To appear the Digital Ethics Lab 2020 Yearbook

On The Risks of Trusting Artificial Intelligence: The Case of Cybersecurity*

Mariarosaria Taddeo
Oxford Internet Institute, University of Oxford
Alan Turing Institute
mariarosaria.taddeo@oii.ox.ac.uk

Abstract

In this chapter, I draw on my previous work on trust and cybersecurity to offer a definition of trust and trustworthiness to understand to what extent trusting AI for cybersecurity tasks is justified and what measures can be put in place to rely on AI in cases where trust is not justified, but the use of AI is still beneficial.

Keywords: Artificial Intelligence, Cybersecurity, Digital Ethics, Governance, Reliability, Trust.

1. Introduction

To argue that trust is a key component of individual lives and also of social systems, Luhmann said that "a complete absence of trust would prevent even getting up in the morning" (Luhmann 1979, 4). It is because we trust other parts of society, for example, to work properly that we can delegate to others key tasks and focus only on the activities that we prefer or are trained to do. Without trust, this delegation would be much more problematic as it would require supervision. Imagine not trusting your GP, your children's teacher, or your mechanic. This would require spending a significant portion of your time and resources either performing their tasks or controlling the way in which they perform their tasks. Trust is a facilitator of interactions among the members of a system. This is the case when we consider human systems (e.g. a families), artificial systems (e.g. smart grids), and hybrid systems that involves both human and artificial agents, as in the case of information societies.

As members of mature information societies, we expect to be able to rely on digital technologies (Floridi 2016), in many cases we also have come to *trust* (by delegating and not supervising, as we shall see in the next section) digital technologies with important tasks. We trust artificial intelligence (AI) to identify the information that we like to receive while searching the web; to indicate the best decision to make when hiring a future colleague or when granting parole

* This chapter is based on previously published work on trust (Taddeo 2009; 2010a) and AI applications in cybersecurity (Taddeo, McCutcheon, and Floridi 2019; Taddeo 2019).

during a criminal trial; diagnose diseases and identify possible cure, and more broadly to foster well-being (Burr, Taddeo, and Floridi 2020) and deliver socially good outcomes (Luciano Floridi et al. 2020). We trust robots to take care of our elderly and toddlers, to patrol borders, and to drive or fly us around the globe. We even trust digital technologies to simulate experiments and provide results that advance our scientific knowledge and understanding of the world.

In mature information societies, trust in digital technology is widespread and is resilient. It is only reassessed (almost never broken) in view of serious negative effects. On the one side, digital technologies are so pervasive that trusting them is essential for our societies to work properly. Supervising each run of a machine learning algorithm used to make a decision would require significant time and resources, to the point that it would become disadvantageous and unfeasible to resort to these technologies altogether. On the other side, the tasks that we now delegate to these technologies are of such relevance that a complete lack of supervision may lead to serious risks for our safety and security, as well for the rights and values underpinning our societies (Yang et al. 2018a). This poses the question as to what level of trust is the correct one when considering information societies and specifically trust in digital technologies (Taddeo 2017b).

Digital technologies are not just a tool to preform actions. Rather they are an interface through which we interact, change, perceive, and understand others and the environment surrounding us (Floridi 2014). At the same time, as these technologies share with the environment and with human agents the same informational nature (Floridi 2011), and for this reason they blend in the *infosphere* (Floridi 2002) to the point of becoming an *invisible interface* (Taddeo and Floridi 2018b), one that we trust and about which we forget, until something goes (badly) wrong and recalls our attention onto the interface. The *trust and forget* dynamic is problematic, as it erodes human control on digital technologies and may lead to serious risks for our democracies and the security of our societies.

In this chapter, I draw on my previous work on trust (Taddeo 2009; 2010a) and cybersecurity (Taddeo 2014; Taddeo, McCutcheon, and Floridi 2019; Taddeo 2019) to offer a definition of trust and trustworthiness to understand to what extent trusting AI for cybersecurity tasks is justified and what measures can be put in place to rely on AI in cases where trust is not justified, but the use of AI is still beneficial.

