Preprint

The article has been published in Nature Machine Intelligence

DOI : 10.1038/s42256-019-0109-1

Trusting Artificial Intelligence in Cybersecurity, a Double-Edged Sword

Mariarosaria Taddeo,[1,2,3] Tom McCutcheon,[4] Luciano Floridi[1,2]

[1]Oxford Internet Institute, University of Oxford, 1 St Giles', Oxford, OX1 3JS, UK

[2]Alan Turing Institute, British Library, 96 Euston Rd, London NW1 2DB, UK

[3]Department of Computer Science, 15 Parks Rd, Oxford, OX1 3QD, UK

[4]Defence Science and Technology Laboratories, Salisbury, SP40JQ, UK

Abstract

Applications of artificial intelligence (AI) for cybersecurity tasks are attracting increasingly more attention from the private and the public sectors. Estimates indicate that the market for AI in cybersecurity will grow from US $1billion in 2016 to a US $34.8 billion net worth by 2025. The latest national cybersecurity and defence strategies of several governments mention explicitly AI capabilities. At the same time, initiatives to define new standards and certification procedures to elicit users' trust in AI are emerging on a global scale. However, trust in AI (both machine learning and neural networks) to deliver cybersecurity tasks is a double-edged sword: it can improve significantly cybersecurity practices, but can also facilitate new forms of attacks to the AI applications themselves, which may pose severe security threats. In this article, we argue that trust in AI for cybersecurity is unwarranted and that, to reduce security risks, some form of control to ensure the deployment of *reliable AI* for cybersecurity is necessary. To this end, we offer three recommendations focusing on the design, development, and deployment of AI for cybersecurity.

The 2019 Global Risks Report of the World Economic Forum ranks cyber attacks among the top five most likely source of severe risks on a global scale [1]. The report is in line with other analyses [2] [3] about the escalation in frequency and impact of cyber attacks. For example, in the first half of 2018 cyber attacks compromised 3.3 billion records, almost 70% more the records (2.7 billion) compromised during the entire 2017 [4]. Attacks are also becoming faster in reaching their targets and more mutable. A Microsoft study shows that 60% of the attacks in 2018 lasted less than an hour and relied on new forms of malware [5].

Artificial Intelligence (AI) can lower these figures, and the associate human capital and efficiency costs that cybersecurity teams face, in three ways (later, we shall refer to them as the 3R: *robustness*, *response*, and *resilience*). First, AI can improve a system's *robustness,* that is, the capacity of a system to keep behaving as expected even when it processes erroneous inputs, thanks to self-testing and self-healing software [6]. Second, AI can advance a system's *response*, that is, the capacity of a system to defeat an attack autonomously, refine future strategies on the basis of the achieved success, and possibly launch more aggressive counter operations with each iteration [7]. AI systems that support responses to attacks, generating decoys and honeypots for attackers, are already available on the market [8]. Third, AI can increase a system's *resilience*, that is, the ability of a system to withstand attacks, by facilitating threat and anomaly detection (TAD) – data indicate that by 2022, AI will deal with 50% of TAD tasks [9] – and supporting security analysts in retrieving information about cyber threats [10].

Because of its impact on the 3R, applications of AI in cybersecurity offer a tactical and a strategic advantage. Tactically, AI can improve the security of systems and reduce its vulnerability to attacks. Strategically, AI can alter the dynamics that facilitate offence over defence in cyberspace. For example, the use of AI to improve systems' robustness may have a knock-on effect and decrease the impact of *zero-days* attacks (these leverage vulnerabilities of a system that are exploitable by attackers as long as they remain unknown to the system providers or there is no patch to resolve them), thus reducing their value on the black market. At the same time, AI systems able to launch counter responses to cyber attacks independently of the identification of the attackers could enable defence to respond to attacks even when they are anonymous [11].

