ARI 139/2020
9 December 2020

Artificial Intelligence and Cybersecurity: A Promising but Uncertain Future

Matteo E. Bonfanti | Senior Researcher in the Cyberdefense Project with the Risk and Resilience Team at the Center for Security Studies (CSS) at ETH Center for Security Studies, Zurich | @Teobonf05

Theme[1]

The adoption of AI-based solutions to achieve cyber-defensive/offensive objectives ("AI for Cybersecurity") is promising but challenging too.

Summary

Governments and private corporations are paying attention to the transformative capacity of AI but despite some early and promising applications of AI to cyber-security there are many open questions on what to expect in the near-term future. The present paper presents a possible scenario based on the on the review of selected scientific and technical literature. It discusses the existing challenges and opportunities deriving from the adoption of AI-based solutions to achieve cyber-defensive/offensive objectives ("AI for Cybersecurity") or counter cyber information and influence operations ("AI and Cyber-influence").

Analysis

Artificial Intelligence (AI) refers to a field of research and an enabling technological system.[2] 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.[3] It is also described as an enabling technology because it can be deployed across many different domains, for civil and military purposes, as well

[1] The present paper draws on a more comprehensive study on Artificial Intelligence and Cybersecurity that is contributed by the author to an edited volume, to be published in 2021.

[2] T. Coombs (2018), *Artificial Intelligence & Cybersecurity for Dummies*, IBM, https://hosteddocs.ittoolbox.com/ai_cybersecurity_dummies.pdf. S. Rusell and P. Norvig (2016), *Artificial Intelligence: A Modern Approach*, Pearson. High-Level Expert Group on Artificial Intelligence, AIHLEG (2019), *A definition of AI: Main capabilities and scientific disciplines*, https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

[3] 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. The MITRE Corporation (2017), "Perspectives on Research in Artificial Intelligence and Artificial General Intelligence Relevant to DoD", JSR-16-Task-003, January, https://fas.org/irp/agency/dod/jason/ai-dod.pdf.

as to do good or harm. Unsurprisingly, it can also be applied to achieve cybersecurity-related goals.[4]

Over the last couple of years, AI has become significantly attractive for both governmental and private cybersecurity stakeholders across the world. The growing interest in this technological field is somehow reflected by the increasing number of dedicated initiatives they have recently promoted. These are aimed at boosting the acquisition/provision of AI capabilities at sustaining/accelerating research and development of these technologies and fostering their application for (among others) cybersecurity purposes. Promising applications include cyber threats detection, analysis, and, possibly, prevention and response. Western governments in general, and the Defence Advanced Research Projects Agency (DARPA) in particular, have been funding fundamental research and development in AI and computer networks since the 1960s. However, driven by profit and visions of transformative AI, today, the private sector invests the largest amount of resources (human, technological, organisational, financial) into the development and commercialisation of artificial intelligence. Subsequently, multinational technology firms, other private organizations, and academia are often at the forefront of AI innovation, whereas militaries and government agencies work towards leveraging these advances. As per Governments, they have explicitly sustained AI advancements through multiple policy mechanisms at least since 2016. They have invested in AI infrastructures, encouraged academic education and professional training, funded scientific research, incentivised public-private partnerships, and collaborations, as well as promoted standards through procurement or other policies. In consultation with the private sector and the broad civil society, they have in some cases sponsored the adoption of guiding principles or basic norms (e.g. fundamental rights, data privacy) to sustain "responsible" or "trustworthy" innovation in this technological field.[5] In many countries –e.g. China, the United Kingdom, Canada, India, Japan, France– Governments orient their actions toward the acquisition of AI capabilities according to wide-scope national AI strategies, most of which address cyber-security as one promising field of application.[6] These strategies are then complemented by further policy instruments or other technical documentation tackling sectorial applications of AI. In general, Governmental policies and their implementing actions pursue the threefold objective of encouraging the uptake of AI, maximising its benefits, and minimising the associated risks. As far as cybersecurity is concerned, policies aspire to make AI capabilities available to relevant national cybersecurity stakeholders (mainly public and private organisations) and ensure they can resort to these capabilities to gain an advantage over their competitors. An advantage which can make the difference in terms of power relations, *i.e.* in terms of the capacity of such stakeholders to safeguard their assets and promote their interests in or through the cyberspace.

