Three Ethical Challenges of Applications of Artificial Intelligence in Cybersecurity

Article *in* Minds and Machines · June 2019

DOI: 10.1007/s11023-019-09504-8

1 author:

Mariarosaria Taddeo
University of Oxford
122 PUBLICATIONS **2,390** CITATIONS

Minds and Machines
https://doi.org/10.1007/s11023-019-09504-8

LETTER TO THE EDITOR

Three Ethical Challenges of Applications of Artificial Intelligence in Cybersecurity

Mariarosaria Taddeo[1,2]

In 2017, the WannCry and NotPeya showed that attacks targeting the cyber component of infrastructures (e.g. attacks on power plants), services (e.g. attacks to banks or hospitals servers), and endpoint devices (e.g. attacks on mobiles and personal computers) have a great disruptive potential and could cause serious damage to our information societies. WannaCry crippled hundreds of IT systems. And NotPetya costed pharmaceutical giant Merck, shipping firm Maersk and logistics company FedEx around US$300 million each. At a global level, cyber crime causes multi-billion dollar losses to businesses, with average losses per organization running from US$3.8 to US$16.8 million in the smallest and largest quartiles respectively (Accenture 2017).

The picture did not improve in 2018. Data show that over the year 2.6 million people encountered newly discovered malware on a daily basis.[1] Attacks ranged over 1.7 million different forms of malware, and 60% of the attacks lasted less than 1 h. Cyber attacks are escalating in frequency, impact, and sophistication. The escalation is due to several factors, for example, attacking in cyberspace is easier than defending; most attacks remain unattributed and, therefore, unpunished. Moreover, as defences are porous, cyber attacks are more likely to succeed than not (Taddeo 2017b). Artificial intelligence (AI)[2] could help to improve defences and reduce the impact of cyber attacks. This is why initiatives to develop applications of AI for cybers security applications are attracting increasing attention both within the private and public sector (The 2019 Official Annual Cybercrime Report 2019).

[1] https://www.microsoft.com/security/blog/2018/08/09/protecting-the-protector-hardening-machine-learning-defenses-against-adversarial-attacks/.

[2] AI as a form an autonomous, self-learning, interactive agency poses a plethora of ethical issues, that Luciano Floridi and I addressed here (Floridi and Taddeo 2016; Yang et al. 2018).

✉ Mariarosaria Taddeo
mariarosaria.taddeo@oii.ox.ac.uk

[1] Oxford Internet Institute, University of Oxford, Oxford, UK

[2] Alan Turing Institute, London, UK

Published online: 03 June 2019

AI enters in the cybersecurity scenario bringing both bad and good news (Taddeo and Floridi 2018a, b). The bad news is that AI, both in the forms of machine learning and deep learning, risks facilitating the escalation process of attacks, for it enables better targeted, faster, and more impactful attacks. AI can identify systems vulnerabilities that often escape human experts and exploit them to attack a given target. We learned about this potential during the 2016 DARPA Cyber Grand Challenge, when seven AI systems, engaged in a war game and were able to identify and target their opponents' vulnerabilities, while finding and patching their own. Pervasive distribution of AI systems, multiple interactions, and fast-pace execution will make the control of the performance of AI systems progressively less effective, while increasing the risks for unforeseen consequences and errors. Regulations may extenuate the lack of control, by ensuring proportionality of responses, legitimate targets, and responsible behaviour. So it is crucial to start shaping and enforcing policies and norms for the use of AI in cybersecurity both for the private and public sector, now, while the technology is still nascent. Complications may be expected—for example, in 2017 the UN Group of Governmental Experts on Information Security failed to agree on regulations for responsible State's conduct in cyberspace—but they must not hinder the process. The alternative is an AI-weaponised, unsafe, and unstable cyberspace.

Luckily there is also some good news. For AI can improve significantly cyber security and defence measures, and foster stability of cyberspace. When considering the role of AI in cybersecurity from systems level, there are three areas of great impact: system robustness, system resilience, system responses. In all three cases, the potential of AI is coupled with serious ethical risks. If left unaddressed, these risks could hinder the adoption of AI in cybersecurity or pose significant problems for our societies (see footnote 2). Let me delve into each case.

