non-AI-integrated cyber defense. In particular, AI-based cyber capabilities may support activities aimed at protecting from or executing computer-network operations, be it attacks or exploitation. They will also likely support defense from or execution of so-called cyber information and influence operations.

Both defense and offense can benefit from the deployment of AI to produce cyber intelligence, i.e. actionable knowledge to support decision making on cyberspace-related issues. Indeed, AI is able to integrate several functions of the cyber intelligence process, in particular the collection, processing, and analysis of information. It can boost information gathering and widen its scope to multiple sources and several end points. It may also enhance the selection of information and corroborate it with additional data provided by other sources. AI can also support analysis by finding hidden patterns and correlations in processed data. By integrating AI capabilities into these functions, the cyber intelligence process will likely advance in terms of automation and speed.

AI-based Malware Detection

Training a neural network on a large dataset of files that are labelled as goodware or malware allow it to have decent intuition about whether a new file is malicious without relying on manually updated lists. This will likely **improve the discovery of modern and emerging malwares**, which can automatically generate novel variants to elude traditional rule-based identification approaches and help in attributing these variants to the correct malware family. At the same time, this binary classification task is far from easy. The prevalence of malware within all files is very low, which makes such **classifiers prone to trigger false positives** and block executable files on legitimate software. As a workaround, some companies have whitelisted harmless families of files. Yet in turn researchers have shown that it is easy to append whitelisted files to malware so that it goes undetected. Hence, for the foreseeable future, **AI-based malware detection is a complement and not a substitute to traditional methods**.

© 2020 Center for Security Studies (CSS), ETH Zurich

Computer Network Operations

The ability of AI components to produce cyber intelligence will translate into specific defensive applications at the tactical/technical and operational level of cybersecurity. Operationally, AI could be used to retrieve and process data gathered from network security analysis programs and correlate them against other available information. Tactically, AI will increasingly support cyber threat detection, analysis, and, possibly, prevention. In particular, it will upgrade Intrusion Detection Systems (IDS) aimed at discovering illicit activities within a computer or a network. The same goes for spam and phishing detection systems as well as malware detection and analysis tools (see Box). AI components will also integrate multi-factor authentication or verification systems. These will help detect a pattern of behaviour for a particular user to identify changes in those patterns. Another promising target for tactical defensive application of AI is automated vulnerability testing, also known as fuzzing.

AI applications will also be used for cyber offensive purposes, i.e. to compromise a target organization or user, its networks, and the data processed. They will enable more numerous and sophisticated cyberattacks. As in the case of defense, AI applications may generate cyber intelligence to prepare and implement attacks. They may improve the selection and prioritization of targets for cyberattacks involving social engineering. These are attacks employing psychological manipulation of target users to get them to reveal specific information or perform a specific action for illegitimate reasons. Thanks to AI, potential victims' online information can be harvested and processed to automatically generate custom malicious websites, emails, and links based on profiling.

AI components will also enhance adversarial vulnerability discovery and exploitation. They will prompt sophistication in malware designing and functioning, as well as support their obfuscation. AI-powered malware can evade detection and respond creatively to changes in the target's behaviour. They will function as an autonomous and adaptive implant that learns from the host in order to remain undetected; search for and classify interesting content for exfiltration; search for and infect new targets; and discover new pathways or methods for moving through a network and finding the key data that are the ultimate target of an attack. Already in 2018, IBM researchers developed a malware of this type, dubbed Deeplocker. Finally, AI will also be deployed to spoof authentication or verification systems, such as those integrating biometric identifiers.

Cyber Information and Influence

AI will likely enhance the planning and running of cyber information and influence operations. By supporting automation, it will boost digital information gathering as well as surveillance of targets' online behaviour. It will increase the set of tools available to inform and influence adversaries through and within cyberspace, especially by leveraging social media platforms. In social media, AI can improve bots and social bots management as well as allow the production of messages targeted at those most susceptible to them. Following an ongoing trend, AI-based solutions, especially those integrating deep generative adversarial neural networks, will help to create manipulated digital content. Such content, known as synthetic media or deepfakes, consists of hyper-realistic video, audio, imagery, or text that are not easily recognizable as fake through manual or other conventional forensic techniques. Once generated, synthetic media may be abused. Harmful employment is already abundant and documented in the media. For the most part, it consists of the deployment of AI-doctored videos for targeted online cyber bullying, stalking, and defamation. Probably on the rise in the near-term is the weaponization of synthetic media for cyber-enabled blackmailing, scamming, corporate sabotage via market or other types of manipulative operations, and for political propaganda.