2. Trustworthiness and Trust

Let me begin with a definition of trust:

> "Assume a set of first-order relations functional to the achievement of a goal and that two AAs are involved in the relations, such that one of them (the trustor) has to achieve the given goal

and the other (the trustee) is able to perform some actions in order to achieve that goal. If the trustor chooses to achieve its goal by the action performed by the trustee, and if the trustor rationally selects the trustee on the basis of its trustworthiness, then the relation has the property of minimising the trustor's effort and commitment in the achievement of that given goal. Such a property is a second-order property that affects the first-order relations taking place between AAs, and is called e-trust ", (Taddeo 2010a, 249).[2]

Successful instances of trust rest on an appropriate assessment of the trustworthiness of the trustee. In the relevant literature, trustworthiness has been defined as the set of beliefs that the trustor holds about the potential trustee's abilities, and the probabilities that the trustor assigns to those beliefs (Taddeo 2009). However, this definition is only partially correct, as it overlooks a key aspect of the assessment of trustworthiness. Trustworthiness is both a prediction of the probability that the trustee will behave as expected given the trustee's past behaviour and a measure of the risk that the trustor faces, should the trustee behave differently. In this sense,

> "trustworthiness is […] a measure that indicates to the trustor the probability of her gaining by the trustee's performances and, conversely, the risk to her that the trustee will not act as she expects" (Taddeo 2010a, 247).

Trustworthiness is not a mere assessment of one's own beliefs, neither it the mere reputation of the potential trustee. Rather, it is the guarantee required by the trustor that the trustee will act as it is expected to do without any supervision and that, should the trustee behave differently, the risks for the trustor are still acceptable. In an ideal scenario, rational agents choose to trust only the most trustworthy agent for the execution of a given task. When the value of trustworthiness is low, the risk for the trustor is too high, and trust is unjustified.

Recalling the definition of trust, trust is related to, and affects, pre-existing relations, like for example a relation of communication where 'Alice informs Bob that it's cloudy' (Taddeo 2010a; 2010b). Trust is not to be considered a relation itself. Rather, it is a *property of relations*, something that changes the way relations occur. Consider, Alice and Bob. There is a first-order relation, the communication which ranges over the two agents, and there is the second-order property of trust that ranges over the first-order-relation and affects the way it occurs (Primiero and Taddeo 2012). If Bob trusts Alice to communicate the weather correctly, he will not double-check the weather, nor will he ask how she knows that it is cloudy. He will simply act on the basis of that information.

[2] This definition assumes Kantian regulative ideal of a rational agent, able to choose the best option for itself, given a specific scenario and a goal to achieve. While this approach may not be correct when considering individual decision to trust, it is justified when analysing the decision to trust a specific technology made, for example, by policy makers or public institutions, which are expected to act as rational, informed, agents.

As a property of relations, trust *facilitates* the way relations occur by minimising the trustor's effort and commitment to achieve a given goal. It does so in two ways. First, the trustor can avoid performing the action necessary to achieve his goal himself, because he can count on the trustee to do it. Second, the trustor can decide not to supervise the trustee's performance. Delegation without supervision characterises the presence of trust (Taddeo 2010). This holds true both of trust among human agents and of trust between human and technological artefacts, especially digital technologies.

As we shall see in the next sections, the facilitating effect of trust motivates the growing use of AI to perform cybersecurity tasks. I will argue that defining and developing standards and certification procedures for AI in cybersecurity centred on trust is conceptually misleading and may lead to severe security risks. Let us first describe the current applications of AI in cybersecurity to then analyse the implication of trust in AI in this domain.

3. AI for Cybersecurity Tasks

Analyses of tendencies in cybersecurity ('The 2019 Official Annual Cybercrime Report' 2019; Borno 2017) consistently show an escalation of the frequency and impact of cyber attacks. For example, a Microsoft research shows that 60% of the attacks occurred in 2018 lasted less than a hour and relied on new forms of malware.[3] This is why initiatives to develop applications of AI[4] are attracting increasing attention both within the private and public sector ('The 2019 Official Annual Cybercrime Report' 2019).

AI enters in this scenario bringing both bad and good news (Taddeo and Floridi 2018a; 2018b). The bad news is that AI, both in the forms of machine learning and deep learning, will facilitate the escalation process, for it enables better targeted, faster, and more impactful attacks (Yang et al. 2018b). AI can identify systems vulneraries 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 called 'capture the flag' and where able to identify and target their opponents' vulnerabilities, while finding and patching their own.