Tactical and strategic advantages explain the growing trust in AI applications in cybersecurity, from the private and the public sectors. Estimates indicate that the market for AI in cybersecurity will grow from US $1billion in 2016 to a US $34.8 billion net worth by 2025 [12]. The latest national cyber security and defence strategies of several governments (Australia, China, Japan, Singapore, UK, and US) mention explicitly AI capabilities, which are already deployed to improve the security of critical national infrastructures, like transport, hospitals, energy, and water supply. However, trust in AI (both machine learning and neural networks) to deliver the 3R advantages is a double-edged sword. It can significantly improve cybersecurity practices, but can also facilitate new forms of attacks to the AI applications themselves, which may generate new categories of vulnerabilities posing severe security threats.

In this article, we distinguish (both conceptually, in terms of theory and understanding, and operationally, in terms of actual policies, procedures and strategies) *trust* from *reliance*: while trust is a form of delegation of a task with no (or a very minimal level of) control on the way the delegated task is performed [13], reliance envisages some form of control over the execution of a given task [14], including, most importantly, its termination. We argue that trust in AI for 3R is unwarranted and that, to reduce security risks, some form of control to ensure the deployment of *reliable* AI in cybersecurity is necessary. To this end, we offer three recommendations focusing on the design, development, and deployment of AI for 3R.

Vulnerabilities of AI

Previous generations of cyber attacks aimed mostly at stealing data (*extraction*) and breaking systems (*disruption*). New forms of attacks on AI systems seek to gain *control* of the targeted system and change its behaviour, thus undermining the potential of AI to improve the 3R.

To gain control, three types of attacks are particularly relevant: *data poisoning*; *tempering* of categorization models; and *backdoors* [15].[1] All of them exploit the learning ability of AI systems to change their behaviour. For example, attackers may introduce carefully crafted, erroneous data among the legitimate data used to train the system in order to alter its behaviour. A recent study showed that, by adding 8% of erroneous data to an AI system for drug dosage, attackers could cause a 75.06% change of the dosages for half of the patients relying on the system for their treatment [16]. Similar results can be achieved by manipulating the categorization models of neural networks. Using pictures of a specially 3-D printed turtle, researchers exploited the learning method of an AI system to deceive it into classify turtles as rifles [17]. Similarly, backdoor-based attacks rely on hidden associations (triggers) added to the AI model to override correct classification and make the system perform unexpectedly [18]. In a famous study, images of stop signs with a special sticker were added to the training set of a neural network and labelled as speed limit sign [19]. This tricked the model to classify any stop sign with that sticker on as a speed limit sign. The trigger would cause autonomous vehicles to speed through, rather than stopping at, crossroads, thus posing severe safety risks.

Once launched, attacks on AI are hard to detect. The networked, dynamic, and adaptive nature of AI systems makes it problematic to explain their internal processes (this is known as lack of *transparency*) and to reverse-engineer their behaviour to understand what exactly has determined a given outcome, whether this is due to an attack, and of which kind. Furthermore, attacks on AI can be deceptive. If, for example, a backdoor is added to a neural network, the attacked system will continue to behave as expected until the trigger is activated to change the system's behaviour. But even when the trigger is activated, it may be difficult to understand when the compromised system is showing some 'wrong' behaviour, because a skilfully crafted attack may determine only a minimal divergence between the actual and the expected behaviour. The difference could be too small to be noticed, yet sufficient to enable attackers to achieve their goals. For example, it is possible [20] to trick an AI image recognition system to misclassify subjects wearing specially-crafted eyeglasses. Arguably, a similar attack could target a system controlling access to a facility and enable access to malicious actors without raising any alert for a security breach. This is why it is crucial to ensure robustness of an AI system, so that it continues to behave as expected even when their inputs or model are perturbed by an attack. Unfortunately, assessing the robustness of a system requires testing for all possible input perturbations. This is practically impossible, because the number of possible perturbations is often exorbitantly large. For instance, in the case of image classification, imperceptible perturbations at pixel-level can lead the system to misclassify an object with high-level confidence [21], [22]. So it turns out that assessing the robustness of AI is often a computationally intractable problem: it is unfeasible to foresee exhaustively all possible erroneous inputs to an AI system, and then measure the divergence of the related outputs from the expected ones. The assessment of the robustness of AI systems at design and development stages remains only partially, if all, indicative of their actual robustness once deployed. A different approach is required, as we shall argue below.