System robustness AI for software testing is a new area of research and development. It is defined as 'emerging field aimed at the development of AI systems to test software, methods to test AI systems, and ultimately designing software that is capable of self-testing and self-healing".[3] AI can help with verification and validation of software, liberating human experts from tedious jobs, and offering a faster and more accurate testing of a given system. In this sense, AI can take software testing to a new level, making systems more robust. This may have a snowball effect and, for example, reduce the value of exploits and zero-days, slowing down the race to acquire vulnerabilities. However, we should be careful as societies in the way we use AI in this context, for delegating testing to AI could lead to a complete deskilling of experts. This would imprudent. Cybersecurity experts need to keep testing systems, for the same reason doctors need to keep reading X-ray scans, so that they still can if AI can't or gets it wrong.

System resilience AI is increasingly deployed for threat and anomaly detection (TAD). TAD can make use of existing security data to train their pattern recognition. Although, more advanced TAD systems claim not to need historical threat information to function. Many of them offer the ability to flag and prioritize threats

[3] www.aitesting.org.

according to the level of risk. These services analyse malware and viruses and some are able to quarantine threats and portions of the system for further investigation. In certain cases, threat scanners access and monitor files, emails, mobile and endpoint devices, or even traffic data on a network. Monitoring extends to users as well. AI can be used to authenticate users by monitoring behaviour and generating biometric profiles, like for example, the unique way in which a user moves her mouse around (BehavioSec: Continuous Authentication Through Behavioral Biometrics 2019). Sometimes, this may imply tracking "sensor data and human-device interaction from your app/website. Every touch event, device motion, or mouse gesture is collected".[4] The risk is quite clear here. AI can improve system resilience to attacks, but this requires extensive monitoring of the system and comprehensive data collection. This poses users' privacy under a sharp pressure, exposes users to extra risks, should data confidentiality be breached, and leads to creating a mass-surveillance effect (Taddeo 2013, 2014b).

System response On the one hand, AI will expand the targeting ability of attackers and offer new means to deliver an attack, enabling them to use more complex and richer attacks. On the other hand, AI will improve response and countering measures. Autonomous and semi-autonomous cybersecurity systems endowed with a "playbook" of pre-determined responses to an attack are already available on the on the market (DarkLight Offers First of Its Kind Artificial Intelligence to Enhance Cybersecurity Defenses 2017). Autonomous systems able to learn adversarial behaviour and generate decoys and honeypots, thus actively luring threat actors (Acalvio Autonomous Deception 2019), are also being commercialised. And AI-enabled cyber weapons have already been prototyped including autonomous malware, corrupting medical imagery, and attacking autonomous vehicles (Mirsky et al. 2019; Zhuge et al. 2007). For example, IBM prototype of an autonomous malware, DeepLocker, uses a neural network to select its targets and disguise itself until it reaches its destination (DeepLocker: How AI Can Power a Stealthy New Breed of Malware 2018). This capability has great relevance for national cyber defence strategies, for it will support active cyber defence strategies, which are already deployed by several state actors, and may help the development of deterrence strategies in cyberspace (Taddeo 2017b; Taddeo and Floridi 2018a).

However, applications of AI for system responses pose some of the most pressing challenges to the stability of cyberspace. It was on the basis of these new AI capabilities that in 2018 a number of US Senators proposed to allow companies to hack back cyber attackers in response to a cyber attack. The proposal was not approved, but the path that it open is dangerously. AI can refine strategies and launch more aggressive counter operations. This may snowball into an intensification of cyber attacks and responses, which, in turn, may lead to kinetic (physical) consequences and pose serious risks of escalation (Taddeo 2017a, b). As I argued elsewhere (Taddeo 2017b; Taddeo and Floridi 2018a), to mitage this risk the international community needs to define a regime of norms regulating state behaviour in cyberspace and to establish an authority able to (1) convene agreement about international norms,

[4] http://www.unbotify.com.

(2) verify states compliance with the norms at national and international level, (3) launch investigations into suspected state-run (or state-sponsored) cyber attacks to ascertain attribution, (4) expose breaches of the norms and the sources of illegitimate cyber attacks, and (5) impose adequate sanctions and punishments.