Luckily there is also some good news. For AI can also foster and improve significantly cyber security and defence measures. This explains the ever growing effort to apply AI capabilities

[3] https://www.gemalto.com/press/pages/data-breaches-compromised-4-5-billion-records-in-first-half-of-2018.aspx

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

in cybersecurity. Indeed, the latest national cyber security and defence strategies of the US,[5] UK,[6] Chinese,[7] Singapore,[8] Japanese,[9] and Australian[10] government all mention explicitly AI capabilities. When considering the role of AI in cybersecurity from systems level, there are three areas of great impact: system robustness, system resilience, system responses (3R). Let me delve into each case.

Consider system robustness first. 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".[11] AI can help with the 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. 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. Radiologists may need to keep reading x-ray scans for the same reason cyber security experts need to keep testing systems, so that they still can, if AI can't or gets it wrong.

AI is also increasingly deployed system resilience, i.e. for threat and anomaly detection (TAD). TAD can make use of existing security data to train their pattern recognition, and some more advanced systems claim not to need historical threat information to function. Many of them offer the ability to flag and prioritize threats according to the level of risk and transform threat information into visualizations for users. 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 have access to 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 ('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".[12] 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 to train the AI. This may pose users' privacy under

[5] https://media.defense.gov/2019/Feb/12/2002088963/-1/-1/1/SUMMARY-OF-DOD-AI-STRATEGY.PDF
[6] https://www.gov.uk/government/publications/future-force-concept-jcn-117
[7] https://futureoflife.org/ai-policy-china/
[8] https://www.csa.gov.sg/~/media/csa/documents/publications/singaporecybersecuritystrategy.pdf
[9] https://www.nisc.go.jp/eng/pdf/cs-senryaku2018-en.pdf
[10] https://www.business.gov.au/news/budget-2019-20
[11] www.aitesting.org
[12] http://www.unbotify.com

a sharp devaluative pressure, expose users to extra risks should data confidentiality be breached, and lead to creating a mass-surveillance effect (Taddeo 2013; 2014).

Finally let us focus on system response. AI will expand the targeting ability of attackers, enabling them to use more complex and richer data. Enhancing current methods of attack is an obvious extension of existing technology, however using AI within malware can change the nature and delivery of an attack. Autonomous and semi-autonomous cybersecurity systems endowed with a "playbook" of pre-determined responses to an activity, constraining the agent to known actions 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 created a prototype autonomous malware, DeepLocker, which 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 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).

AI can perform successfully 3R tasks and, thus, its adoption in cybersecurity contributes to improve cybersecurity responses. This is why there is a widespread effort to foster trust in AI for 3R tasks. Trust is an important element of the US executive order on AI and a focal one of the European Commission's guidelines for AI (High Level Expert Group on Artificial Intelligence 2019). It is also central in the 2017 IEEE report on the development of standards for AI and ML in cybersecurity (IEEE 2017). However, AI remains a vulnerable and opaque technology, whose trustworthiness is hard to assess.

4. The Vulnerability of AI

If, on the one hand, AI drastically improves cybersecurity practices; on the other, AI systems are not immune from attacks. Their vulnerabilities open avenues for new forms attacks, which may threaten national security and defence, as AI is increasingly deployed to guarantee the security of national critical infrastructures, like transport,[13] hospitals,[14] energy[15] and water supply.[16]

[13] https://www.darktrace.com/en/industries/#transportation
[14] https://www.darktrace.com/en/press/2019/277/
[15] https://gridmod.labworks.org/projects/1.5.01
[16] https://www.sciencedaily.com/releases/2018/07/180718082113.htm

Three types of attacks are particularly relevant when considering the application of AI in cybersecurity: data poisoning; tempering of categorization models; and backdoors (Biggio and Roli 2018).[17] For example, attackers may introduce poisoning data among the legitimate data processed by an AI system to alter its behaviour. A recent study showed that by adding 8% of poisoning data to an AI system for drug dosage, attackers could cause a 75% change of the dosages for half of the patients relying on the system for their treatment (Jagielski et al. 2018). Similar results can be achieved by manipulating the categorization models of neural networks. Using pictures of a 3-D printed turtle, researchers deceived a system into classify turtles as rifles (Athalye et al. 2017). Similarly, backdoor-based attacks rely on hidden associations (triggers) added to the AI model to override correct classification and make the system perform an unexpected behaviour (Liao et al. 2018). 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 (Eykholt et al. 2018). 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.

