GCSP
Geneva Centre for Security Policy

JUNE 2022 | ISSUE 25

Strategic Security Analysis

Securing AI-based Security Systems

Sandra Scott-Hayward

www.gcsp.ch

Key Points

- Fundamental weaknesses of AI include brittleness, embedded bias, catastrophic forgetting and lack of explainability.
- Although research is under way to address some of these issues, the adoption of AI techniques and models in security systems exposes potentially critical security systems to weaknesses/vulnerabilities such as these.
- Adversarial training is one strongly recommended approach to increase the robustness (i.e. reduce the brittleness) of the AI model. In this approach, the training dataset is extended to include *adversarial examples* representative of potential attacks on the system. However, the implementation of adversarial training is currently ad hoc.
- Given the evidence of AI weaknesses, the omission of adversarial training and similar hardening techniques for AI-based security systems is unacceptable. Standardised testing and evaluation of AI-based security systems is recommended. From a governance perspective, evidence of adversarial robustness evaluation should be a minimum requirement for the acceptance of an AI-based security system.
- The production of strong adversarial samples does not account for "black swan" events, i.e. random and unexpected events that have an extreme impact. Given that security systems tend to be designed to detect "old" or "known" types of attack, ways need to be found to manage the occurrence of "new" attacks.

About the author

Sandra Scott-Hayward is a GCSP Polymath Fellow and a Senior Lecturer (Associate Professor) with the School of Electronics, Electrical Engineering and Computer Science, and a Member of the Centre for Secure Information Technologies at Queen's University Belfast (QUB). She is also the Director of the QUB Academic Centre of Excellence in Cyber Security Education.

About this publication

This publication is part of a special series of Strategic Security Analysis under the Polymath Initiative supported by the Didier and Martine Primat Foundation. For more information, please visit the Polymath Initiative website: https://www.gcsp.ch/the-polymath-initiative

Introduction

Together with the innovations in systems and services based on advances in artificial intelligence (AI), the vulnerabilities associated with its increasing use in a broad range of areas have been reported. Of these vulnerabilities, embedded bias or algorithmic discrimination is well recognised, such as racial and gender biases in algorithmic tools used for recruitment decision-making, criminal risk assessment, health-care resource allocation, etc. To address the issue of embedded bias, steps have been proposed such as identifying the algorithms used, understanding the target of the solution (e.g. considering the diversity and representativeness of end users and/or subjects in the data), assessing performance toward that goal (e.g. testing for specific target groups or cases of problematic use), retraining based on the performance assessment and introducing oversight bodies.

Beyond bias, AI systems are also recognised to suffer from brittleness, catastrophic forgetting and lack of explainability.

Beyond bias, AI systems are also recognised to suffer from brittleness (the inability to generalise or adapt to conditions outside a narrow set of assumptions),[1] catastrophic forgetting (when a model has to process new data and can no longer classify the old data),[2] and lack of explainability (the absence of details and reasons given by a model to make its functioning clear or easy to understand).[3]

This GCSP Strategic Security Analysis paper addresses the question of AI robustness[4] when AI techniques and models are adopted in security systems. Robustness refers to the reliable operation of a system across a range of conditions (including attacks). Firstly, the distinction between AI and machine learning (ML) is highlighted, with reference to the *Artificial Intelligence and UK National Security* report.[5] Whereas "general AI" refers to machine intelligence with the agency, reasoning and adaptability of a human brain, "narrow AI" refers to machine intelligence trained to perform narrowly defined cognitive tasks, such as playing chess, driving a car or translating documents. This paper addresses "narrow AI", for which the terms AI and ML are used interchangeably.

Secure machine learning

In Europe, AI systems come under the scope of the EU Cybersecurity Act (2019), which introduces an EU-wide cyber security certification framework[6] for ICT products, services and processes.[7] As part of this effort, the 2021 EU Agency for Cybersecurity (ENISA) *Securing Machine Learning Algorithms* report[8] maps out a suite of security controls appropriate for ML algorithms. These include conventional information system security controls such as compliance with data security requirements, identity management, authentication and access control. The importance of maintaining security levels for ML systems at a standard that we should expect from traditional information systems should not be underestimated. We can learn a lesson here from the proliferation of Internet of Things (IoT) devices with inadequate basic security controls. It has been a struggle to retrospectively require security controls in the IoT domain with the UK code of practice for consumer IoT security only issued in 2018[9] and the European standard on connected device security following in June 2020[10], long after the widespread deployment of IoT devices.

