CEPS Task Force Report

Artificial Intelligence and Cybersecurity

Technology, Governance and Policy Challenges

Rapporteurs
Lorenzo Pupillo
Stefano Fantin
Afonso Ferreira
Carolina Polito

CEPS

Artificial Intelligence and Cybersecurity

Technology, Governance and Policy Challenges

Final Report of a CEPS Task Force

Rapporteurs:

Lorenzo Pupillo

Stefano Fantin

Afonso Ferreira

Carolina Polito

Centre for European Policy Studies (CEPS)

Brussels

May 2021

The Centre for European Policy Studies (CEPS) is an independent policy research institute based in Brussels. Its mission is to produce sound analytical research leading to constructive solutions to the challenges facing Europe today.

Lorenzo Pupillo is CEPS Associate Senior Research Fellow and Head of the Cybersecurity@CEPS Initiative. Stefano Fantin is Legal Researcher at Center for IT and IP Law, KU Leuven. Afonso Ferreira is Directeur of Research at CNRS. Carolina Polito is CEPS Research Assistant at GRID unit, Cybersecurity@CEPS Initiative.

ISBN 978-94-6138-785-1

CEPS
Place du Congrès 1, B-1000 Brussels
Tel: 32 (0) 2 229.39.11
e-mail: info@ceps.eu
internet: www.ceps.eu

Contents

List of Figures

List of Tables

List of Abbreviations

AI	Artificial Intelligence
AIST	Artificial Intelligence for software testing
ANDES	Analysis of Dual Use Synergies
API	Application programming interface
BGP	Border Gateway Protocol
CAGR	Compound annual growth rate
CAICT	China Academy of Information and Communication Technology
CAPTCHA	Completely Automated Public Turing test to tell Computers and Humans Apart
CASE	Computer-aided software engineering
CIA	Confidentiality, integrity, availability
CJEU	European Court of Justice
CPS	Cyber-Physical Systems
CRISP-DM	Cross-Industry Standard Process for Data Mining
CSA	Cybersecurity Act
CSIRTs	Computer Security Incidents Response Teams
CSO	Chief security officer
DARPA	Defense Advanced Research Projects Agency
DevOps	Development and Operations
DNS	Domain name system
DSP	Digital service provider
EDA	European Defence Agency
EMEA	Europe, Middle East, Africa
ENISA	European Union Agency for Cybersecurity
GAN	Generative adversarial network
GARD	Guaranteeing AI Robustness against Deception
GDPR	General Data Protection Regulation
GPS	Global Positioning System
ICT	Information and Communications Technology
IEC	International Electrotechnical Commission
IEEE	Institute of Electrical and Electronics Engineers
IoT	Internet of Things
IP	Internet Protocol
ISO	International Organization for Standardization
JRC	European Joint Research Centre

LAWS	Lethal autonomous weapons systems
ML	Machine learning
NISD	Network and Information Security Directive
NIST	National Institute for Standards and Technology
OECD	Organisation for Economic Co-operation and Development
OESs	Operators of essential services
OMB	White House Office of Management and Budget
PAI	Partnership on AI
TAD	Threat and anomalies detection
TC260	National Information Security Standardization Technical Committee
TFEU	Treaty on the Functioning of the European Union
TTP	Tactics, techniques, and procedures
SAPPAN	Sharing and Automation for Privacy Preserving Attack Neutralization
SMEs	Small and medium-sized enterprises
SQL	Structured query language
SVM	Support-vector machine
V2V	Vehicle-to-vehicle

Preface

This report is based on discussions in CEPS' Task Force on Artificial Intelligence and Cybersecurity. The Task Force was composed of industry experts, representatives of EU and international institutions, academics, civil society organisations and practitioners (see list of participants in Annex II). The activity of the group started in autumn 2019, met on four separate occasions and continued online during the Covid-19 breakout, until March 2021.

As Coordinator of the Task Force, I would like to acknowledge the invaluable contributions of all the participants in this work. Particular thanks go to the members of the Advisory Board: Joanna Bryson at the Hertie School for Governance, Berlin, Mariarosaria Taddeo at the University of Oxford, Jean-Marc Rickli at the Geneva Centre for Security Policy and Marc Ph. Stoecklin at IBM Research Center, Zurich. I also wish to acknowledge the substantial work done by my fellow rapporteurs, Stefano Fantin, Afonso Ferreira and Carolina Polito. This work has been a collective endeavour and, as indicated in the text itself, other Task Force participants or members of the advisory board directly contributed their expertise by personally drafting selected sections of the report, namely Joanna Bryson, Mariarosaria Taddeo, Jean-Marc Rickli, David Clark, Rob Spiger, Chris Hutchins, Matt Walmsley, Sebastian Gerlach, and Matti Aksela. I am also grateful to members who kindly volunteered to review earlier versions of this report, especially Marc Radice, Carol Mezat, Alex Sangers, Rachel Azafrani, Philip Graefen, Giuseppe Giovanni Daquino, Federica Russo, Nineta Polemi, Wide Hogenhout and Miguel Gonzales-Sancho-Bodero. Thanks also go to the invited speakers who contributed to the Task Force discussions.

Lorenzo Pupillo, Coordinator and Rapporteur of the Task Force
Associate Senior Research Fellow and Head of the Cybersecurity@CEPS Initiative
CEPS Brussel, May 2021

Executive Summary

The Centre for European Policy Studies (CEPS) launched a Task Force on Artificial Intelligence (AI) and Cybersecurity in the autumn of 2019. The goal of this Task Force was to draw attention to the technical, ethical, market and governance challenges posed by the intersection of AI and cybersecurity. The Task Force, multistakeholder by design, was composed of seventeen private organisations, eight European Union (EU) institutions, one international and one multilateral organisation, five universities and think tanks, and two civil society organisations (see a list of participants in Annex II). Meeting on four separate occasions and continuing to work remotely when the Covid-19 lockdown started, the group explored ways to formulate practical guidelines for governments and businesses to ease the adoption of AI in cybersecurity in the EU while addressing the cybersecurity risks posed by the implementation of AI systems. These discussions led to policy recommendations being addressed to EU institutions, member states, the private sector and the research community for the development and deployment of secure AI systems.

AI is playing an increasingly central role in people's everyday lives. The benefits of implementing AI technology are numerous, but so are the challenges. The adoption of AI in cybersecurity could be hampered or even lead to significant problems for society if the security and ethical concerns are not properly addressed through governmental processes and policies. This report aims to contribute to EU efforts to establish a sound policy framework for AI. Its specific objectives are to:

- provide an overview of the current landscape of AI in terms of beneficial applications in the cybersecurity sector and the risks that stem from the likelihood of AI-enabled systems being subject to manipulation
- present the main ethical implications and policy issues related to the implementation of AI as they pertain to cybersecurity
- put forward constructive and concrete policy recommendations to ensure the AI rollout is securely adopted according to the objectives of the EU digital strategy.

The report raises several issues about policy implications. It suggests that, because of the lack of transparency and the learning abilities of AI systems, it is hard to evaluate whether a system will continue to behave as expected in any given context. Therefore, some form of control and human oversight is necessary. Furthermore, the point is made that AI systems, unlike brains, are designed, and so all the decisions based on these systems should be auditable. Talk about brains or consciousness has become rather a means to evade regulation and oversight. Poor cybersecurity in the protection of open-source models could also lead to hacking opportunities for actors seeking to steal such information. Limitations on the dissemination and the sharing of data and codes could therefore enable a more complete assessment of the related security risks. It should be noted that the overview is not exhaustive and other policy issues and ethical implications are raised throughout the report.

Policy recommendations from the Task Force

Based on an extensive review of the existing literature and the contributions from participants, the Task Force suggests the following recommendations to policymakers, the private sector, and the research community:

AI for cybersecurity

Specific EU policy measures that would ease the adoption of AI in cybersecurity in Europe include:

1. Enhancing collaboration between policymakers, the technical community and key corporate representatives to better investigate, prevent and mitigate potential malicious uses of AI in cybersecurity. This collaboration can be informed by the lessons learned in the regulation of cybersecurity, and from bioethics.
2. Enforcing and testing the security requirements for AI systems in public procurement policies. Adherence to ethical and safety principles should be regarded as a prerequisite for the procurement of AI applications in certain critical sectors. This would help to advance discussions on AI and safety in organisations, including at the board level.
3. Encouraging information sharing of cybersecurity-relevant data, for example data to 'train' models according to established best practice. Private sector-driven, cross-border information sharing should also be supported by providing incentives for cooperation and ensuring a governance framework that would enable legal certainty when exchanging data.
4. Focusing on supporting the reliability of AI, rather than its trustworthiness, in standards and certification methods. The following developing and monitoring practices are suggested to ensure reliability and mitigate the risks linked to the lack of predictability of AI systems' robustness:
    - Companies' in-house development of AI applications models and testing of data
    - Improving AI systems' robustness through adversarial training between AI systems
    - Parallel and dynamic monitoring or punctual checks of AI systems through a clone system as control, which would be used as a baseline comparison to assess the behaviour of the original system.
5. Supporting and internationally promoting proactive AI cybersecurity certification efforts, to be coordinated by ENISA. These should demand that assessment actions be taken prior to deployments and during the whole life cycle of a product, service, or process.
6. Envisaging appropriate limitations to the full openness policy for research output, such as algorithms or model parameters,[1] to enable a more complete assessment of the security risks related to the technology and its dissemination, balanced with the EU policy objective of fostering innovation.