While previous generation of cyber attacks aimed mostly at data extraction and system disruption; attacks to AI systems are geared to gain control of the targeted system, and change its behaviour. For this reason, cyber attacks targeting AI systems underpinning the security of critical infrastructures can have severe, negative impact on national security and defence. This is why it is crucial to ensure robustness of these systems, that is making sure that a system continues to show the expected behaviour even when the inputs or the model have been perturbed by an attack.

However, developing and assessing robust AI systems is problematic. This is in part due to the lack of transparency - opaqueness - of AI systems. Opaqueness makes it hard to explain why a given system produces a certain output, and this hinders the identification of anomalies and vulnerabilities that may be linked it. Other problems emerge because attacks to AI systems, like backdoors, do not exploit existing vulnerabilities of the system, they leverage system's autonomy, learning and refinement abilities to create *new* vulnerabilities that will only become evident at deployment stage.

Robustness is a measure of the divergence of the actual behaviour of a system from its expected behaviour when system processes erroneous inputs (e.g. poisoning data). Assessing robustness, thus, requires testing for all possible input perturbations. For AI systems the number

[17] Attacks of these types are feasible as developers often outsource training to commercial services; while this makes the development cheaper and faster, it also facilitates third-party access to training dataset and models (Gu, Dolan-Gavitt, and Garg 2017).

of possible perturbations is astronomically 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 (Szegedy et al. 2013; Uesato et al. 2018). This makes assessing robustness an computationally intractable problem: it unfeasible to foresee all possible erroneous inputs to an AI system, and then measure the divergence of the related outputs from the expected ones. For this reason, the assessment of the robustness of AI systems at design and development stages is only partially, if all, indicative of their actual robustness.

At the same time, attacks on AI are quite deceptive. Once attacked, for example a backdoor is added to a neural network, the system will continue to behave as expected, until the trigger is activated causing a change of behaviour. And even when the trigger is activated, it may be hard to understand when a system is showing a wrong behaviour. For a skilfully crafted attack may cause a minimal divergence between the actual and the expected behaviour. The difference could be too small to be noticed, but sufficient to allow attackers to achieve their goals. A study (Sharif et al. 2016), for example, showed that it is possible to trick an AI image recognition system to misclassify subjects wearing specially-crafted eyeglasses. It is not hard to imagine that a similar attack could target a system controlling access to a facility and enable access to one or few malicious actors without raising an alert for a security breach.

The vulnerabilities of AI pose serious limitations to its otherwise great potential to improve cybersecurity. New testing methods able to grapple with the opaqueness of AI systems and the dynamic nature of cyber attacks targeting them are necessary to overcome these limits. Indeed, this is why initiatives to define new standards and certification procedures to assess the robustness of AI systems are emerging on a global scale. For example, 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) will focus on the assessment of the robustness neural networks.

In the US, DARPA launched in 2019 a new research program to Guaranteeing AI Robustness against Deception to foster the design and development of more robust AI applications.[18] 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 later in May 2019, the U.S. Department of Commerce's National Institute of Standards and Technology issued a formal request of information for the development of these standards.

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

[18] https://www.darpa.mil/news-events/2019-02-06

Electronics Standardization Institute established three working groups - 'AI and open source', 'AI standardization system in China', and 'AI and social ethics' - which are expected to publish their guidelines by the end of year.

The European Union (EU) may lead by example international efforts to develop certifications and standards for cybersecurity, for 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 in the EU. The Cybersecurity Act, in particular, 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 (European Union 2019, 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.

All these initiatives are still nascent, so it is hard to assess now the effectiveness of the standards and procedures that they will develop. But their approach is quite clear, for they are all geared to elicit human trust in AI systems. However, the opaqueness and learning abilities of AI systems, and 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. This is because 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 been detected. This impairs the assessment of trustworthiness. As long as the assessment of trustworthiness remains problematic, trust in AI applications for cybersecurity is unwarranted. This is not tantamount to say that we should not delegate cyber security tasks to AI, especially when AI proves to be able to perform them efficiently. Delegation can and should still occur, but some forms of control are necessary to mitigate the risks linked to the opaqueness of AI systems and lack of predictability of their robustness.