Standards and Certification Procedures

The vulnerabilities of AI pose serious limitations to its great potential to improve cybersecurity. New testing methods able to grapple with the lack of transparency of AI systems, and the deceptive nature of cyber attacks targeting them, are necessary in order to overcome these limits. Initiatives to define new standards and certification procedures to assess the robustness of AI systems are emerging on a global scale.

The International Standardisation Organisation (ISO) has established a committee (ISO/IEC JTC 1/SC 42) to work specifically on AI standards. One of these standards (ISO/IEC NP TR 24029-1) concerns the assessment of the robustness of neural networks.

In the US, DARPA launched in 2019 a new research program, called 'Guaranteeing AI Robustness against Deception', to foster the design and development of more robust AI applications. In the same vein, the 2019 US executive order on AI mandated the development of national standards for *reliable*, *robust*, and *trustworthy* AI systems. And in May 2019, the U.S. Department of Commerce's National Institute of Standards and Technology issued a formal request for comments with the aim of defining these standards by the end of 2019.

China is also investing resources to foster standards for robust AI. Following the strategy delineated in the New Generation Artificial Intelligence Development Plan, in 2019 the China Electronics Standardization Institute established three working groups: 'AI and open source', 'AI standardization system in China', and 'AI and social ethics'. They are also expected to publish their guidelines by the end of 2019.

The European Union (EU) may lead by example the international efforts to develop certifications and standards for cybersecurity, because the 2017 Cybersecurity Framework and the 2019 Cybersecurity Act established the infrastructure to create and enforce cybersecurity standards and certification procedures for digital technologies and services available on the EU market. In particular, the Cybersecurity Act mandates the EU Agency for Network and Information Security (ENISA) to work with member states to finalise cybersecurity certification frameworks. Interestingly, a set of pre-defined goals will shape ENISA work in this area [16, Art. 51]. They refer to vulnerability identification and disclosure, access and control of data, especially sensitive or personal data, but none of the pre-defined goals mentions AI. Yet, it is crucial that ENISA will focus also on AI systems, otherwise the certification framework will at best improve only partially the security of digital technologies and services available on the EU market.

The aforementioned initiatives are still embryonic, so it is too early to assess their effectiveness. However, they all share the same goal, for they all seek to elicit human trust in AI systems. Trust is an important element of the US executive order on AI, of the European Commission's Cybersecurity Act, and a focal one of Commission's guidelines for AI [24]. Trust is also central in the 2017 IEEE report on the development of standards for AI in cybersecurity [25]. Users' trust in technology is important to foster adoption [26]. However, defining and developing standards and certification procedures with the goal of developing *trustworthy* AI in cybersecurity is conceptually misleading, and may lead to severe security risks.

Philosophical analyses of trust qualify trust as the decision to delegate a task, without any form of control or supervision over the way the task is executed [13]. Successful instances of trust rest on an appropriate assessment of the trustworthiness of the agent to which the task is delegated (the trustee). Trustworthiness is both a prediction about the probability that the trustee will behave as expected, given the trustee's past behaviour, and a measure of the risk run by the trustor, should the trustee behave differently. When the probability that the expected behaviour will occur is either too low or not assessable, the risk is too high and trust is unjustified. This is the case with trust in AI systems for cybersecurity. The lack of transparency and the learning abilities of AI systems, as well as the nature of attacks to these systems make it hard to evaluate whether the same system will continue to behave as expected in any given context. Records of past behaviour of AI systems are neither predictive of the systems' robustness to future attacks, nor are they an indication that the system has not been corrupted by a dormant attack (e.g., has a backdoor) or by an attack that has not yet been detected. This impairs the assessment of trustworthiness. And as long as the assessment of trustworthiness remains problematic, trust in AI applications for cybersecurity is unwarranted. This is not to say that we should not delegate 3R tasks to AI, especially when AI proves to be able to perform them efficiently and efficaciously. On the contrary, delegation can and should still occur. However, some forms of controls are necessary to mitigate the risks linked to the lack of transparency of AI systems and the lack of predictability of their robustness. Policy strategies seeking to elicit users' trust fail to address this crucial issue.