[1] Models are often made public and 'open source' having successfully led to AI applications performing tasks with a broad general interest.

7. Promoting further study and regulatory interpretation of the General Data Protection Regulation (GRPR) provisions, even at the national level (for instance, with respect to Recitals 49 and 71, on data-sharing practices for information security aims), in the context of both AI for cybersecurity and applications aimed at making AI secure.
8. Addressing the challenges of adequately enforcing the personal data protection rules posed by datasets of mixed personal and non-personal data.
9. Evaluating how the use of AI systems in cybersecurity research and operations could be impacted by the current (and future) dual-use and export control regulatory framework;[2] drawing up clear rules that respect EU (treaty-based) values without hampering trade and sacrificing openness; establishing an EU-level regulated dual-use technology transfer mechanism, through the support of the industry and within the boundaries fixed by the Wassenaar Agreement, for defining a possible dual-use technology transfer mechanism and creating an avenue for developing a common approach among institutions dealing with dual-use technologies.
10. Enhancing the cooperation between military and civilian entities in AI-based development topics by applying capability development concepts from the military sector (which reflect strong cybersecurity requirements) to civilian AI applications, or by defining a reference architecture for cybersecurity specifically for AI applications, to be used in both civilian and military domains.
11. Addressing the skills shortage and uneven distribution of talents and professionals among market players. The public sector, as well as security-related agencies, should be ready to offer AI-related career paths and to train and retain cybersecurity skills and talents. The transformation of the cybersecurity sector should be monitored while ensuring that AI tools and their use are incorporated into existing cybersecurity professional practice and architectures.

Cybersecurity for AI

Ways to make AI systems safe and reliable when developing and deploying them include:

12. Promoting suitability testing before an AI system is implemented in order to evaluate the related security risks. Such tests, to be performed by all stakeholders involved in a development and/or a deployment project, should gauge value, ease of attack, damage, opportunity cost and alternatives.[3]
13. Encouraging companies to address the risk of AI attacks once the AI system is implemented. General AI safety could also be strengthened by putting detection mechanisms in place. These would alert companies that adversarial attacks are

[2] Wassenaar Agreement and European Commission Regulation No 428/2009.

[3] Some Task Force participants raised concerns about the feasibility of this requirement. A particular argument was that, given the fast pace of adoption of AI systems, innovation would be stifled if a suitability test were required for each and every AI system implemented.

occurring, that the system in question is no longer functioning within specified parameters in order to activate a fallback plan.[4]

14. Suggesting that AI systems follow a secure development life cycle, from ideation to deployment, including runtime monitoring and post-deployment control and auditing.
15. Strengthening AI security as it relates to maintaining accountability across intelligent systems, by requiring adequate documentation of the architecture of the system, including the design and documentation of its components and how they are integrated.[5] Strengthening measures include:
    - Securing logs related to the development/coding/training of the system: who changed what, when, and why? These are standard procedures applied for revision control systems used in developing software, which also preserve older versions of software so that differences and additions can be checked and reversed.
    - Providing cybersecure pedigrees for all software libraries linked to that code.
    - Providing cybersecure pedigrees for any data libraries used for training machine learning (ML) algorithms. This can also show compliance with privacy laws and other principles.
    - Keeping track of the data, model parameters, and training procedure where ML is used.
    - Requiring records that demonstrate due diligence when testing the technology, before releasing it. These would preferably include the test suites used so that they can be checked by the company itself or by third parties and then reused where possible.[6]
    - Maintaining logs of inputs and outputs for AI-powered operating systems, depending on the capacities of the system and when feasible, and assuming these are cybersecure and GDPR compliant.
    - Requiring in-depth logging of the AI system's processes and outcomes for life-critical applications such as automated aeroplanes, surgical robots, autonomous weapons systems, and facial recognition for surveillance purposes. For non-critical applications, the volume of input data should be evaluated before requiring an in-depth logging strategy. This is to avoid unfair competition between big and small players due to implementation costs.
    - Enhancing AI reliability and reproducibility by using techniques other than logging such as randomisation, noise prevention, defensive distillation, and ensemble learning.

[4] Some Task Force participants raised concerns about the maturity of AI technology, which at the current state of the art might not allow for effective detection mechanisms to be put in place.

[5] This should not be regarded as an exhaustive list of cybersecurity requirements for AI, for which further study will be required.

[6] Some Task Force participants raised concerns about the proportionality and intrusiveness of this requirement, especially in terms of compliance with the GDPR provisions.

16. Suggesting that organisations ensure models are fully auditable at time/point of failure, and to make the information available for subsequent analysis (e.g. analysis required by courts).[7] New methods of auditing systems should also be encouraged, such as restricting them to a trusted third party, rather than openly pushing datasets.

17. Suggesting that organisations develop an attack incident-response plan, and create a map showing how the compromise of one asset, dataset, or system affects other AI systems, for example how systems can exploit the same dataset or model once the attack has occurred. Policymakers should support the development and sharing of best practice. Validating data collection practices could guide companies in this process, for example in identifying potential weaknesses that could facilitate attacks or exacerbate the consequences of attacks.

[7] Some Task Force participants raised concerns about the feasibility and economic burden of this requirement.

PART I.
INTRODUCTION

1. Introduction

The Covid-19 pandemic is marking our lives in unprecedented ways. Since the outbreak in Wuhan, China in 2020, the virus has spread consistently and continuously across the globe. International organisations and scientists have increasingly started to apply new technologies such as Artificial Intelligence to track the pandemic, predict where the virus might appear and develop an effective response.

First, several institutions are using AI to assess and discover drugs or treatments that would help to treat Covid-19, and to develop prototype vaccines. AI has also been used to help detect whether people have new coronaviruses by identifying visual signs of Covid-19 on images from lung scans. It has monitored changes in body temperature through the use of wearable sensors and has provided open-source data platforms to track the spread of the disease.[8] In the early phase of the pandemic, DeepMind used its AlphaFold AI system to predict and publish protein structures associated with coronavirus.[9] Now that Pfizer, Moderna and AstraZeneca vaccines have been approved and are finally being administered across the globe, AI and other new technologies are also being deployed to manage this monumental effort. For example, the UK Medicines and Healthcare products Regulatory Agency (MHRA), in partnership with the UK unit of Genpact, the global professional services firm specialising in digital transformation, is using AI to track possible adverse effects of the vaccines on different population segments.

AI has been used in applications other than medical, too. It has helped in the fight against disinformation by mining social media, tracking down words that are sensational or alarming and identifying reliable and authoritative online references. AI applications have been adopted by several countries around the world to support the enforcement of lockdown measures, such as facial recognition systems to identify people not wearing masks or mobile applications tracking people's social contacts.

However, in the fight against Covid-19, AI has also revealed its inherent limitations. Current systems learn by finding patterns in data. To achieve the expected performance, systems must be trained with high-quality inputs that model desired behaviours. While this process has been successful in AI applications with staged situations and clear parameters, the process is much less predictable in real-life scenarios. Covid-19 is so new and complex, and the clinical and biological datasets needed to train AI systems are still scarce.[10]

Similar limitations in the use of AI have been observed in the financial world. March 2020 was the most volatile month in the history of the stock market. It is no surprise that the pandemic caused trillions of dollars to be wiped out in market capitalisation. The market shock, however, also hit dollar-neutral quant trading strategies (those that hold equally long and short

[8] European Parliamentary Research Service (2020), "What if we could fight coronavirus with Artificial Intelligence?", March.

[9] DeepMind (2020), "Computational predictions of protein structures associated with COVID-19", August (https://deepmind.com/research/open-source/computational-predictions-of-protein-structures-associated-with-COVID-19).