The importance of maintaining security levels for ML systems at a standard that we should expect from traditional information systems should not be underestimated.

However, while established technical, organisational and policy security controls can be applied to mitigate some of the vulnerabilities of ML algorithms, additional security controls specific to ML systems will be required. For example, as highlighted in the ENISA report, one of these ML-specific security controls could be the requirement to "include adversarial examples to training datasets".[11] This issue and the proposed security control are discussed in the following section.

Adversarial machine learning

Adversarial machine learning (AML) is concerned with the design of ML algorithms that can resist security challenges, the study of the capabilities of attackers, and understanding the consequences of an attack.[12] The U.S. National Cybersecurity Center of Excellence describes AML as "the process of extracting information about the behaviour and characteristics of an ML system and/or learning how to manipulate the inputs into an ML system in order to obtain a preferred outcome".[13]

The term "adversarial attack" refers to an attack crafted by an adversary targeting a learning system in order to cause its malfunction. There are two main approaches: (1) targeting the training phase, e.g. through "poisoning", which is an adversary's corruption of the ML system's training data (in practice, the attacker might add false data to the original training data, but with an incorrect label such as adding a malicious software sample that is labelled as benign), thus reducing the likelihood of an attack being detected; and (2) targeting the testing (inference) phase, e.g. through "evasion", which allows attackers to evade detection by manipulating (making small perturbations/modifications to) input samples to cause misclassification at test time. In practice, in an evasion attack the attacker might add some noise to an image to cause the ML system to misclassify the image, e.g. from that of a train to that of a duck.

Vulnerability of AI-based security systems

Given that the objective of a security system is to increase security, it is concerning that the adoption of AI techniques and models in such systems exposes potentially critical security systems to vulnerabilities. Essentially, the introduction of AI components to a system changes its cyber security by expanding the attack surface with each element in the AI processing chain exposed to threats (e.g. data exposed to poisoning or the classification system exposed to an evasion attack). A detailed discussion of AI's security vulnerability is presented in Comiter's paper on "Attacking AI".[14] However, Comiter's recommendation to "improve intrusion detection systems to better detect when assets have been compromised and to detect patterns of behavior indicative of an adversary formulating an attack" fails to recognise that these systems themselves are vulnerable to an adversarial attack.

Given that the objective of a security system is to increase security, it is concerning that the adoption of AI techniques and models in such systems exposes potentially critical security systems to vulnerabilities.

These vulnerabilities are perhaps best illustrated with some examples.

One of the key security mechanisms in communication networks is the network intrusion detection system (NIDS). ML-based NIDSs are currently increasingly used to analyse large volumes of network traffic and detect previously unseen network attacks. Even with no direct access to the NIDS or its ML model (i.e. a black-box attack), an adversary could craft an evasion attack by generating network traffic with perturbed/modified features to resemble those of benign traffic (the ML model uses features of the data to determine whether there is an attack or not). In this way, the adversary evades detection by the NIDS. The vulnerability of Distributed Denial-of-Service (DDoS) detection systems to adversarial attack has been variously demonstrated.[15] DDoS attacks are among the most severe threats to the security of Internet-connected systems. For example, the 2016 Mirai botnet infected nearly 65,000 IoT devices and overwhelmed websites, game servers, telecoms and anti-DDoS providers with massive DDoS attacks. Today, the largest attacks achieve several terabits per second and millions of packets per second, which can saturate communication links and exhaust network and online service resources. Their daily occurrence has a heavy impact on Internet service providers, financial institutions, retail services and supply chains.[16] The implication of adversarial attacks on ML-based DDoS detection systems is that the very systems designed to protect the network from DDoS attack are themselves vulnerable to attack. Unless these security systems are robustly designed and evaluated, the risk of DDoS to Internet-connected systems will not be mitigated.