Making AI in Cybersecurity Reliable

Nascent standards and certification methods for AI in cybersecurity should focus on supporting the *reliability of AI*, rather than trust. Conceptually and operationally, supporting the reliability of AI is different from fostering its trustworthiness. For reliability of AI implies that the technology can, technically, perform cybersecurity tasks successfully, but the risks that the technology may behave differently from what expected are too high to forgo any form of control or monitoring over execution of the delegated task. Thus, supporting the reliability of AI for 3R task implies envisaging forms and degrees of (operational) control adequate to the learning nature of the systems, their lack of transparency, and the dynamic nature of the attacks, but also feasible in terms of resources, especially time and hence computational feasibility. Below, we suggest three requirements that specify developing and monitoring practices to mitigate the vulnerabilities of AI systems and improve their reliability with respect to the 3R.

1) In-house development. The most common forms of attacks to AI systems are facilitated by the use of commercial services offering support for development and training of AI, such as cloud, virtual machines, natural language processing, predictive analytics and deep learning [27]. A breach in a cloud system, for example, may provide the attacker with access to the AI model and the training data. Therefore, standards for AI applications for the security of national critical infrastructures should ensure that reliable suppliers design and develop their models in house, and that data for system training and testing are collected, curated, and validated by the systems providers directly, and maintained securely. Although this requirement would not eliminate all the possibilities of attacks, it would rule out many forms of attacks leveraging internet connections to access data and models.

2) Adversarial training. AI improves its performances using feedback loops, which enable it to adjust its own variables and coefficients with each iteration. This is why adversarial training between AI systems can help improving their robustness as well as facilitate the identification of vulnerabilities of the system. This is a well-known method to improve system robustness [28]. However, research also shows that its effectiveness depends on the refinement of the adversarial model [22], [29]. Standards and certification processes should mandate adversarial training but also establish appropriate levels of refinement of models. In this case too, it is essential that models are developed in house and specifically for the task at hand.

3) Parallel and dynamic monitoring. The limits in assessing the robustness of AI systems, the deceptive nature of attacks, and learning abilities of the targeted systems require some form of *constant* (not merely *regular*, i.e. at time intervals, but continuous, 24 hours a day, 7 days a week) monitoring during deployment. Monitoring is necessary to ensure that divergence between the expected and actual behaviour of a system is captured early and promptly, and addressed adequately. To do so, providers of AI systems should maintain a clone system as control system. The clone system should not be considered a 'digital twin' [30] of the deployed system. For the clone is not a virtual simulation of the AI system, it is rather the same system deployed in controlled environmental conditions. And its behaviour is not a simulation of the original system, it is the benchmark (the baseline) against which the behaviour of the original system is assessed.

The clone should go through regular adversarial exercises, simulating real world attacks to establish a baseline behaviour against which the behaviour of the deployed system can be benchmarked. Divergences between the clone and the deployed system should flag degrees of security alerts. A divergence threshold, commensurate to the security risks, should be defined on a case by case basis. It should be noted that too sensitive a threshold (e.g. a 0% threshold) may make monitoring and controlling unfeasible, while too high a threshold would make the system unreliable. However, for systems that satisfy requirements (1) and (2), minimal divergence should not occur frequently and is less likely to be indicative of false positives. Thus, a 0% threshold for these systems may not pose severe limitations to their operability, while it would allow the system to flag concrete threats.