[4] M. E. Bonfanti and K. Kohler (2020), «Artificial Intelligence for Cybersecurity", *CSS Analyses in Security Policy* nº 265, 2020, Center for Security Studies (CSS), ETH Zurich.

[5] European Commission (2019), "Building Trust in Human-Centric Artificial Intelligence", COM(2019)168, April 28, https://ec.europa.eu/digital-single-market/en/artificial-intelligence.

[6] J. Cussins Newman (2019), "Toward AI Security. Global Aspirations For A More Resilient Future", *CLTC White Paper Series,* Centre for Long-Term Cybersecurity, p. 34, https://cltc.berkeley.edu/wp-content/uploads/2019/02/CLTC_Cussins_Toward_AI_Security.pdf. OECD Policy Observatory, OECD.AI (2020), *National AI policies & strategies*, https://oecd.ai/policy-areas.

The above-described attention toward AI suggests Governments (and private corporations) largely believe in the transformative capacity of these technologies and are aware of the importance of mastering it in the coming years. It shows there are expectations for the role AI can play in shaping cybersecurity (and security in general). However, there are many issues that are still open for discussion. Yet to be fully understood it is how and/or to what extent: AI will enhance the protection of individuals, organisations, nations and their cyber-dependent assets from hostile threat-actors; it will introduce novel vulnerabilities and enable additional typologies of actions; it will induce cyber-security stakeholders to adapt to changing risk scenarios and opportunities; it will require them to adopt additional measures on the technological, policy, regulatory, or other levels.[7] The listed overarching issues can be broken down in to more specific questions to be tackled from a variety of angles.[8] The most relevant are: how will AI likely impact on the cyber threat landscape? To which degree will AI reshape the so-called cyber kill chain? Will it make cyber capabilities more powerful and destructive? Will it also improve cyberattacks detection, interdiction, and attribution? Can AI be fooled, and how to prevent such risk? How can AI fail and what are the possible consequence of such failures? How will AI change the risk of accidents in cyber operations? How can AI support digital information manipulation and computational propaganda/disinformation? Will AI benefit cyberattack more than cyberdefence, or vice versa? Are AI-based solutions more likely to be proliferated than previous types of cyber tools? Will AI shape strategic stability in cyber operations?

Providing a definitive answer to the above questions is not trivial for different reasons.[9] AI innovation keeps on developing quite fast under the pressure of several forces that are displayed by multiple actors (public/private researchers, developers and providers; policy and regulatory authorities at the domestic or supranational level; security/military agencies; the broad civil society). Making foresights on the mid/long-term outcomes of such development is hard. From a technical point of view only, there have been unimaginable improvements in the AI core infrastructures and components, *i.e.* computing power, algorithms design, standard software frameworks for faster replication of experiments, and the availability of large datasets. Improvements that will probably continue quite fast, boosted by growing public and commercial investments in the field.[10] In addition, AI is not the only technological component which promises to change cyber-security. There are further technologies and techniques displaying a similar transformative capacity (*e.g.* quantum computing, homomorphic encryption). AI will interact with these technologies in a way that can be hardly predicted from now. Therefore, there are few chances to establish *a priori* the whole spectrum of possible interactions of AI with these technologies and foresee the overall impact on cyber-

[7] M. E. Bonfanti (2021 forthcoming), "Artificial Intelligence and the Offence-Defence Balance in Cyber Security", in M. Dunn Cavelty and A. Wenger (Eds.), *Cyber Security Politics: Socio-Technological Transformations and Political Fragmentation*, London, Routledge.

[8] B. Buchanan (2020), "A National Security Research Agenda for Cybersecurity and Artificial Intelligence", *CSET Issue Brief*, Centre for Security and Emerging Technology, May 2020, https://cset.georgetown.edu/research/a-national-security-research-agenda-for-cybersecurity-and-artificial-intelligence/.

[9] M. E. Bonfanti (2021).