[10] N. Benaich (2020), "AI has disappointed on Covid", *Financial Times*, September (www.ft.com/content/0aafc2de-f46d-4646-acfd-4ed7a7f6feaa).

positions), even though most hedge funds were using AI to identify their portfolio composition.[11] In fact, quant funds that were using overly complex AI models may have suffered the most. The reason for AI's poor performance is that it is not matched for rare events like Covid-19; with few such shocks having ever occurred in the market, the system could not learn from past data.[12]

AI's role in the fight against Covid-19 is, therefore, two-edged. On the one hand, AI can support operators in their responses to this unprecedented health crisis. On the other hand, the inner limitations of these systems need to be considered and appropriately countered before they can be relied upon. This double-edged relationship between AI and Covid-19 can offer the reader a useful metaphor for understanding the interplay between AI and cybersecurity. As much as in the fight against the pandemic, AI can both empower and disrupt cybersecurity. In the case of the pandemic, shortcomings in the application of AI are mainly caused by the current unavailability of enough quality data. In the case of cybersecurity, however, the risks are inherently dependent on the way AI functions and learns and often result from the sophistication of the underlying AI technology. Overall, this report will argue, AI can substantially improve cybersecurity practices but can also facilitate new forms of attacks and further exacerbate security threats. The report will shed light on this dynamic and suggest which measures should be envisaged to counter these risks.

2. Where does Europe stand on the AI and cybersecurity interplay discussion?

The Joint Research Centre of the European Commission's report on AI in the European Union,[13] published in 2018, addressed different aspects of AI adoption, from an economic to a legal perspective, including cybersecurity. The report acknowledges the dual nature of AI and cybersecurity and the potential dangers to the security of the systems. Recognising that ML is often not robust against malicious attacks, it suggests that *"further research is needed in the field of adversarial ML to better understand the limitations in the robustness of ML algorithms and design effective strategies to mitigate these vulnerabilities."*[14]

On 19 February 2020, the European Commission published the White Paper on Artificial Intelligence. This outlined a strategy that aimed to foster an AI ecosystem in Europe. According to the White Paper, the EU will allocate funding that, combined with private resources, is expected to reach €20 billion per year. Moreover, it envisaged the creation of a network of centres of excellence to improve the EU digital infrastructure, and the development of mechanisms to allow small and medium-sized enterprises (SMEs) to better reimagine their business model to incorporate AI. Based on the recommendations of the High-Level Expert Group on AI, the EU also defined the fundamental requirements for AI implementation.

[11] Z. Kakushadze (2020), Quant Bust 2020, April.

[12] W. Knight (2020), "Even the Best AI Models Are No Match for the Coronavirus", *Wired*, July (www.wired.com/story/best-ai-models-no-match-coronavirus/).

[13] M. Craglia (ed.), A. Annoni, et. al. (2018), *Artificial Intelligence – A European Perspective,* EUR 29425 EN, Publications Office, Luxembourg.

[14] Ibid.

According to the White Paper the requirements for high-risk AI applications could consist of the following key features:

- training data
- data and record-keeping
- information to be provided
- robustness and accuracy
- human oversight
- specific requirements for specific AI applications, such as those used for remote biometric identification purposes.[15]

The AI White Paper contemplated the adoption of a flexible and agile regulatory framework limited to 'high-risk' applications, in sectors such as healthcare, transport, police and the judiciary. A follow-up Regulation to the White Paper on AI was published on 21 April 2021, after a public consultation that ran between 23 July and 10 September 2020.

The European Commission's "Regulation on a European Approach for Artificial Intelligence" fosters ad hoc protection for high-risk AI systems, based on a secure development life cycle. However, when it comes to cybersecurity, the proposed text could state more clearly some additional and necessary steps to achieve security of AI systems. The proposed requirements concern high-quality datasets, documentation and record-keeping, transparency and provision of information, human oversight, robustness, accuracy, and cybersecurity.[16]

As far as cybersecurity is concerned, the Regulation provides that high-risk AI systems *"shall be resilient to attempts by unauthorised third parties to alter their use or performance by exploiting the system vulnerabilities."*[17] It also stipulates that the technical solutions aimed at ensuring the cybersecurity of high-risk AI should encompass measures to prevent and control attacks trying to manipulate the training dataset inputs ('data poisoning') designed to cause the model to make a mistake ('adversarial examples'), or model flaws. These requirements represent a fundamental step towards assuring the necessary level of protection of AI systems.

This CEPS Task Force supports this approach and proposes a series of recommendations to provide further concrete guidance on how to secure AI systems.

Enhancing the AI sector in a timely fashion is particularly relevant for Europe. Given that the established market model is characterised by strong network and scale effects, first-mover gains in adopting AI technologies are particularly strong. While fostering its AI ecosystem, the EU has to both define how to make AI systems safe and reliable, and address what cybersecurity roadmap should be considered at the EU policy level to make the most out of such an AI ecosystem.

---

[15] European Commission, White Paper On Artificial Intelligence - A European approach to excellence and trust, COM(2020) 65 final, Brussels, 19.2.2020.

[16] European Commission, Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts, COM(2021) 206 final, Brussels, 21.4.2021.

[17] Ibid.

3. Some definitions

While the literature is state of the art, a shared definition of what AI is seems to be lacking. The definitions below give a better understanding of how AI has been conceptualised for the purposes of this report.

The Organisation for Economic Co-operation and Development (OECD) defines an AI system as a *"machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments."*[18] This definition has also been adopted by the European Commission in the "Regulation on a European Approach for Artificial Intelligence."

In this study we distinguish between symbolic and non-symbolic AI. In symbolic (or traditional) AI, programmers make use of programming languages to generate explicit rules to be hard coded into the machine. Non-symbolic AI does not rely on the hard coding of explicit rules. Instead, machines are able to process an extensive set of data, deal with uncertainty and incompleteness, and autonomously extract patterns or make predictions.

Machine learning is the major tool in today's AI systems. According to the OECD, ML is *"[...] a set of techniques to allow machines to learn in an automated manner through patterns and inferences rather than through explicit instructions from a human. ML approaches often teach machines to reach an outcome by showing them many examples of correct outcomes. However, they can also define a set of rules and let the machine learn by trial and error."*[19] ML algorithms are usually divided into three large categories: supervised learning, unsupervised learning, and reinforcement learning. In supervised learning, the data that are given to the ML algorithm already contain the correct answer (e.g., is this email spam?) whereas in unsupervised learning, algorithms cluster the data without prior information on how to break them down into groups.[20] Both systems are able to learn and make predictions based on this information. Reinforcement learning instead entails creating a system of rewards within an artificial environment to teach an artificial agent how to move through different states and act in a given environment.[21]

Neural networks are a sub-category of ML. These systems are characterised by layers that compute information in parallel and are formed by interconnected nodes that pass information to each other. The patterns of this knowledge represent the knowledge in these systems. According to the OECD: *"Neural networks involve repeatedly interconnecting thousands or millions of simple transformations into a larger statistical machine that can learn sophisticated*

[18] See OECD (2019), AI Policy Observatory, 22 May (www.oecd.ai/ai-principles).

[19] OECD (2019a), "Artificial Intelligence in Society", OECD Publishing, Paris (https://doi.org/10.1787/eedfee77-en.)

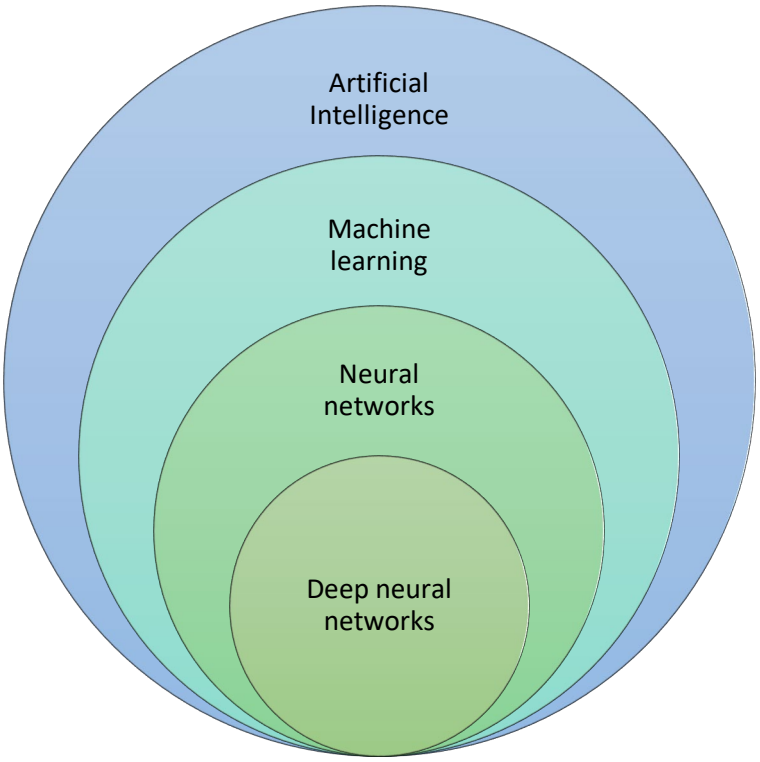
[20] B. Buchanan and T. Miller (2017), "Machine Learning for Policymakers, What It Is and Why It Matters", Belfer Center for Science and International Affairs Harvard Kennedy School, June.