Facial recognition technology (FRT) is a type of biometric security that has been widely adopted for the identification of individuals. The uses of FRT range from individual access control systems (e.g. to confirm an employee's identity to enter a secured workplace) to national and international security applications. The rapid adoption of FRT systems for both private and public use has encountered pushback in recent years due to controversy regarding both the collection of images used in FRT systems[17] and remote use of FRT systems in public places[18] without the consent of those whose images have been collected or used. Some of the issues relate to embedded bias, as previously discussed. There are multiple examples of the inaccuracy of FRT in identifying women and people of colour.[19] In law-enforcement applications, this has led to erroneous identification and false arrests.

FRT compares an individual's facial features to available images for verification or identification purposes. The technology is based on ML image classification, and studies have proved the viability of poisoning and evasion attacks against image classification.[20] The possible consequences of an adversarial attack on an FRT system pose a significant security threat. The verification or identification processes could be undermined, enabling fraudulent access to systems. Taking the adversarial attack one step further, we encounter the concept of deepfakes (deep learning + fakes) that use deep learning techniques to create visual and audio content with a high potential to deceive.[21] Deepfakes are increasingly realistic and credible, placing pressure on systems that rely on the veracity of an image, such as an FRT system. Deepfakes have already been used in spear phishing (a targeted technique to steal information or money or to infect a target's device).[22] Furthermore, similar to the monetisation of DDoS attacks through the sale of off-the-shelf DDoS-as-a-service tools to those with limited know-how to launch DDoS attacks, the proliferation of deepfake-as-a-service[23] can be expected to increasingly monetise the use of deepfakes.

A standardisation approach to hardening AI-based security systems is specifically recommended.

As with other security systems, it is inevitable that advances in AI will be applied in military systems. The development and use of autonomous weapons systems (AWSs) is one example of this. Issues related to emerging technologies in the areas of LAWS (lethal AWSs) in the context of the objectives and purposes of the UN Convention on Certain Conventional Weapons are examined through the open-ended Group of Governmental Experts on emerging technologies in the area of LAWS. The vulnerabilities of AI are no different in the AWSs application. However, the consequences of the exploitation of such vulnerabilities could be much greater. Consider the potential impact of an adversarial attack leading to misclassification and incorrect "target" identification.

A particular issue regarding an adversarial attack is the difficulty of detecting it. In the example of the NIDS (see above), an evasion attack in which the adversary disguises an intrusion as benign traffic could delay the detection of the intrusion, thus enabling the attacker to move through the network and inflict further damage. Lateral movement has been automated and used in various ransomware attacks and exploits, e.g. WannaCry, EternalBlue and NotPetya.

To achieve their security goals, AI-based security systems must be highly robust. As previously noted, Comiter outlines the security vulnerability of AI-based systems.[24] Content filters, the military, law enforcement, traditionally human-based tasks being replaced by AI, and civil society are highlighted as attractive targets for attack. AI security compliance programmes are proposed to protect against AI attacks modelled on compliance programmes in other industries, such as payment card industry compliance for securing payment transactions. It is further proposed that regulators should mandate compliance for government and high-risk uses of AI.

In agreement with these recommendations, a standardisation approach to hardening AI-based security systems is specifically recommended.

Adversarially robust ML models: a standardisation approach

There is extensive research on the topic of achieving adversarially robust ML models[25] (which is also described as hardening). Techniques include data sanitisation to prevent data poisoning, using ensemble methods to combine the results of multiple classifiers in decision-making, training the model with features that cannot be manipulated by an attacker, and adversarial training (training against known adversarial examples[26]) to generate more robust classifiers.

In terms of the adversarial training approach, the training dataset is augmented with inputs containing adversarial perturbations/modifications, but with correct output labels to minimise classification errors regarding adversarial examples. One of the problems (which also applies to the development of all ML models) is whether the proposed adversarially robust model is, in fact, robust, or if it only demonstrates robustness with respect to selected adversarial attacks/examples. With RobustBench, Croce et al. attempt to address this issue with a benchmarking approach aimed at *standardised* adversarial robustness evaluation.[27] This builds on earlier adversarial attack libraries and benchmarking tools[28] with the goal of offering an honest, worst-case robustness evaluation based on strong, standardised attacks, with transparency regarding the reliability of the evaluation, and an open system that can expand to include new defences and evaluations using adaptive attacks. Although RobustBench focuses on image classification, the goals of the benchmarking framework are equally applicable to other domains such as the security systems described previously in this paper.