5. Making AI in Cybersecurity Reliable

Nascent standards and certification methods for AI in cybersecurity should focus on fostering *reliance* on AI, rather than trust. This implies envisaging forms of control adequate to the learning nature of the systems, their opaqueness, and the dynamic nature of attacks, but also feasible in terms of time and resources spent controlling. We suggest three requirements that should become essentials for AI systems deployed for the security of national critical infrastructures. While the three requirements may pose too high a cost for average commercial AI applications for

cybersecurity; the national security and defence risks that attacks to AI systems underpinning critical infrastructures may pose justify the need for more extensive controlling mechanisms.

i. 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 (e.g. cloud, virtual machines, natural language processing, predictive analytics and deep learning) (Gu, Dolan-Gavitt, and Garg 2017). A breach in a cloud system, for example, may provide the attacker with access to the AI model and the training data. Standards for AI applications for the security of national critical infrastructures should envisages 'in-house' development of models, and ensure that data for system training and testing are collected, curated, and validated by the systems providers directly, and maintained securely in isolated (air-gapped) repositories. While this would not eliminate the possibilities of attacks, it would rule out most forms attacks leveraging internet connections to access data and models.

ii. Adversarial training. AI improves its performances using feedback loops, which enable it to adjust its own variables and coefficients at 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. Indeed, this is a well-known method to improve system robustness (Sinha, Namkoong, and Duchi 2017). But research also shows that it effectiveness depends on the refinement of the adversarial model (Carlini and Wagner 2017; Uesato et al. 2018). Standards and certification processes should mandate adversarial training but also establish appropriate levels of refinement of adversarial models.

iii. Parallel and dynamic control. The limits in assessing robustness of AI systems, the deceptive nature of attacks, and learning abilities of these systems require some form of monitoring during deployment. Monitoring is necessary to ensure that divergence between the expected and actual behaviour of a system is captured promptly and addressed adequately. To do so, providers of AI systems should maintain a clone, air-gapped, system as control system. The clone should go through regular red team exercise, 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 a security alert. A divergence threshold, commensurate to the security risks, should be defined from case to case. It should be noted that too sensitive a threshold (e.g. a 0% threshold) may make monitoring and controlling unfeasible, too high a threshold would make the system unreliable. However, for systems that satisfy requirements (i) and (ii) minimal divergence would not occur frequently and is less likely to be indicative of false positives. Thus, a 0% threshold for these systems would not pose severe limitations to their operability, while it would allow the system to flag concrete threats.

AI systems are autonomous, self-learning agents interacting with the environment (Yang et al. 2018b). Their robustness depends as much from the inputs they are fed and interactions with other agents, as much as from their design and training. Standards and certification procedures focusing on the robustness of these systems will be effective 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.

6. Conclusion

AI systems are autonomous, self-learning agents interacting with the environment (Yang et al. 2018b). 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 (Organisation for Economic Co-operation and Development) principles on AI, which refer explicitly to the need for continuous monitoring and assessment of threats for AI systems. 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 an 'in-house', 'adversarial' and 'always-on' strategy, the safer the AI applications for 3R will be.

The analysis of the risks of trusting AI for cybersecurity tasks is indicative of the level of trust in digital technologies that mature information societies should foster. While trust is necessary for systems to function, not all systems require the same level of trust. In some cases too little trust may encroach the internal dynamics of the system and limit its development; but too much trust may pose serious risks or dissolve the system, because it may lead to the lack of any form of control and coordination. When considering mature information societies, it is crucial to understand what is the right level of trust in digital technologies that would foster technological innovation and adoption, without endangering the security of our societies or breaching their fundamental values. The answer should not be found by a trial and error approach. Once the nature of trust and of digital technologies are clear, a governance approach should be defined able to foster the right level of trust, to limit 'trust and forget' dynamics, and to ensure transparency on the way digital technologies are deployed; meaningful human oversight; ascribing liabilities of designers, providers, and users of digital technologies (Floridi 2016c). The alternative is to risk losing

stewardship of the deployment of digital technologies and hence of the development of the societies that rely on them.

References

'Acalvio Autonomous Deception'. 2019. Acalvio. 2019. https://www.acalvio.com/.