[21] Ibid.

relationships between inputs and outputs. In other words, neural networks modify their own code to find and optimise links between inputs and outputs."[22]

Deep learning is a large neural network subset composed of hierarchical layers that increase the complexity of the relationship between input and output. It is an architecture able to implement supervised, unsupervised, and reinforcement learning. It uses networks with layers of nodes that mimic the neurons of the brain. Each layer of neurons uses the data from the layer below it, makes calculations and offers its output to the layers above it.[23] Figure 1 shows the relationship between AI and ML.

Figure 1. Relationship between AI and ML



Source: authors' composition based on Armin Wasicek (2018), "Artificial Intelligence vs. Machine Learning vs. Deep Learning: What's the Difference?", sumo logic, October.

4. AI for cybersecurity and cybersecurity for AI

AI in cybersecurity presents great opportunities but, as with any powerful general purpose, dual-use technology, it also brings great challenges. AI can improve cybersecurity and defence measures, allowing for greater system robustness, resilience, and responsiveness, but AI in the form of ML and deep learning will escalate sophisticated cyberattacks, enabling faster, better-targeted and more destructive attacks.

[22] OECD (2019a), op. cit.

[23] B. Buchanan and T. Miller (2017), op. cit, p. 17.

The application of AI in cybersecurity also poses security and ethical concerns. Among other things, it remains unclear how responsibilities for autonomous response systems should be ascribed, how to make sure systems are behaving according to the expectations, or what the security risks carried by the increasing anthropomorphisation of AI systems are.[24]

This report will therefore explore the twofold nature of the relationship between AI and cybersecurity. On the one hand, the report will explore the possibilities offered by AI adoption of enhancing cybersecurity, of particular importance if one considers the increase in cybersecurity breaches that accompanied the Covid-19 crisis. On the other hand, the report will address how cybersecurity for AI needs to be developed to make systems safe and secure. In this respect, the report will explore the concept of AI attacks, what the likelihood is of AI-enabled systems being subject to manipulation such as data poisoning and adversarial examples, and how to best protect AI systems from malicious attack.

[24] Anthropomorphic language at times appears intrinsic to the field of AI research. According to Salles, Evers and Farisco, *"From Turing's descriptions of his machines to accounts of AlphaZero's intellectual feats it is not uncommon to find terms typically used to describe human skills and capacities when referring to AIs and focusing on alleged similarities between humans and machines."* A. Salles, K. Evers and M. Farisco (2020), "Anthropomorphism in AI", *AJOB Neuroscience,* Vol. 11, No. 2.

PART II.
ARTIFICIAL INTELLIGENCE IN CYBERSECURITY

1. Introduction

According to many security analysts, security incidents reached the highest number ever recorded in 2019.[25] From phishing to ransomware, from dark web as a service economy to attacks on civil infrastructure, the cybersecurity landscape involved attacks that grew increasingly sophisticated during the year.[26] This upwards trend continued in 2020. The volume of malware threats observed averaged 419 threats per minute, an increase of 44 threats per minute (12%) in the second quarter of 2020.[27] Cyber criminals managed to exploit the Covid-19 pandemic and the growing online dependency of individuals and corporations, leveraging potential vulnerabilities of remote devices and bandwidth security. According to Interpol, 907,000 spam messages related to Covid-19 were detected between June and April 2020. Similarly, the 2020 Remote Workforce Cybersecurity Report showed that nearly two thirds of respondents saw an increase in breach attempts, with 34% of those surveyed having experienced a breach during the shift to telework.[28] Exploiting the potential for high impact and financial benefit, threat actors deployed themed phishing emails impersonating government and health authorities to steal personal data and deployed malware against critical infrastructure and healthcare institutions.[29]

In 2021 the drive for ubiquitous connectivity and digitalisation continues to support economic progress but also, simultaneously and 'unavoidably', creates a fertile ground for the rise in scale and volume of cyberattacks. Increasing ransomware and diversified tactics, increasingly mobile cyber threats, ever more sophisticated phishing, cyber criminals and nation state attackers targeting the systems that run our day-to day-lives and malicious actors attacking the cloud for every new low-hanging fruit. [30]

2. AI systems' support to cybersecurity

Against this backdrop, organisations have started using AI to help manage a growing range of cybersecurity risks, technical challenges, and resource constraints by enhancing their systems' robustness, resilience, and response. Police dogs provide a useful analogy to understand why companies are using AI to increase cybersecurity. Police officers use police dogs' specific abilities to hunt threats; likewise, AI systems work with security analysts to change the speed

[25] In the first quarter of 2019, businesses detected a 118% increase in ransomware attacks and discovered new ransomware families such as Anatova, Dharma and GandCrab, which use innovative techniques to target and infect enterprises, MacAfee (2019), "McAfee Labs Threats Report", August.

[26] M.Drolet (2020), "The Evolving Threat Landscape: Five Trends to Expect in 2020 and Beyond", Forbes Technology Council; Orange Business Service (2020), "2020 Security Landscape".

[27] MacAfee (2020), "McAfee Labs Threats Report", November.

[28] Fortinet (2020), Enterprises Must Adapt to Address Telework Security Challenges: 2020 Remote Workforce Cybersecurity Report", August.

[29] INTERPOL (2020), "INTERPOL report shows alarming rate of cyberattacks during COVID-19", August (www.interpol.int/en/News-and-Events/News/2020/INTERPOL-report-shows-alarming-rate-of-cyberattacks-during-COVID-19).

[30] Splunk (2019), "IT Security Predictions 2020"; ENISA (2020), "Emerging Trends – ENISA Threat Landscape", 20 October (www.enisa.europa.eu/publications/emerging-trends)

at which operations can be performed. In this regard, the relationship between AI systems and security operators should be understood as a synergetic integration, in which the unique added value of both humans and AI systems are preserved and enhanced, rather than as a competition between the two.[31]

Estimates suggest that the market for AI in cybersecurity will grow from $3.92 billion in 2017 to $34.81 billion by 2025, at a compound annual growth rate (CAGR) of 31.38% during the forecast period.[32] According to a recent Capgemini survey, the pace of adoption of AI solutions for cybersecurity is skyrocketing. The number of companies implementing these systems has risen from one fifth of the overall sample in 2019, to two thirds of companies planning to deploy them in 2020. 73% of the sample tested AI applications in cybersecurity. The most common applications are network security, followed by data security, and endpoint security. Three main categories can be identified in AI use in cybersecurity: detection (51%), prediction (34%), and response (18%).[33]

The driving forces that are boosting the use of AI in cybersecurity comprise:[34]

1. *Speed of impact:* In some of the major attacks, the average time of impact on organisations is four minutes. Furthermore, today's attacks are not just ransomware, or just targeting certain systems or certain vulnerabilities; they can move and adjust based on what the targets are doing. These kinds of attacks impact incredibly quickly and there are not many human interactions that can happen in the meantime.

2. *Operational complexity:* Today, the proliferation of cloud computing platforms and the fact that those platforms can be operationalised and deliver services very quickly – in the millisecond range – means that you cannot have a lot of humans in that loop, and you have to think about a more analytics-driven capability.

3. *Skills gaps in cybersecurity remain an ongoing challenge:* According to Frost & Sullivan,[35] there is a global shortage of about a million and a half cybersecurity experts. This level of scarcity pushes the industry to automate processes at a faster rate.

AI can help security teams in three ways: by improving systems' *robustness*, *response,* and *resilience.* The report defines this as the 3R model.[36] First, AI can improve systems' *robustness*, that is, the ability of a system to maintain its initial assumed stable configuration even when it

[31] K. Skapinetz (2018), "Overcome cybersecurity limitations with artificial intelligence", June (www.youtube.com/watch?time_continue=10&v=-tIPoLin1WY&feature=emb_title).

[32] MarketsandMarkets, "Artificial Intelligence in Cybersecurity Market by Technology Machine Learning, Context Awareness - 2025", MarketsandMarkets (www.marketsandmarkets.com/Market-Reports/ai-in-cybersecurity-market-224437074.html).