Standardised evaluation requires agreed datasets, threat models, evaluation techniques and metrics.

Standardised evaluation requires agreed datasets, threat models, evaluation techniques and metrics.[29] A group of 12 industry and academic research groups have partnered to develop the Adversarial ML Threat Matrix, ATLAS.[30] The goal is a knowledge base of how ML systems can be attacked. The structuring in the style of the MITRE ATT&CK matrix supports a recognised cyber security framework of adversary tactics and techniques.[31] From a security perspective, this approach feeds into benchmarking using a suite of defined threat models. Furthermore, the case studies offer example implementations for raising awareness of the threats to ML systems. More research is required to develop the appropriate benchmarking for security systems.

In its 2020 technical report on *Robustness and Explainability of Artificial Intelligence*, the European Commission's science and knowledge service, the Joint Research Centre (JRC) presents a policy-oriented description of the current perspectives of AI and its implications in society, along with a technical discussion of the current risks (and their mitigations) associated with AI focused on the aspects of robustness and explainability.[32] The authors' conclusions specifically recommend the establishment of good practices through "introducing standardized methodologies to assess the robustness of AI models, in particular to determine their field of action with respect to the data that have been used for the training, the type of mathematical model, or the context of use, amongst other factors".

Given the evidence of weaknesses of AI, in the first instance, consideration should be given as to whether AI is appropriate for the security solution.[33] Following such a risk assessment and the decision to proceed with an AI-based security system, there must be standardised testing and evaluation of AI-based security systems. By way of motivation, it can be noted that in its 2021 report on *Face Recognition Vendor Testing*,

the National Institute of Standards and Technology (NIST) found that many FRT systems functioned without any racial or gender bias, indicating that fair and unbiased FRT systems are possible with sufficient effort and testing.[34]

As well as conforming with the requirements of established information system security controls (technical, organisational and policy), from a governance perspective, evidence of adversarial robustness evaluation should be a minimum requirement for accepting an AI-based security system.

Education, awareness and responsibilities

As well as conforming with the requirements of established information system security controls, from a governance perspective, evidence of adversarial robustness evaluation should be a minimum requirement for accepting an AI-based security system.

The Organisation for Economic Co-operation and Development (OECD) AI Principles promote the innovative and trustworthy use of AI.[35] OECD AI Principle 1.4 states that "AI systems should be robust, secure and safe throughout their entire lifecycle so that, in conditions of normal use, foreseeable use or misuse, or other adverse conditions, they function appropriately and do not pose unreasonable safety risk". This principle reflecting robustness, security and safety is also identified in the Microsoft Responsible AI Principles[36] and in the Alan Turing Institute's FAST Track Principles.[37] While the general notion of risk analysis applies to gauge the robustness and resilience of the system, the specific risk of adversarial attack is emphasised with the recommendation to employ model-hardening techniques to mitigate the risk.

To fulfil these principles and to achieve our stated goal of standardised testing and evaluation, the community must be aware of them. This is recognised in the OECD AI Principles: "AI actors should ensure traceability, including in relation to datasets, processes and decisions made during the AI system lifecycle ... appropriate to the context and consistent with the state of the art"[38] and in the JRC report recommendation to "[raise] awareness among AI practitioners through the publication of good practices regarding to known vulnerabilities of AI models and technical solutions to address them".[39]

We have an opportunity to do this through our education systems and industry-focused training programmes. In our teaching and training, we should develop awareness in our students of not only the scientific and technical aspects of AI, but also the importance of applying economic, legal, social, ethical and environmental considerations to issues surfaced by AI technologies. Awareness is the first of the principles[40] on algorithmic transparency and accountability set out by the Association for Computing Machinery US Public Policy Council.[41] Higher education programme accreditation bodies already require this approach.[42] Unfortunately, both students and academics can undervalue the importance of these aspects of engineering and computer science.