Athalye, Anish, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. 2017. 'Synthesizing Robust Adversarial Examples'. *ArXiv:1707.07397 [Cs]*, July. http://arxiv.org/abs/1707.07397.

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

Biggio, Battista, and Fabio Roli. 2018. '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*, 2154–56. Toronto, Canada: ACM Press. https://doi.org/10.1145/3243734.3264418.

Borno, Ruba. 2017. 'The First Imperative: The Best Digital Offense Starts with the Best Security Defense'. 2017. https://newsroom.cisco.com/feature-content?type=webcontent&articleId=1843565.

Burr, Christopher, Mariarosaria Taddeo, and Luciano Floridi. 2020. 'The Ethics of Digital Well-Being: A Thematic Review'. *Science and Engineering Ethics*, January. https://doi.org/10.1007/s11948-020-00175-8.

Carlini, N., and D. Wagner. 2017. 'Towards Evaluating the Robustness of Neural Networks'. In *2017 IEEE Symposium on Security and Privacy (SP)*, 39–57. https://doi.org/10.1109/SP.2017.49.

'DarkLight Offers First of Its Kind Artificial Intelligence to Enhance Cybersecurity Defenses'. 2017. Business Wire. 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/.

European Union. 2019. '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)'.

Eykholt, Kevin, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. 2018. 'Robust Physical-World Attacks on Deep Learning Visual Classification'. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1625–34. Salt Lake City, UT, USA: IEEE. https://doi.org/10.1109/CVPR.2018.00175.

Floridi, L. 2002. 'On the Intrinsic Value of Information Objects and the Infosphere'. *Ethics and Information Technology* 4 (4): 287–304.

Floridi, L. 2014. *The Fourth Revolution, How the Infosphere Is Reshaping Human Reality*. Oxford: Oxford University Press.

Floridi, Luciano. 2011. *The Philosophy of Information*. Oxford ; New York: Oxford University Press.

———. 2016a. 'Mature Information Societies—a Matter of Expectations'. *Philosophy & Technology* 29 (1): 1–4. https://doi.org/10.1007/s13347-016-0214-6.

———. 2016b. ‘Faultless Responsibility: On the Nature and Allocation of Moral Responsibility for Distributed Moral Actions’. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374 (2083): 20160112. https://doi.org/10.1098/rsta.2016.0112.

Floridi, Luciano, Josh Cowls, Thomas C. King, and Mariarosaria Taddeo. 2020. ‘How to Design AI for Social Good: Seven Essential Factors’. *Science and Engineering Ethics* 26 (3): 1771–96. https://doi.org/10.1007/s11948-020-00213-5.

Floridi, Luciano, and Mariarosaria Taddeo. 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.

Gu, Tianyu, Brendan Dolan-Gavitt, and Siddharth Garg. 2017. ‘BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain’. *ArXiv:1708.06733 [Cs]*, August. http://arxiv.org/abs/1708.06733.

High Level Expert Group on Artificial Intelligence. 2019. ‘Ethics Guideline for Trustworthy AI’. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

IEEE. 2017. ‘Artificial Intelligence and Machine Learning Applied to Cybersecurity’. https://www.ieee.org/about/industry/confluence/feedback.html.

Jagielski, Matthew, Alina Oprea, Battista Biggio, Chang Liu, Cristina Nita-Rotaru, and Bo Li. 2018. ‘Manipulating Machine Learning: Poisoning Attacks and Countermeasures for Regression Learning’. *ArXiv:1804.00308 [Cs]*, April. http://arxiv.org/abs/1804.00308.

Liao, Cong, Haoti Zhong, Anna Squicciarini, Sencun Zhu, and David Miller. 2018. ‘Backdoor Embedding in Convolutional Neural Network Models via Invisible Perturbation’. *ArXiv:1808.10307 [Cs, Stat]*, August. http://arxiv.org/abs/1808.10307.

Luhmann, N. 1979. *Trust and Power : Two Works*. Chichester; New York: Wiley.

Mirsky, Yisroel, Tom Mahler, Ilan Shelef, and Yuval Elovici. 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.

Primiero, Giuseppe, and Mariarosaria Taddeo. 2012. ‘A Modal Type Theory for Formalizing Trusted Communications’. *Journal of Applied Logic* 10 (1): 92–114. https://doi.org/10.1016/j.jal.2011.12.002.