[33] CAP Gemini (2019), "Reinventing Cyber security with Artificial Intelligence. The new frontier in digital security", Research Institute.

[34] This section is taken from McAfee's contribution to the kick-off meeting of the CEPS Task Force.

[35] Frost & Sullivan (2017), "2017 Global Information Security Workforce Study", Center for Cyber Safety and Education.

[36] See M. Taddeo, T. McCutcheon and L. Floridi (2019) on this, "Trusting artificial intelligence in cybersecurity is a double-edged sword", *Nature Machine Intelligence*, November.

processes erroneous inputs, thanks to self-testing and self-healing software. This means that AI systems can be used to improve testing for robustness, delegating to the machines the process of verification and validation. Second, AI can strengthen systems' *resilience*, i.e. the ability of a system to resist and tolerate an attack by facilitating threat and anomaly detection. Third, AI can be used to enhance system *response,* i.e. the capacity of a system to respond autonomously to attacks, to identify vulnerabilities in other machines and to operate strategically by deciding which vulnerability to attack and at which point, and to launch more aggressive counterattacks.

Identifying when to delegate decision-making and response actions to AI and the need of an individual organisation to perform a risk-impact assessment are related. In many cases AI will augment, without replacing, the decision-making of human security analysts and will be integrated into processes that accelerate response actions.

2.1 System robustness

The need to respond to cyberattacks spurs companies to build systems that are self-learning, i.e., able to establish local context and distinguish rogue from normal behaviour.

Robustness can be understood as the ability of a system to resist perturbations that would fundamentally alter its configuration. In other words, a system is robust when it can continue functioning in the presence of internal or external challenges without changing its original configuration.

Artificial Intelligence for software testing (AIST) is a new area of AI research aiming to design software that can self-test and self-heal. Self-testing refers to "*the ability of a system or component to monitor its dynamically adaptive behaviour and perform runtime testing prior to, or as part of the adaptation process"*.[37] Hence, this area of research involves methods of constructing software that it is more amenable to autonomous testing, and knows when to deploy such systems and how to validate their correct behaviour.[38] These systems are able to check and optimise their state continuously and respond quickly to changing conditions. AI-powered behavioural analytics help compare how a system should run with how it is currently running and what the trigger corrections are.[39]

System robustness implies that AI is able to perform anomaly detection and profiling of anything that is generically different. It should be noted, however, that this approach can create a lot of noise from benign detections and false negatives when sophisticated attackers hide by blending in with normal observed behaviours. As such, more robust and accurate approaches focus on detecting attacker's specific and immutable behaviours.

[37] T.M. King et. al. (2019), "AI for testing today and tomorrow: Industry Perspective", IEEE International Conference on Artificial Intelligence Testing, IEEE, pp. 81-88.

[38] See AISTA, Self-Testing (www.aitesting.org/self-testing-ai).

[39] Wired Insider, "Fighting Cybercrime with Self-Healing Machines", *Wired*, October 2018.

System robustness can also be enhanced by incorporating AI in the system's development to increase security controls, for example via vulnerability assessment and scanning. Vulnerability assessment can be either manual, assistive, or fully automated. Fully automated vulnerability assessment leverages AI techniques and allows for considerable financial gains and time reductions. ML has been used to build predictive models for vulnerability classification, clustering, and ranking. Support-vector machines (SVMs), Naive Bayes, and Random Forests are among the most common algorithms. Various evaluation metrics are used to determine the performance, such as precision,[40] recall[41] and f-score.[42] Among other techniques, ML can be used to create risk-analysis models that proactively determine and prioritise security loopholes.[43] Automated planning has also been successfully applied for vulnerability assessment, mainly in the area of generating attack plans that can assess the security of underlying systems. The real-time steps of an attacker are modelled through automated planning, for example by simulating realistic adversary courses of action or focusing on malicious threats represented in the form of attack graphs. Khan and Parkinson suggest that if attack plans are generated by an AI system, there is greater potential to discover more plans than if they are generated by human experts.[44]

Code review is another area of application for enhancing system robustness. Peer code review is a common best practice in software engineering where source code is reviewed manually by one or more peers (reviewers) of the code author. Automating the process by using AI systems can both reduce time and allow a greater number of bugs to be discovered than ones discovered manually. Several AI systems are being developed for code review support. In June 2020, for example, the Amazon Web Services' AI-powered code reviewer from CodeGuru was made publicly available.[45]

The use of AI to improve system robustness not only has a tactical effect (i.e. improving the security of systems and reducing their vulnerability) but also a strategic one. Indeed, it decreases the impact of zero-days attacks. Zero-days attacks leverage vulnerabilities that are exploitable by attackers as long as they remain unknown to the system providers or as long as there is no patch to resolve them. By decreasing the impact of zero-days attacks, AI reduces their value on the black market.[46]

---

[40] Precision is a metric that quantifies the number of correct positive predictions made.

[41] Recall is a metric that quantifies the number of correct positive predictions made out of all positive predictions that could have been made.

[42] F-Measure provides a way to combine both precision and recall into a single measure that captures both properties.

[43] For more on ML techniques for performing fully automated vulnerability assessment, see S. Khan and S. Parkinson (2018), "Review into State of the Art of Vulnerability Assessment using Artificial Intelligence", *Guide to Vulnerability Analysis for Computer Networks and Systems*, Springer, Cham, pp.3-32.

[44] Ibid.

[45] See Amazon, CodeGuru (https://aws.amazon.com/it/codeguru/).

[46] M. Taddeo T. McCutcheon and L. Floridi (2019), "Trusting artificial intelligence in cybersecurity is a double-edged sword", *Nature Machine Intelligence*, November, pp. 1-4.

2.2 System resilience

Resilience can be understood as the ability of a system to resist and tolerate an attack by facilitating threat and anomaly detection. In other words, a system is resilient when it can adapt to internal and external challenges by changing its methods of operations while continuing to function. System resilience implies, unlike system robustness, some fundamental shift in the core activities of the system that has to adapt to the new environment. Threat and anomalies detection (TAD) is today the most common application of AI systems. Indeed:

- There are now approximately 592,145 new unique malware files every day, and possibly even more.
- Classification of new threats by humans alone is impossible, and besides, threats are becoming more complicated and better dissimulated.
- In the past, it was common to use signatures to classify malicious attacks, leveraging databases of known threats. Such measures, however, are becoming considerably less effective against the latest strains of advanced malware, which evolve by the second.[47]

AI solutions for cybersecurity enable a fundamental shift from a signature-based detection to a more flexible and continuous monitoring of the network as it shifts from its normal behaviours. "*AI algorithms can detect any changes that appear abnormal – without needing an advance definition of abnormal.*"[48] AI can also provide insights into potential attacks by performing deep packet traces through internal or external sensors or pieces of monitoring software.[49]

Companies use AI to automate cyber defences against spam and phishing and to detect malware, fraudulent payments, and compromised computers and network systems.[50] Furthermore, AI is used for critical forensics and investigative techniques. In particular, AI is used to create real-time, customer-specific analysis, improving the total percentage of malware identified and reducing false positives. Hence, AI data processing helps cybersecurity threat intelligence become more effective. Finally, organisations are using AI-based predictive analytics to determine the probability of attacks, enhancing an organisation's network defence through near real-time data provisions. Predictive analytics can help in processing real-time data from various sources and identifying attack vectors by helping manage big data; in filtering and parsing the data before they are analysed; in automatically filtering out duplicates; in categorising information; and by suggesting which incident to prioritise. In this way predictive analytics reduces human errors and the workload for security analysts.[51]

---

[47] This section is taken from Palo Alto Network's contribution to the fourth meeting of the CEPS Task Force.

[48] R. Goosen et al. (2018), "Artificial intelligence is a threat to cybersecurity. It's also a solution", The Boston Consulting Group.

[49] Ibid.

[50] Companies like McAfee have access to 1bn sensors via their end points, web gateway, cloud, and CASB protection services and use ML to transform raw data into analytics and insight.

[51] WhoisXML API (2019), "The importance of Predictive Analytics and Machine Learning in Cybersecurity", CircleID, September.

While the use of AI in cybersecurity is increasingly indispensable, AI systems will continue to require a rather collaborative environment between AI and humans, at least for the foreseeable future. While completely autonomous systems do exist, their use is as yet relatively limited, and systems still often require human intervention to function as intended.