Responsibility for robust, secure, and safe AI-based security systems spans the system lifecycle from designers and developers, through users and deployers of the systems, to those evaluating and governing them. The "Assessment List for Trustworthy AI for Self-Assessment" details questions to be considered in the self-evaluation of the achievement of technical robustness and safety that can help meet this responsibility.[43]

"Black swan" events

A "black swan" event is a random and unexpected phenomenon with an extreme impact. The terrorist strikes on 11 September 2001 and the 2008 financial crisis are examples of such events. In the context of security systems, we generally refer to previously unseen exploits or attacks as zero-day attacks.[44] For example, a new type of malware or malicious software that neither matches an existing malware signature nor presents a pattern resembling an existing class of malware would be described as zero-day malware. It would not have been previously known about or anticipated. An example of this is the June 2017 NotPetya cyber attack. This attack (using the EternalBlue and EternalRomance exploits) began with the compromise of a legitimate software application. A malicious data encryption tool was inserted into the software, and when organisations updated the application, the Petya code was initiated. Once an organisation's machine was infected, the encryption tool was designed to spread rapidly across the targeted enterprise, rebooting and starting the encryption process. The malware was not designed to be decrypted, i.e. there was no way for victims to recover data once it had been encrypted. The consequence of this zero-day attack has been millions of dollars in losses to companies such as Maersk, FedEx and Merck, and incalculable impact due to data loss on, for example, hospitals and health-care institutions.

In the evolution of AI-based security systems, we should recognise and acknowledge uncertainty.

When designing and developing AI-based security systems, the testing and evaluation process (as described previously in this paper) determines the system's capability to detect such previously unseen attacks. This is usually evaluated by using a test dataset on which the system has not been trained. However, this approach does not account for black swans. Current AI-based security systems are predominantly designed to detect "old" attacks, i.e. those that match (at least some of) the pattern of a previously categorised attack. Using adversarial training increases the attack detection capability, but is still tuned to the original attack. How should we prepare for the detection of "new" attacks?

In the evolution of AI-based security systems, we should recognise and acknowledge uncertainty. In the research towards adaptive systems in the image of 'general AI', adaptability may enable us to better detect 'new' attacks. Such considerations should be taken account of in the design and evolution of benchmarking. This will not change the nature of a black swan event but may reduce the risk of cyber security black swans. Given the impact of AI systems on cyber security, this is of paramount importance.

Conclusions

The vulnerabilities that affect AI techniques strongly impact the robustness of AI-based systems. The concern is that such systems can potentially exhibit uncontrolled behaviour; enable malicious or adversarial actors to mislead them, thus reducing their effectiveness; or cause a malfunction that disrupts their operation. This is of particular concern for AI-based security systems. Whether an evasion attack leads to a network intrusion going undetected, a deepfake leads to fraud or the violation of an access control, or misclassification leads to the incorrect identification of an AI-controlled weapon's "target", these are all unacceptable outcomes.

One approach to reduce the likelihood of such outcomes is to introduce standardised adversarial robustness benchmarking. This would involve required testing and evaluation based on agreed datasets, threat models, evaluation techniques and metrics. Responsibility for the design, deployment, and maintenance of robust, secure, and safe AI-based security systems will lie with all stakeholders, ranging from the educators teaching and training in AI system design, through developers building AI systems, to policymakers and governors specifying, selecting, or promoting the adoption and use of AI-based systems.

Responsibility for the design, deployment, and maintenance of robust, secure, and safe AI-based security systems will lie with all stakeholders.

Endnotes

1. M.L. Cummings, *The Surprising Brittleness of AI*, Technical Report, Women Corporate Directors, 2020, https://www.womencorporatedirectors.org/WCD/News/JAN-Feb2020/Reality%20Light.pdf.

2. J. Kirkpatrick et al., "Overcoming Catastrophic Forgetting in Neural Networks", *Proceedings of the National Academy of Sciences*, Vol.114(13), 2017, pp.3521-3526.

3. A.B. Arrieta et al., "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI", *Information Fusion*, Vol.58, 2020, pp.82-115.

4. Robustness is one of the three components of trustworthy AI as identified by the European Commission High-Level Expert Group on AI.

5. A. Babuta et al., *Artificial Intelligence and UK National Security: Policy Considerations*, 2020, https://rusi.org/explore-our-research/publications/occasional-papers/artificial-intelligence-and-uk-national-security-policy-considerations.

6. In 2020 the Commission and the European Union (EU) Agency for Cybersecurity created the Stakeholders Cybersecurity Certification Group to advise them on strategic issues regarding cyber security certification. Its aim, arising from the 2019 EU Cybersecurity Act, is to create market-driven certification schemes and help to reduce fragmentation among various existing schemes in EU member states.