AI can improve the 3R only insofar as it is reliable. Imagine, for example, deploying an AI system for a TAD task without being able to exclude the presence of backdoors in the AI system itself, and hence the possibility that attackers could gain control of the AI system and ensure that a specific attack on the monitored system goes undetected. The three requirements we advocate are preconditions for AI systems performing any of the 3R tasks in a reliable way, and should become essential preconditions for AI systems deployed for the security of national critical infrastructures. Their implementation may be too expensive for average commercial AI applications for cybersecurity. This is why one may imagine that small and medium seize enterprises may adopt these requirements only in part; this may depend, for example, on the nature of their business and the nature of the system to be protected. However, these requirements should be met fully when considering national security and defence. The risks posed by attacks to AI systems underpinning critical infrastructures justify the need for more extensive controlling mechanisms, and hence higher investments.

AI systems are autonomous, self-learning agents interacting with the environment [31]. Their robustness depends as much on the inputs they are fed and interactions with other agents once deployed, as on their design and training. Standards and certification procedures focusing on the robustness of these systems will be effective only insofar as they will take into account the dynamic and self-learning nature of AI systems, and start envisaging forms of monitoring and control that span from the design to the development stages. This point has also been stressed in the OECD principles on AI, which refer explicitly to the need for continuous monitoring and assessment of threats for AI systems [32]. In view of this, defining standards for AI in cybersecurity that seek to elicit trust (and thus forgo monitoring and control of AI) is risky. The sooner we focus standards and certification procedures on developing *reliable* AI, and the more we adopt a 'in house', 'adversarial', and 'always on' strategy, the safer the AI applications for 3R will be.

Funding

Floridi's and Taddeo's work was supported by Privacy and Trust Stream - Social lead of the PETRAS Internet of Things research hub - PETRAS is funded by the Engineering and Physical Sciences Research Council (EPSRC), grant agreement no. EP/N023013/1, Google UK Limited, and Facebook Inc. Funding from Defence Science and Technology Laboratories and the Alan Turing Institute supported the organisation of the research workshop on the 'Ethics of AI in Cybersecurity', which inspired this article.

Acknowledgment

We are grateful for their feedback to Marco Ramili, YOROI, and to the participants in the workshop 'Ethics of AI in Cybersecurity' hosted in March 2019 by the Digital Ethics Lab, OII, University of Oxford and the UK Defence Science and Technology Laboratories.

References

[1] World Economic Forum, 'The Global Risks Report 2018', World Economic Forum, 2018.
[2] 'The 2019 Official Annual Cybercrime Report', Herjavec Group, 2019.
[3] R. Borno, 'The first imperative: The best digital offense starts with the best security defense', 2017. .
[4] Gemalto, 'Breach Level Index', Belcamp, USA, 2018.
[5] Microsoft Defender ATP Research Team, 'Protecting the protector: Hardening machine learning defenses against adversarial attacks', 2018. .
[6] T. M. King, J. Arbon, D. Santiago, D. Adamo, W. Chin, and R. Shanmugam, 'AI for Testing Today and Tomorrow: Industry Perspectives', in *2019 IEEE International Conference On Artificial Intelligence Testing (AITest)*, Newark, CA, USA, 2019, pp. 81–88.
[7] 'DarkLight Offers First of its Kind Artificial Intelligence to Enhance Cybersecurity Defenses', 26-Jul-2017. [Online]. Available: https://www.businesswire.com/news/home/20170726005117/en/DarkLight-Offers-Kind-Artificial-Intelligence-Enhance-Cybersecurity. [Accessed: 23-Feb-2019].
[8] 'Acalvio Autonomous Deception', *Acalvio*, 2019. [Online]. Available: https://www.acalvio.com/. [Accessed: 08-Mar-2019].
[9] IDC FutureScape, 'IDC FutureScape: Worldwide IT Industry 2019 Predictions', 2018.
[10] S. Mittal, A. Joshi, and T. Finin, 'Cyber-All-Intel: An AI for Security related Threat Intelligence', *arXiv:1905.02895 [cs]*, May 2019.
[11] M. Taddeo and L. Floridi, 'Regulate artificial intelligence to avert cyber arms race', *Nature*, vol. 556, no. 7701, pp. 296–298, Apr. 2018.
[12] MarketsandMarkets, 'Artificial Intelligence in Cybersecurity Market by Technology Machine Learning, Context Awareness - 2025 | MarketsandMarkets', *MarketsandMarkets*. [Online]. Available: https://www.marketsandmarkets.com/Market-Reports/ai-in-cybersecurity-market-224437074.html. [Accessed: 07-Mar-2019].
[13] M. Taddeo, 'Modelling Trust in Artificial Agents, A First Step Toward the Analysis of e-Trust', *Minds and Machines*, vol. 20, no. 2, pp. 243–257, Jun. 2010.
[14] M. Taddeo, 'Trust in Technology: A Distinctive and a Problematic Relation', *Know Techn Pol*, vol. 23, no. 3–4, pp. 283–286, Dec. 2010.
[15] B. Biggio and F. Roli, 'Wild Patterns: Ten Years After the Rise of Adversarial Machine Learning', in *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security - CCS '18*, Toronto, Canada, 2018, pp. 2154–2156.
[16] M. Jagielski, A. Oprea, B. Biggio, C. Liu, C. Nita-Rotaru, and B. Li, 'Manipulating Machine Learning: Poisoning Attacks and Countermeasures for Regression Learning', *arXiv:1804.00308 [cs]*, Apr. 2018.
[17] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok, 'Synthesizing Robust Adversarial Examples', *arXiv:1707.07397 [cs]*, Jul. 2017.
[18] C. Liao, H. Zhong, A. Squicciarini, S. Zhu, and D. Miller, 'Backdoor Embedding in Convolutional Neural Network Models via Invisible Perturbation', *arXiv:1808.10307 [cs, stat]*, Aug. 2018.
[19] K. Eykholt *et al.*, 'Robust Physical-World Attacks on Deep Learning Visual Classification', in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 1625–1634.
[20] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, 'Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition', in *Proceedings of the 2016*