In this respect, the people involved have to keep monitoring the system (for accuracy, to change request, etc.). Some models still have to be retrained every single day just to stay ahead of the attackers, as attacks change in response to the defences being built. Finally, there are communities of security practitioners that continue to work together to establish a common understanding of what is malicious and what is not.[52]

2.3 System response

System resilience and response are deeply intertwined and logically interdependent, as, to respond to a cyberattack, you need to detect what it is occurring and develop and deploy an appropriate response by deciding which vulnerability to attack and at which point, or by launching counterattacks. During the 2014 Defence Advanced Research Projects Agency (DARPA) Cyber Grand Challenge seven AI systems fought against each other, identifying and patching their own vulnerabilities while exploiting their opponents' flaws without human instructions. Since then, prevention of cyberattacks is increasingly going in the direction of systems able to deploy real-time solutions to security flaws. AI can help to reduce cybersecurity experts' workloads by prioritising the areas that require greater attention and by automating some of the experts' tasks.[53] This aspect is particularly relevant if one considers the shortage in the supply of cybersecurity professionals, which is currently estimated at four million workers.[54]

AI can facilitate attack responses by deploying, for example, semi-autonomous lures that create a copy of the environment that the attackers are intending to infiltrate. These deceive them and help understand the payloads (the attack components responsible for executing an activity to harm the target). AI solutions can also segregate networks dynamically to isolate assets in controlled areas of the network or redirect an attack away from valuable data.[55] Furthermore, AI systems are able to generate adaptive honeypots (computer systems intended to mimic likely targets of cyberattacks) and honeytokens (chunks of data that look attractive to potential attackers). Adaptive honeypots are more complex than traditional honeypots insofar as they change their behaviour based on the interaction with attackers. Based on the attacker's reaction to the defences, it is possible to understand its skills and tools. The AI solution gets to learn the attacker's behaviour via this tool so that it will be recognised and tackled during future attacks.

---

[52] This section is taken from Palo Alto Network's contribution to the fourth meeting of the CEPS Task Force.

[53] R. Goosen et al. (2018), "Artificial intelligence is a threat to cybersecurity. It's also a solution", The Boston Consulting Group.

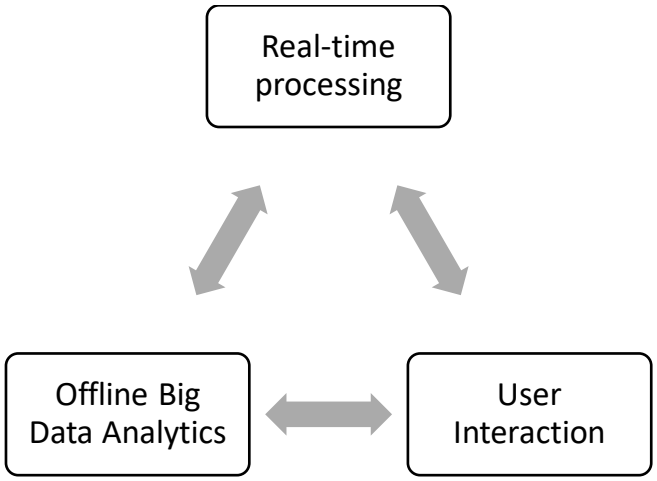
[54] (ISC)[2] (2019), "Cybersecurity Workforce Study Strategies for Building and Growing Strong Cybersecurity Teams" (www.isc2.org/-/media/ISC2/Research/2019-Cybersecurity-Workforce-Study/ISC2-Cybersecurity-Workforce-Study-2019.ashx?la=en&hash=1827084508A24DD75C60655E243EAC59ECDD4482).

[55] Ibid.

2.4 Major techniques in the use of AI for system robustness, resilience, and response

Whenever AI is applied to cyber-incident detection and response the problem solving can be roughly divided into three parts, as shown in Figure 2. Data is collected from customer environments and processed by a system that is managed by a security vendor. The detection system flags malicious activity and can be used to activate an action in response.

Figure 2. AI cyber incidents detection and response



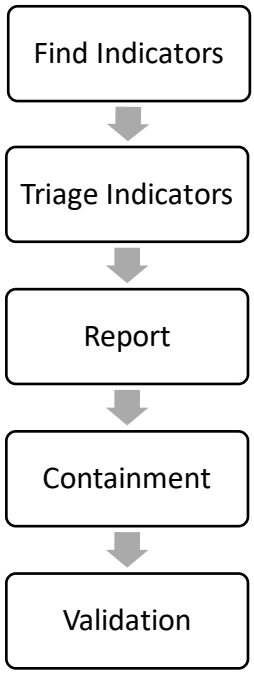
Source: Palo Alto Network contribution to the fourth meeting of the CEPS Task Force.

Companies today recognise that the attack surface is growing massively because of the adoption of the Internet of Things (IoT) and the diffusion of mobile devices, compounded by a diverse and ever-changing threat landscape. Against this backdrop, there are two measures that can be implemented:

1. speed up defenders
2. slow down attackers.

With respect to speeding up defenders, companies adopt AI solutions to automate the detection and response to attacks already active inside the organisation's defences. Security teams traditionally spend a lot of time dealing with alerts, investigating if they are benign or malicious, reporting on them, containing them, and validating the containment actions. AI can help with some of the tasks that security operations teams spend most of their time on. Notably, this is also one of the primary and most common uses of AI in general.

In particular, security operations teams can use AI to solve the following five fundamental questions:



1. Classify items: What does this item represent? How mature is it?
2. What is the scale of it?
3. How is it organised?
4. Is it anomalous? Is it a behaviour an active attacker would manifest?
5. What actions should be taken?

In the past, the industry focused first on labelling and on categorising malware, but nowadays companies are using models that are not looking for individual pieces of malware; rather they are looking at how attackers behave. It is therefore becoming more common to use ML threat-detection models that are behavioural in their analysis, and in turn they are becoming durable and potentially capable of detecting zero-day attacks. The goal is to identify the very subtle yet immutable attack behaviour with high fidelity, and low noise.

The following are practical examples of the benefits of using AI and ML for cybersecurity detection and response.[56]

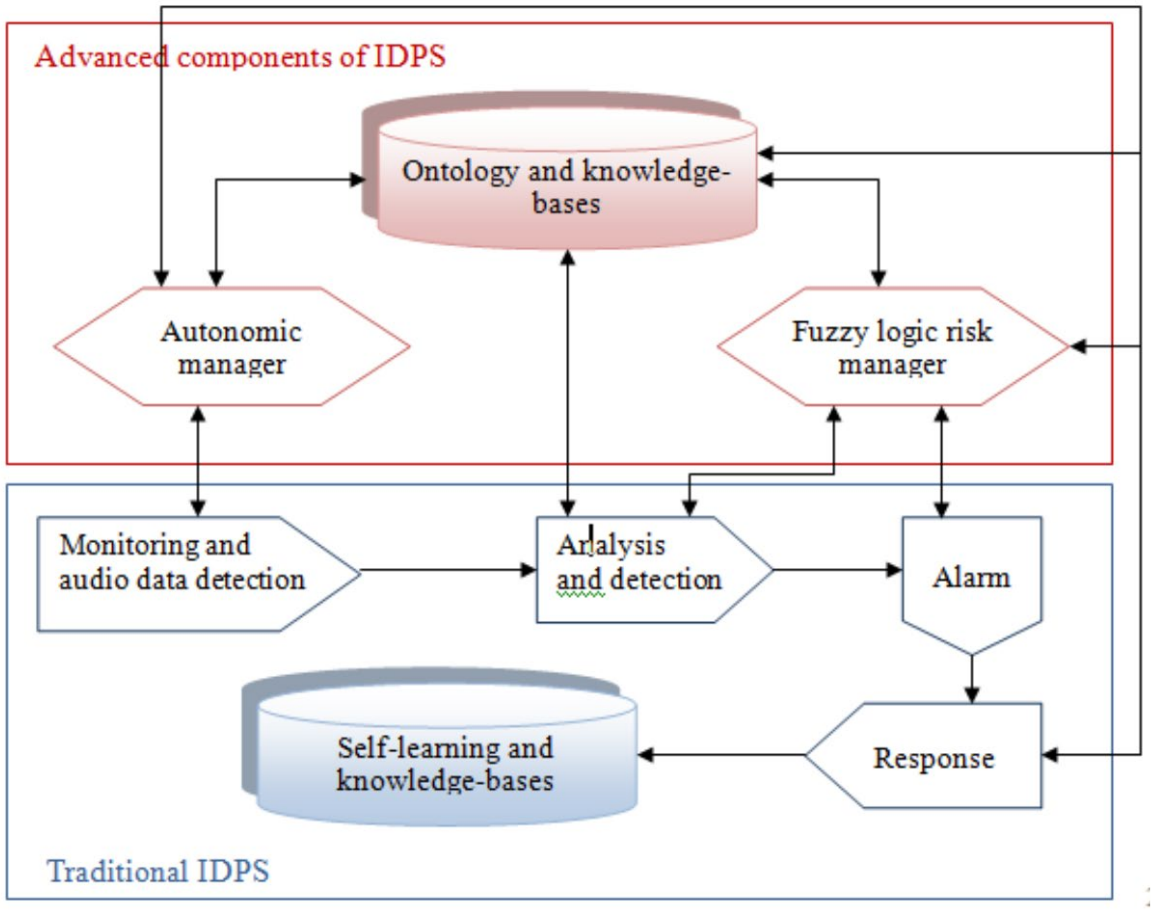
- ML trained on user interaction provides a way of understanding local context and knowing what data to focus on; models trained to identify those more likely to be malicious improve the efficiency of a system by triaging the information to process in real time. In this way, using ML is cost saving but also allows for faster reaction in the most critical situations.
- ML can be useful in detecting new anomalies by learning robust models from the data they have been fed with. ML is particularly good at identifying patterns and extracting algorithms in large sets of data where humans are lost.
- ML can be useful for asynchronous user profiling and for measuring deviation from common behaviours as well as going back to much larger data volumes to understand behaviour.
- ML trained on immutable attacker 'Tactics, Techniques, and Procedures' (TTP) behaviours (those identified in the Mire Attack framework)[57] can support durable and broad attacker detection.