7. European Parliament and the Council, "Regulation (EU) 2019/881 of the European Parliament and of the Council of 17 April 2019 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, https://eur-lex.europa.eu/eli/reg/2019/881/oj.

8. ENISA (EU Agency for Cybersecurity), *Securing Machine Learning Algorithms*, December 2021, https://www.enisa.europa.eu/publications/securing-machine-learning-algorithms.

9. UK Department of Digital, Culture, Media and Sport, *Code of Practice for Consumer Internet of Things (IoT) Security*, 2018, https://www.gov.uk/government/publications/code-of-practice-for-consumer-iot-security.

10. ETSI, *Cyber Security for Consumer Internet of Things: Baseline Requirements*, ETSI EN 303 645, June 2020. In the United States the National Institute of Standards and Technology also released guidance for IoT device manufacturers in 2020.

11. ENISA, 2021.

12. L. Huang et al., "Adversarial Machine Learning", *Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence*, 2011, pp.43-58, https://people.eecs.berkeley.edu/~tygar/papers/SML2/Adversarial_AISEC.pdf.

13. NCCoE (National Cybersecurity Center of Excellence), *Artificial Intelligence: Adversarial Machine Learning*, n.d., https://www.nccoe.nist.gov/ai/adversarial-machine-learning.

14. M. Comiter, "Attacking Artificial Intelligence: AI's Security Vulnerability and What Policymakers Can Do About It", Belfer Center Paper, August 2019, https://www.belfercenter.org/sites/default/files/2019-08/AttackingAI/AttackingAI.pdf.

15. See J. Aiken and S. Scott-Hayward. "Investigating Adversarial Attacks against Network Intrusion Detection Systems in SDNs", *IEEE Conference on Network Function Virtualization and Software Defined Networks (NFV-SDNs)*, 2019, pp.1-7, https://pure.qub.ac.uk/en/publications/investigating-adversarial-attacks-against-network-intrusion-detec; M. Abdelaty et al., "GaDoT: GAN-based Adversarial Training for Robust DDoS Attack Detection", Ninth IEEE Conference on Communications and Network Security, 2021, https://pure.qub.ac.uk/en/publications/gadot-gan-based-adversarial-training-for-robust-ddos-attack-detec.

16. O. Yoachimik and V. Ganti, "DDoS Attack Trends for Q4 2021", 2022, https://blog.cloudflare.com/ddos-attack-trends-for-2021-q4/; S. Hilton, "Dyn Analysis Summary of Friday October 21 Attack", Oracle Dyn, 2016, https://dyn.com/blog/ dyn-analysis-summary-of-friday-october-21-attack/.

17. ICO (Information Commissioner's Office), "ICO Issues Provisional View to Fine Clearview AI Inc over £17 Million", November 2021, https://ico.org.uk/about-the-ico/news-and-events/news-and-blogs/2021/11/ico-issues-provisional-view-to-fine-clearview-ai-inc-over-17-million/.

18. ICO, "ICO Investigation into how the Police Use Facial Recognition Technology in Public Places", October 2019, https://cy.ico.org.uk/media/about-the-ico/documents/2616185/live-frt-law-enforcement-report-20191031.pdf.

19. P. Grother et al., *Face Recognition Vendor Test (FRVT): Part 3: Demographic Effects*, NISTIR 8280, 2019, https://nvlpubs.nist.gov/nistpubs/ir/2019/NIST.IR.8280.pdf.

20. H. Kwon et al., "Multi-targeted Adversarial Example in Evasion Attack on Deep Neural Network", *IEEE Access*, Vol.6, 2018, pp.46084–46096.

21. J. Kietzmann et al., "Deepfakes: Trick or Treat?" *Business Horizons*, Vol.63(2), 2020, pp.135-146.

22. In 2021, a bank manager received a call from a company director whose voice the manager (falsely) recognised. The call was a request to transfer funds for an acquisition. The manager agreed.

23. https://deepfakesweb.com/; https://github.com/iperov/DeepFaceLab.

24. Comiter, 2019.

25. N. Carlini, "A Complete List of All (arXiv) Adversarial Example Papers", 15 June 2019, https://nicholas.carlini.com/writing/2019/all-adversarial-example-papers.html.