AI-generated synthetic media can be used for blackmailing, scamming, sabotage, and for political propaganda.

Although AI will integrate and enable the above activities, it will also contribute to countering them. From a defensive point of view, AI can support the detection of and response to cyber influence and information operations. It can help monitor the online environment, such as social media platforms, identify the early signs of malicious operations, such as increasing bots or socialbots activities, as well as discover altered digital content, including synthetic media.

Further Readings

Matteo E. Bonfanti, Artificial Intelligence and the Offence-Defence Balance in Cyber Security, in: Dunn Cavelty, M. & Wenger A. (Eds), Cyber Security Politics: *Socio-Technological Transformations and Political Fragmentation* (Routledge, 2020, forthcoming).

Matteo E. Bonfanti, Cyber Intelligence: In Pursuit of a Better Understanding for an Emerging Practice, in: I*NSS Cyber, Intelligence, and Security* 2:1 (2018), pp. 105–121.

Miles Brundage et al., *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation,* February 2018.

Ben Buchanan, A National Security Research Agenda for Cybersecurity and Artificial Intelligence. Center for Security and Emerging Technology, *CSET Issue Brief,* May 2020.

A Matter of Governance

AI will affect cybersecurity in the coming years. It will enrich the cyber threat landscape – both in quantitative and qualitative terms. It will likely increase the number of cyber threat actors, offer them additional exploitable vulnerabilities and targets, as well as boost their malevolent actions. Conversely, AI will contribute to defense from those threats by enabling the discovery of unknown vulnerabilities, the detection of malicious cyber activities, and the implementation of countermeasures. It will support both cyber defense and offense. It is difficult to establish whether defensive or offensive applications will benefit more. This will likely depend on the capacity of public or private cybersecurity stakeholders to master and leverage AI. It will also depend on their overall ability to identify, understand, and address the risks, threats, and opportunities stemming from the deployment of these technologies.

Governments can play a significant role in addressing these risks and opportunities by managing and steering the AI-induced transformation of cybersecurity. So far, they have sustained AI innovation through multiple policy mechanisms. They have invested in AI infrastructures, encouraged academic education and professional training, funded scientific research, incentivized public-private partnership and collaboration, as well as promoted standards through procurement policies. In consultation with the private sector and civil society, they have sponsored the adoption of guiding principles or basic norms, such as funda-

© 2020 Center for Security Studies (CSS), ETH Zurich

mental rights and data privacy, to sustain responsible and trustworthy innovation in this technological field.

In many countries, governments orient their actions toward the acquisition of AI capabilities according to wide-scope national AI strategies, most of which address cybersecurity as one promising field of application. These strategies are then complemented by sectoral policy instruments or other technical documentation. In general, they aspire to make AI capabilities available to relevant national cybersecurity stakeholders and ensure the latter can employ AI to gain an advantage over their competitors.

Governments can play a significant role by managing the AI-induced transformation of cybersecurity.

To influence the AI-induced transformation of cybersecurity, governments can also establish dynamic testing, validation, and certification standards of AI tools for cyber-related applications. At the international level, they can work towards common norms around AI research and development, and consider smart constraints on the proliferation of knowledge and capabilities in this technological domain. Furthermore, they can foster a positive and inclusive governance of AI by operationalizing high-level principles, such as those adopted by the EU and the Organisation for Economic Co-operation and Development (OECD) for trustworthy AI.

For more on perspectives on Cyber Security Politics, see CSS core theme page.

Matteo E. Bonfanti is a Senior Researcher in the Risk and Resilience Team.

Kevin Kohler is a researcher in the Risk and Resilience Team.

** Solution of the puzzle on page 1: Actually, both photos are AI generated. These "children" do not exist.*

CSS Analyses in Security Policy is published by the Center for Security Studies (CSS) at ETH Zurich. The CSS is a center of competence for Swiss and international security policy. Each month, two analyses are published in German, French, and English.

Editor: Benno Zogg
Language editing: Lisa Watanabe
Layout: Rosa Guggenheim

Feedback and comments: analysen@sipo.gess.ethz.ch
More editions and online subscription: www.css.ethz.ch/cssanalyses

© 2020 Center for Security Studies (CSS), ETH Zurich
ISSN: 2296-0244; DOI: 10.3929/ethz-b-000417116