Sharif, Mahmood, Sruti Bhagavatula, Lujo Bauer, and Michael K. Reiter. 2016. ‘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*, 1528–40. Vienna, Austria: ACM Press. https://doi.org/10.1145/2976749.2978392.

Sinha, Aman, Hongseok Namkoong, and John Duchi. 2017. ‘Certifying Some Distributional Robustness with Principled Adversarial Training’. *ArXiv:1710.10571 [Cs, Stat]*, October. http://arxiv.org/abs/1710.10571.

Szegedy, Christian, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. ‘Intriguing Properties of Neural Networks’. *ArXiv:1312.6199 [Cs]*, December. http://arxiv.org/abs/1312.6199.

Taddeo, Mariarosaria. 2009. 'Defining Trust and E-Trust: From Old Theories to New Problems'. Article. International Journal of Technology and Human Interaction (IJTHI). 1 April 2009. www.igi-global.com/article/defining-trust-trust/2939.

———. 2010a. 'Modelling Trust in Artificial Agents, A First Step Toward the Analysis of e-Trust'. *Minds and Machines* 20 (2): 243–57. https://doi.org/10.1007/s11023-010-9201-3.

———. 2010b. 'An Information-based Solution for the Puzzle of Testimony and Trust'. *Social Epistemology* 24 (4): 285–99. https://doi.org/10.1080/02691728.2010.521863.

———. 2013. 'Cyber Security and Individual Rights, Striking the Right Balance'. *Philosophy & Technology* 26 (4): 353–56. https://doi.org/10.1007/s13347-013-0140-9.

———. 2014. 'The Struggle Between Liberties and Authorities in the Information Age'. *Science and Engineering Ethics*, September, 1–14. https://doi.org/10.1007/s11948-014-9586-0.

———. 2017a. 'The Limits of Deterrence Theory in Cyberspace'. *Philosophy & Technology*. https://doi.org/10.1007/s13347-017-0290-2.

———. 2017b. 'Trusting Digital Technologies Correctly'. *Minds and Machines*, November. https://doi.org/10.1007/s11023-017-9450-5.

———. 2019. 'Three Ethical Challenges of Applications of Artificial Intelligence in Cybersecurity'. *Minds and Machines* 29 (2): 187–91. https://doi.org/10.1007/s11023-019-09504-8.

Taddeo, Mariarosaria, and Luciano Floridi. 2018a. 'Regulate Artificial Intelligence to Avert Cyber Arms Race'. *Nature* 556 (7701): 296–98. https://doi.org/10.1038/d41586-018-04602-6.

———. 2018b. 'How AI Can Be a Force for Good': *Science* 361 (6404): 751–52. https://doi.org/10.1126/science.aat5991.

Taddeo, Mariarosaria, Tom McCutcheon, and Luciano Floridi. 2019. 'Trusting Artificial Intelligence in Cybersecurity Is a Double-Edged Sword'. *Nature Machine Intelligence* 1 (12): 557–60. https://doi.org/10.1038/s42256-019-0109-1.

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

Uesato, Jonathan, Brendan O'Donoghue, Aaron van den Oord, and Pushmeet Kohli. 2018. 'Adversarial Risk and the Dangers of Evaluating Against Weak Attacks'. *ArXiv:1802.05666 [Cs, Stat]*, February. http://arxiv.org/abs/1802.05666.

Yang, Guang-Zhong, Jim Bellingham, Pierre E. Dupont, Peer Fischer, Luciano Floridi, Robert Full, Neil Jacobstein, et al. 2018a. 'The Grand Challenges of Science Robotics'. *Science Robotics* 3 (14): eaar7650. https://doi.org/10.1126/scirobotics.aar7650.

———. 2018b. 'The Grand Challenges of *Science Robotics*'. *Science Robotics* 3 (14): eaar7650. https://doi.org/10.1126/scirobotics.aar7650.

Zhuge, Jianwei, Thorsten Holz, Xinhui Han, Chengyu Song, and Wei Zou. 2007. 'Collecting Autonomous Spreading Malware Using High-Interaction Honeypots'. In *Information and Communications Security*, edited by Sihan Qing, Hideki Imai, and Guilin Wang, 438–51. Lecture Notes in Computer Science. Springer Berlin Heidelberg.