[56] This section is taken from Vectra's contribution to the kick-off meeting of the CEPS Task Force.

[57] See MITRE ATT&CK (https://attack.mitre.org).

To better illustrate the use of AI and ML for cybersecurity detection and response, Figure 3 presents an intrusion detection and prevention system that combines software and hardware devices inside the network. The system *"can detect possible intrusions and attempt to prevent them. Intrusion detection and prevention systems provide four vital security functions: monitoring, detecting, analysing and responding to unauthorized activities."*[58]

Figure 3. Intrusion detection and prevention system



Source: Dilek (2015).

There are a variety of AI techniques that can be used for intrusion prevention, detection, and response. Table 1 illustrates examples of the main advantages of some of these techniques.[59]

[58] S. Dilek, H. Caku and M. Aydin, (2015), "Applications of Artificial Intelligence Techniques to Combating Cyber Crime: A Review", *International Journal of Artificial Intelligence & Applications*, p. 24.

[59] Please note that the list does not aim to be comprehensive for all the possible AI techniques for intrusion prevention, detection and response.

Table 1. Examples of AI techniques for intrusion prevention, detection and response

Technology	Advantages
Artificial Neural Networks[60]	Parallelism in information processing Learning by example Nonlinearity – handling complex nonlinear functions Resilience to noise and incomplete data Versatility and flexibility with learning models
Intelligent Agents[61]	Mobility Rationality – in achieving their objectives Adaptability – to the environment and user preferences Collaboration – awareness that a human user can make mistakes and provide uncertain or omit important information; thus they should not accept instructions without consideration and checking the inconsistencies with the user
Genetic Algorithms[62]	Robustness Adaptability to the environment Optimisation – providing optimal solutions even for complex computing problems Parallelism – allowing evaluation of multiple schemas at once Flexible and robust global search
Fuzzy Sets[63]	Robustness of their interpolative reasoning mechanism Interoperability – human friendliness

Source: Dilek (2015).

All these intrusion detection AI-powered technologies help in reducing the dwell time – the length of time a cyberattacker has free reign in an environment from the time they get in until they are eradicated.[64] In December 2019, the dwell time in Europe was about 177 days, and attackers were discovered in only 44% of cases because of data breach or other problems. Using AI techniques, the dwell time has been dramatically reduced.[65]

Finally, AI can be also very helpful in enhancing network security. (See Box 1).

[60] First developed in 1957 by Frank Rosenblatt, these techniques rely on the perceptron. By connecting with one another and processing raw data, perceptrons independently learn to identify the entity on which they have been trained. See A. Panimalar et al. (2018), "Artificial intelligence techniques for cybersecurity", *International Research Journal of Engineering and Technology* (IRJET), Vol. 5, No. 3.

[61] Intelligent Agents are defined as entries able to recognise movement through their sensors, to follow up on an environment based on the perceived condition using actuators and to direct their behaviour toward the accomplishment of an objective. They can vary greatly in complexities (thermostats, for example, are intelligent agents). In cybersecurity, they can be used in showdown DDoS attacks, and could potentially be deployed as Cyber Police mobile agents. See A. Panimalar et al. (2018), op. cit.

[62] The genetic algorithm is a method for solving both constrained and unconstrained optimisation problems that is based on natural selection, the process that drives biological evolution.

[63] Fuzzy sets can be considered an extension and simplification of classical sets. They can be understood in the context of set membership. They allow partial membership of elements that have varying degrees of membership in the set.

[64] See Optiv, "Cybersecurity Dictionary, Dwell Time" (www.optiv.com/cybersecurity-dictionary/dwell-time).

[65] M. Walmsley (2019), intervention at the CEPS Cyber Summit 2019, December (www.youtube.com/watch?v=sY16ToU9UiQ [3:05:40]).

Box 1. AI and network security

Example 1. Detecting route hijacking attacks[66]

AI is helpful in enhancing network security. An increasingly popular cyberattack today is hijacking Internet Protocol (IP) addresses. 'Route hijacking' means stealing traffic intended for other destinations. The regions of the Internet in the world are connected through a global routing protocol called the Border Gateway Protocol (BGP), which allows different parts of the Internet to talk to each other. Using the BGP, networks exchange routing information in such way that packets are able to reach the correct destination. Each region announces to its neighbourhood that it holds certain IP addresses. There are about 70,000 regions on the Internet called autonomous systems and about 700,000 distinct announcements. The BGP does not have any security procedures for validating that a message is actually coming from the place it says it's coming from, so hijackers exploit this shortcoming by convincing nearby networks that the best way to reach a specific IP address is through their network. In other words, a rogue region can announce that it contains an IP address that belongs, for instance, to MIT. A malicious router would be advertising a network that does not really belong to its autonomous system (the range of IP addresses that it has authority over). In so doing, the malicious router and related infrastructure can eavesdrop, and redirects the traffic that was supposed to go to MIT to the rogue region. This is happening regularly, for example to send spam and malware or when an actor manages to hijack bitcoin traffic to steal the bitcoins.

In a recent joint project between MIT and the University of California at San Diego, researchers have trained a machine-learning model to automatically identify malicious actors through the patterns of their past traffic. Using data from network operator mailing lists and historical BGP data, taken every five minutes from the global routing tables during a five-year period, the machine-learning model was able to identify malicious actors. Their networks had key characteristics related to the specific blocks of IP addresses they use, namely:

- Volatile changes in activity: if a region announces address blocks and then the announcements disappear in a short time, the likelihood of there being a hijacker becomes very high. The average duration of an announcement for legitimate networks was two years, compared with 50 days for hijackers.
- Multiple address blocks: serial hijackers advertise many more blocks of IP addresses.
- IP addresses in multiple countries: most networks do not have foreign IP addresses, but hijackers are much more likely to announce addresses registered in different countries and continents.

One challenge in developing this system was handling the false positives related to a legitimate short-term address announcement or human error. Indeed, changing the route is sometimes a legitimate way to block an attack.

This model allows network operators to handle these accidents in advance by tracing hijackers' behaviour instead of reacting on a case-by-case basis.

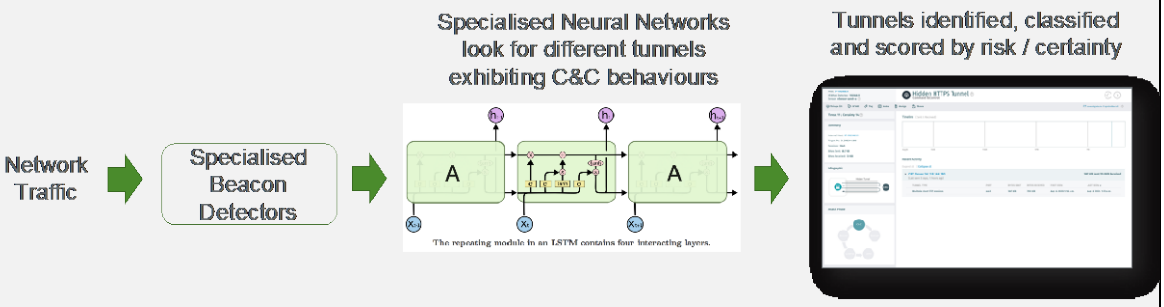
[66] This section draws from the intervention of Professor David Clark from MIT at the third meeting of the CEPS Task Force and from A. Conner-Simons (2019), "Using machine learning to hunt down cybercriminals", MIT CSAIL, October.