Ethical analyses are needed to avoid human deskilling, mass-surveillance, and risks following the lack of control of AI systems. The alternative is that the risks may overcome the benefits and, societies may reject AI applications in cybersecurity, despite their potential to improve security of information systems. Partial adoption of AI applications in cybersecurity will make cyber defence even more porous, and offer a strategic advantage to malicious users who will be able to rely on AI extensively to launch new attacks. Ethical design and deployment of AI is a first necessary step in this direction. But more needs to be done, especially when considering state use of AI for cybersecurity purposes. Regulations are necessary to ensure responsible behaviour, protection of individual rights, and identify legitimate actors and targets (Taddeo 2012; Taddeo 2014a). A coordinated effort by civil society, politics, business, and academia will help to identify and pursue the best strategies to make AI a force for good and unlock its potential to foster human flourishing while respecting human dignity.

References

Acalvio Autonomous Deception. (2019). Acalvio. https://www.acalvio.com/.

Accenture. (2017). 2017 Cost of cyber crime study. Accenture. https://www.accenture.com/gb-en/insight-cost-of-cybercrime-2017.

BehavioSec: Continuous Authentication Through Behavioral Biometrics. (2019). BehavioSec. https://www.behaviosec.com/.

DarkLight Offers First of Its Kind Artificial Intelligence to Enhance Cybersecurity Defenses. (2017). 26 July 2017. https://www.businesswire.com/news/home/20170726005117/en/DarkLight-Offers-Kind-Artificial-Intelligence-Enhance-Cybersecurity.

DeepLocker: How AI Can Power a Stealthy New Breed of Malware. (2018). *Security intelligence* (blog). 8 August 2018. https://securityintelligence.com/deeplocker-how-ai-can-power-a-stealthy-new-breed-of-malware/.

Floridi, L., & Taddeo, M. (2016). What is data ethics? *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374*(2083), 20160360. https://doi.org/10.1098/rsta.2016.0360.

Mirsky, Y., Mahler, T., Shelef, I., & Elovici, Y. (2019). CT-GAN: Malicious tampering of 3D medical imagery using deep learning. *ResearchGate*. https://www.researchgate.net/publication/330357848_CT-GAN_Malicious_Tampering_of_3D_Medical_Imagery_using_Deep_Learning/figures?lo=1.

Taddeo, M. (2012). An analysis for a just cyber warfare. In *2012 4th international conference on cyber conflict (CYCON 2012)* (pp. 1–10).

Taddeo, M. (2013). Cyber security and individual rights, striking the right balance. *Philosophy and Technology, 26*(4), 353–356. https://doi.org/10.1007/s13347-013-0140-9.

Taddeo, M. (2014a). Just information warfare. *Topoi*. https://doi.org/10.1007/s11245-014-9245-8.

Taddeo, M. (2014b). The struggle between liberties and authorities in the information age. *Science and Engineering Ethics*. https://doi.org/10.1007/s11948-014-9586-0.

Taddeo, M. (2017a). The limits of deterrence theory in cyberspace. *Philosophy and Technology*. https://doi.org/10.1007/s13347-017-0290-2.

Taddeo, M. (2017b). Deterrence by norms to stop interstate cyber attacks. *Minds and Machines*. https://doi.org/10.1007/s11023-017-9446-1.

Taddeo, M., & Floridi, L. (2018a). Regulate artificial intelligence to avert cyber arms race. *Nature, 556*(7701), 296–298. https://doi.org/10.1038/d41586-018-04602-6.

Taddeo, M., & Floridi, L. (2018b). How AI can be a force for good. *Science, 361*(6404), 751–752. https://doi.org/10.1126/science.aat5991.

The 2019 Official Annual Cybercrime Report. (2019). Herjavec group. https://www.herjavecgroup.com/the-2019-official-annual-cybercrime-report/.

Yang, G.-Z., Bellingham, J., Dupont, P. E., Fischer, P., Floridi, L., Full, R., et al. (2018). The grand challenges of science robotics. *Science Robotics, 3*(14), 7650. https://doi.org/10.1126/scirobotics.aar7650.

Zhuge, J., Holz, T., Han, X., Song, C., & Zou, W. (2007). Collecting autonomous spreading malware using high-interaction honeypots. In S. Qing, H. Imai, & G. Wang (Eds.), *Information and communications security. Lecture notes in computer science* (pp. 438–451). Berlin: Springer.