[10] S. Fischer and A. Wenger A. (2019), "A Politically Neutral Hub for Basic AI Research", *ETH CSS Policy perspective*, 7(2), pp. 1-4.

security. Furthermore, advances in AI should be understood as socio-technical phenomena that are more than the sum of technological capabilities and scientific/technical knowledge.[11] Progresses made in AI research and applications, and their implications for cyber-security, are inevitably shaped by the models of governance which emerge from the formal/informal, fragmented/coordinated and often unbalanced interactions among public authorities, private organisations and the civil society. Progresses will be also influenced and driven by the above actors' assessment of the risks and opportunities stemming from the deployment of AI for cyber or other security purposes. To note, risks and opportunities are not to be understood in narrow technological terms only, *e.g.* as strictly pertaining to the functioning of AI tools, their safety and efficiency. They are broader and involve further aspects of the communities which are affected by the employment of AI. At the higher level, they involve Nations' economic integrity and well-being, social cohesion, diplomatic relations, or political stability. The governance of such risks and opportunities will therefore reflect individual and collective assessments, visions, values, interests, and challenges. In sum, given the trajectory of AI innovation remains still uncertain and determined by the interaction of multiple players and forces, it is hard to predict how it will impact on cyber-security, especially in the mid/long term.

However, there are some early and promising applications of AI to cyber-security which allow the making of an informed, although general, guess on what to expect in the near-term future. In particular, they allow to speculate on how cybersecurity can be affected by the deployment of AI solutions within the next 3-5 years. Based on the review of selected scientific and technical literature, the present paper presents a possible scenario. It describes how AI can likely affect the cyber-threat landscape. It discusses the existing challenges and opportunities deriving from the adoption of AI-based solutions to achieve cyber-defensive/offensive objectives ("AI for Cybersecurity").[12] It also presents the role AI can play to sustain or counter cyber information and influence operations ("AI and Cyber-influence"). The paper concludes by reflecting on how cybersecurity stakeholders state-actors can govern the risks and opportunities deriving from the AI-induced transformation of cybersecurity.

AI and the cyber threat landscape

The deployment of AI components for cyber-related purposes can affect 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 can carry 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

[11] J. Cussins Newman, 2019, p. 6.

[12] 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.

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.

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. These vulnerabilities might be the cause of incidents or pave the way to both known and unknown malicious forms of exploitation.[13] 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".[14] Experts are aware there is a wide range of potential malicious exploitations that has still to be fully explored. Some of these exploitations could be cyber-related. 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".[15]

Defensive and offensive use: computer network operations

Many features of AI that make it appropriate for cyber defence 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 defence 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 defence. AI-based cyber capabilities may support activities aimed at protecting from or executing computer-network operations, be

[13] A. Patel et al., (2019), "Security Issues, Dangers and Implications of Smart Information Systems", Sherpa Project D1.3, https://dmu.figshare.com/articles/D1_3_Cyberthreats_and_countermeasures/7951292.

[14] T. Gu, B. Dolan-Gavitt and S. Garg (2019), "BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain", March 11, https://arxiv.org/pdf/1708.06733.pdf.

[15] J. M. Spring et al. (2019), "Machine Learning In Cybersecurity: A Guide", *SEI-CMU Technical Report,* nº 5, *Software Engineering Institute-Carnegie Mellon University,* September, p. 11, https://resources.sei.cmu.edu/asset_files/TechnicalReport/2019_005_001_633597.pdf.

they attacks or exploitation. They will also likely support defence from or execution of so-called cyber information and influence operations.[16]

Both defence 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.[17] 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.

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. It will upgrade Intrusion Detection Systems (IDS) aimed at discovering illicit activities within a computer or a network.[18] The same goes for spam and phishing detection systems as well as malware detection and analysis tools. As per the latter, AI will probably improve the discovery of modern and emerging malwares, which can automatically generate novel variants to elude traditional rule-based identification approaches. It will help in attributing these variants to the correct malware family thank to its capacity to recognise some hidden patterns which are invisible to traditional or human-based analysis. 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. Although promising, the described applications for anomaly and threat detection/analysis are tainted with both false negatives and positives.[19] As per the former, pilot testing or early deployment show they are still, and keep on being, a main problem. Even a false positive rate of 0.1% could account to hundreds of false alarms which are unbearable for many organizations.[20]

[16] S. Cordey (2019), "Cyber Influence Operations: An Overview and Comparative Analysis", *CSS Risk and Resilience Reports*, ETH Zurich, October, https://css.ethz.ch/en/services/digital-library/publications/publication.html/c4ec0cea-62d0-4d1d-aed2-5f6103d89f93.