26. An adversarial example is an "ML input sample formed by applying a small but intentionally worst-case perturbation to a clean example, such that the perturbed input causes a learned model to output an incorrect answer" (NCCoE, n.d.).

27. F. Croce et al., "RobustBench: A Standardized Adversarial Robustness Benchmark", arXiv preprint arXiv:2010.09670, 2020, https://arxiv.org/pdf/2010.09670.pdf?ref=https://githubhelp.com.

28. N. Papernot et al., *Technical Report on the Cleverhans v2.1.0 Adversarial Examples Library*, arXiv preprint arXiv:1610.00768, 2016; Y. Dong et al., "Benchmarking Adversarial Robustness on Image Classification", *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp.321-331, https://www.researchgate.net/publication/343455781_Benchmarking_Adversarial_Robustness_on_Image_Classification.

29. It is worth noting that this approach of using shared models/datasets can also raise security issues, e.g. a single point of failure in the case of multiple systems relying on a poisoned dataset, as outlined in Comiter, 2019. However, the "secret" model/dataset approach assuming "security by obscurity" should never be the only security mechanism.

30. MITRE ATLAS™, "Adversarial Threat Landscape for Artificial-Intelligence Systems", n.d., https://atlas.mitre.org/.

31. MITRE ATT&CK®, n.d., https://attack.mitre.org/.

32. R. Hamon et al., *Robustness and Explainability of Artificial Intelligence*, Publications Office of the European Union, 2020, https://publications.jrc.ec.europa.eu/repository/handle/JRC119336.

33. NIST (National Institute of Standards and Technology), "AI Risk Management Framework Concept Paper", 2021, https://www.nist.gov/system/files/documents/2021/12/14/AI%20RMF%20Concept%20Paper_13Dec2021_posted.pdf.

34. P. Grother et al., *Face Recognition Vendor Test (FRVT): Part 7: Identification for Paperless Travel and Immigration*, NISTIR 8381, 2021, https://nvlpubs.nist.gov/nistpubs/ir/2021/NIST.IR.8381.pdf.

35. OECD.AI Policy Observatory, "OECD AI Principles Overview", n.d., https://oecd.ai/en/ai-principles.

36. Microsoft, "Microsoft Responsible AI Principles in Practice", 2022, https://www.microsoft.com/en-gb/ai/responsible-ai?activetab=pivot1%3aprimaryr6.

37. D. Leslie, "Understanding Artificial Intelligence Ethics and Safety: A Guide for the Responsible Design and Implementation of AI Systems in the Public Sector", 2019, https://www.turing.ac.uk/sites/default/files/2019-08/understanding_artificial_intelligence_ethics_and_safety.pdf.

38. OECD.AI Policy Observatory, n.d.

39. Hamon et al., 2020.

40. The publication of model validation and testing results is also a principle.

41. ACM (Association for Computing Machinery) US Public Policy Council, "Statement on Algorithmic Transparency and Accountability", 2017, https://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf.

42. Engineering Council, *Accreditation of Higher Education Programmes*, 2004, https://www.theiet.org/media/1775/accreditation-of-higher-education-programmes-third-edition.pdf.

43. AI HLEG (High-Level Assessment Group on AI), "Assessment List for Trustworthy AI for Self-Assessment", 2019, https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment.

44. "An exploit is a piece of software, a chunk of data, or a sequence of commands that takes advantage of a bug or vulnerability to cause unintended or unanticipated behavior to occur on computer software, hardware, or something electronic" (https://en.wikipedia.org/wiki/Exploit_(computer_security)).

GCSP
Geneva Centre for Security Policy

Where knowledge meets experience

The GCSP Strategic Security Analysis series are short papers that address a current security issue. They provide background information about the theme, identify the main issues and challenges, and propose policy recommendations.

Geneva Centre for Security Policy
Maison de la paix
Chemin Eugène-Rigot 2D
P.O. Box 1295
CH - 1211 Geneva 1
Switzerland
Tel: + 41 22 730 96 00
Fax: + 41 22 730 96 49
e-mail: info@gcsp.ch
www.gcsp.ch

ISBN: 978-2-88947-311-3