ACM SIGSAC Conference on Computer and Communications Security - CCS'16, Vienna, Austria, 2016, pp. 1528–1540.
[21] C. Szegedy *et al.*, 'Intriguing properties of neural networks', *arXiv:1312.6199 [cs]*, Dec. 2013.
[22] J. Uesato, B. O'Donoghue, A. van den Oord, and P. Kohli, 'Adversarial Risk and the Dangers of Evaluating Against Weak Attacks', *arXiv:1802.05666 [cs, stat]*, Feb. 2018.
[23] European Union, 'Regulation of the European Parliament and of the Council on ENISA (the European Union Agency for Cybersecurity) and on information and communications technology cybersecurity certification and repealing Regulation (EU) No 526/2013 (Cybersecurity Act)', 2019.
[24] High Level Expert Group on Artificial Intelligence, 'Ethics Guideline for Trustworthy AI', 2019.
[25] IEEE, 'Artificial Intelligence and Machine Learning Applied to Cybersecurity', 2017.
[26] M. Taddeo, 'Trusting Digital Technologies Correctly', *Minds and Machines*, Nov. 2017.
[27] T. Gu, B. Dolan-Gavitt, and S. Garg, 'BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain', *arXiv:1708.06733 [cs]*, Aug. 2017.
[28] A. Sinha, H. Namkoong, and J. Duchi, 'Certifying Some Distributional Robustness with Principled Adversarial Training', *arXiv:1710.10571 [cs, stat]*, Oct. 2017.
[29] N. Carlini and D. Wagner, 'Towards Evaluating the Robustness of Neural Networks', in *2017 IEEE Symposium on Security and Privacy (SP)*, 2017, pp. 39–57.
[30] E. Glaessgen and D. Stargel, 'The Digital Twin Paradigm for Future NASA and U.S. Air Force Vehicles', in *53rd AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference
20th AIAA/ASME/AHS Adaptive Structures Conference
14th AIAA*, Honolulu, Hawaii, 2012.
[31] G.-Z. Yang *et al.*, 'The grand challenges of *Science Robotics*', *Science Robotics*, vol. 3, no. 14, p. eaar7650, Jan. 2018.
[32] OECD, 'Recommendation of the Council on Artificial Intelligence', 2019.