[17] A. Galyardt et al. (2019), "Artificial Intelligence and Cyber Intelligence: An Implementation Guide", https://resources.sei.cmu.edu/asset_files/EducationalMaterial/2019_011_001_548767.pdf.

[18] A. Buczak and E. Guven (2016), A Survey of Data Mining and Machine Learning Methods, IEEE Communications Survey, vol (18), nº 2, https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7307098

[19] Y. Xin et al. (2018), Machine Learning and Deep Learning Methods for Cybersecurity, IEEE Access, May 15, https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8359287

[20] G. Apruzzese et al. (2018), "On the Effectiveness of Machine and Deep Learning for Cyber Security", 10th Int. Conference on Cyber Conflict, https://weblab.ing.unimore.it/people/apruzzese/papers/apruzzese_cycon18.pdf

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 defence, 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.[21] 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. This AI-powered malware conceals its intent until it reaches a specific victim. It carries out its malicious action as soon as the AI component identifies the target through indicators like facial recognition, geolocation, and voice recognition. What is unique about DeepLocker is that it uses AI (deep neural network) to unlock the attack.[22] Finally, AI will also be deployed to spoof authentication or verification systems, such as those integrating biometric identifiers.[23]

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.[24] With regard to these latter, AI can improve bots and social bots management as well as allow the production of messages targeted at those most susceptible to them (similar to behavioural advertisement).[25] Following an ongoing trend, AI-based solutions, especially those integrating deep generative adversarial neural networks (GAN), will help to create manipulated digital content. Such content, known as synthetic media or deep fakes, consists of hyper-realistic video, audio,

[21] M. Brundage et al. (2018), "The Mallicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation", February, https://www.researchgate.net/publication/323302750_The_Malicious_Use_of_Artificial_Intelligence_Forecasting_Prevention_and_Mitigation

[22] The malicious payload will only be unlocked if the intended target is reached. The AI model is trained to behave normally unless it is presented with a specific input: the trigger conditions identifying specific victims. Marc Ph. Strecklin (2018), "DeepLocker: How AI Can Power a Stealthy New Breed of Malware", *Security Intelligence*, 8/VIII/2020, https://securityintelligence.com/deeplocker-how-ai-can-power-a-stealthy-new-breed-of-malware/

[23] A. Patel et al. (2019).

[24] A. Patel et al. (2019), p. 22.

[25] M. Brundage et al. (2018).

imagery, or text that are not easily recognizable as fake through manual or other conventional forensic techniques.[26] Once generated, synthetic media may be abused.[27] 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.[28] 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.[29] In these cases, synthetic media will play as add-ons to "individual/organisation-oriented" or "communities-oriented" information operations.[30]

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 social bots' activities, as well as discover altered digital content, including synthetic media.

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 defence 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 defence 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. 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

[26] A. Collins (2019), "Forged Authenticity: Governing Deepfake Risks", Lausanne: EPFL International Risk Governance Center.

[27] M. E. Bonfanti (2020), "The weaponization of synthetic media: what threat does this pose to national security?", *CIBER Elcano*, nº 57, July, http://www.realinstitutoelcano.org/wps/portal/rielcano_en/contenido?WCM_GLOBAL_CONTEXT=/elcano/elcano_in/zonas_in/ari93-2020-bonfanti-weaponisation-of-synthetic-media-what-threat-does-this-pose-to-national-security.

[28] R. Chesney and D. K. Citron (2018), "Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security", *SSRN Electronic Journal*, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3213954 .

[29] H. Ajder et al. (2019), "T*he State of Deepfakes: Landscape, Threats, and Impact*", Deeptrace, https://regmedia.co.uk/2019/10/08/deepfake_report.pdf.

[30] M. E. Bonfanti (2020).

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.