The MIT model is particularly relevant when considering more generally that the Internet was not designed as a high-security network. Incremental security improvements primarily address specific attacks, but these might fail to solve the fundamental problems and could also introduce new undesirable consequences (e.g., Border Gateway Protocol Security prevents route hijacking but causes delayed route convergence, and does not support prefix aggregation, which results in reduced scalability).[i]

Example 2. Detecting hidden tunnel attacks[ii]

Identifying attackers who are already operating inside compromised networks is a more complex challenge. Sophisticated attackers use hidden tunnels to carry out command-and-control and exfiltration behaviours. This means that they steal critical data and personally identifiable information (PII) by blending in with normal traffic, remotely controlling the theft of information, and slipping it out through those same tunnels. Because they blend in with multiple connections that use normal, commonly allowed protocols, hidden tunnels are very difficult to detect.

AI can constantly perform a highly sophisticated analysis of metadata from network traffic, revealing subtle abnormalities within a protocol that gives away the presence of a hidden tunnel. Even though messages are disguised within an allowed protocol, the resulting communications that make up the hidden tunnel can't help but introduce subtle attack behaviours into the overall conversation flow. These include slight delays or abnormal patterns in requests and responses.



Based on these behavioural traits, Neural Networks can be used to accurately detect hidden tunnels within, for example, HTTP, HTTPS, and Domain Name System (DNS) traffic without performing any decryption or inspection of private payload data. It doesn't matter what field attackers use to embed communications or whether they use a never-before-seen obfuscation technique. The attacker's variance from normal protocol behaviour will still expose the hidden tunnel's presence to the Neural Networks.

[i] While the contribution of AI/ML to cybersecurity is of relevance, it is critical that cybersecurity be addressed at the root wherever possible. Scalability, Control and Isolation on Next Generation Networks (SCION) is an Internet-compatible (IPv4 and IPv6) architecture that addresses today's network security issues on the Internet (www.scion-architecture.net).

[ii] See "Breaking ground: Understanding and identifying hidden tunnels" (www.vectra.ai/blogpost/breaking-ground-understanding-and-identifying-hidden-tunnel).

3. AI malicious uses

AI developments bring not only extensive possibilities, but also many corresponding challenges. People can use AI to achieve both honourable and malevolent goals.

The impact of AI on cybersecurity is usually described in terms of expanding the threat landscape. The categories of actors and individuals enabled through AI to carry out malicious attacks are proliferating. At the same time, new forms of attacks against AI systems – different in nature from traditional cyberattacks – increase the attack surface of connected systems in an exponential and sometimes unmeasurable way.

As far as these shifts are concerned, researchers agree that AI affects the cybersecurity landscape by:

- expanding existing threats
- introducing new threats
- altering the typical characteristics of threats.[67]

3.1 Expansion of existing threats

The availability of cheap and increasingly effective AI systems for attacks means categories of individuals and groups have the potential to become malicious actors. This means the asymmetry that once existed in the power balance between conventional and unconventional actors is increasingly shrinking. With the widening spectrum of actors capable of meaningfully undertaking a potentially significant attack, such as those against critical infrastructures, the malicious use of AI applications has become one of the most discussed aspects of this technology.

Experts refer to this phenomenon as the 'democratisation of artificial intelligence', meaning both the increasing number of potential actors exploiting AI to perform an attack, and the democratisation of the software and AI systems themselves. Indeed, the ease of access to scientific and engineering works around machine learning partly explains the increasing availability of AI to a greater number of individuals.[68] In modern times, access to software codes has become an increasingly easy task. Open repositories of stored software programming allow anyone with a laptop and the discrete knowledge to be able to explore the source code of a lot of software, including AI. This is even more relevant in a context in which there is already wide disclosure of hacking tools. Furthermore, academic and scientific research on AI is often openly disseminated, and made available to the general public with little review of misuse-prevention measures, and even fewer boundaries[69] on the vulgarisation of such outcomes. The issue of research openness will be further explored in this report.

---

[67] See M. Brundage et al. (2018), "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation", Malicious AI Report, February, p. 18.

[68] As J.-M. Rickli puts it, "*artificial intelligence relies on algorithms that are easily replicable and therefore facilitate proliferation. While developing the algorithm takes some time, once it is operational, it can be very quickly and easily copied and replicated as algorithms are lines of code*", J.-M. Rickli (2018), "The impact of autonomy and artificial intelligence on strategic stability", UN Special, July-August, pp. 32-33.

[69] For instance, *"(…) it is generally much easier to gain access to software and relevant scientific findings. Indeed, many new AI algorithms are reproduced in a matter of days or weeks. In addition, the culture of AI research is characterized by a high degree of openness, with many papers being accompanied by source code.",* M. Brundage (2018), op.cit., p. 17.

The automation of tasks previously undertaken by humans is another effect of the democratisation of AI. As Ferguson puts it, "*Imagine your attacker has the ability to conduct real-time impersonation of both systems and people, no longer harvesting passwords with noisy pen-testing tools, but through real-time adaptive shimming of the very systems it seeks later to exploit.*"[70] As more and more people use ML, the pattern of time-consuming tasks could be speeded up, rendering them more effective, and making cyber capabilities that were once the preserve of large industry players or wealthy governments accessible to small groups and individuals.[71]

The cost-availability nexus is another factor in the democratisation of AI that leads to the widening spectrum of malicious actors. As Comiter points out: *"the proliferation of powerful yet cheap computing hardware means almost everyone has the power to run these algorithms on their laptops or gaming computers. […] it does have significant bearing on the ability for non-state actors and rogue individuals to execute AI attacks. In conjunction with apps that could be made to allow for the automation of AI attack crafting, the availability of cheap computing hardware removes the last barrier from successful and easy execution of these AI attacks."*[72]

To sum up, the spectrum of malicious actors is being widened by the proliferation of cheap computing hardware, the growing availability and decreasing cost of computing capability through the cloud, and the open-source availability of most of the tools that could facilitate model training and potentially malicious activities.

The greater accessibility of AI tools also affects the combination of efficiency and scalability.[73] Some of the AI systems that are replacing tasks once assigned to humans are destined to depart from ordinary human performance. They will run in a faster way, and will execute those tasks a greater number of times.[74] In the cybersecurity context, scalability will allow an attack to reproduce at a level that has not been seen before. By using the example of spear-phishing attacks, Brundage et al point to two basic effects of scalability and efficiency for the actors driving an attack with an AI system.[75] On the one hand, cheap and efficient AI systems will, as mentioned, expand the category of adversaries being able to handily access such applications. On the other hand, actors that were already present in the threat landscape and labelled as

---

[70] R. Ferguson (2019), "Autonomous Cyber Weapons - The Future of Crime?", *Forbes*, 10 September (www.forbes.com/sites/rikferguson1/2019/09/10/autonomous-cyber-weapons-the-future-of-crime/#549591f85b1a).

[71] M.C Horowitz et al. give the example of the 'script kiddies', i.e. "…*relatively unsophisticated programmers, (…) who are not skilled enough to develop their own cyber-attack programs but can effectively mix, match, and execute code developed by others? Narrow AI will increase the capabilities available to such actors, lowering the bar for attacks by individuals and non-state groups and increasing the scale of potential attacks for all actors*.", M.C Horowitz et al. (2018), "Artificial Intelligence and International Security", Center for a New American Security, p. 13.

[72] M. Comiter (2019), "Attacking Artificial Intelligence: AI's Security Vulnerability and What Policymakers Can Do About It", Belfer Center for Science and International Affairs, Harvard Kennedy School, August, p. 53.

[73] OECD (2019a), op. cit., p. 96.

[74] See M. Brundage et al. (2018), op. cit., p. 5 and p. 16. Nonetheless, the devolution of tasks from humans to machines do encounter a certain limits. For instance, see B. Buchanan and T. Miller (2017), "Machine Learning for Policymakers", Belfer Center for Science and International Affairs, Harvard Kennedy School, p. 20; See also K. Grace et al. (2017), *When Will AI Exceed Human Performance? Evidence from AI Experts*, ArXiv.

[75] See also OECD (2019a), op. cit., p. 96.

potential malicious attackers will be able to benefit from AI systems to which they already had access, with a much higher efficiency rate.[76]

The wider distribution of AI systems not only multiplies the opportunities for cyberattacks – by increasing their speed and volume – but also allows them to become more sophisticated, for example by making their attribution and detection harder. AI also allows for the discovery of flaws that were never discovered before. Attackers, for instance, are able to more easily discover vulnerabilities generating new payloads fuzzing to discover new issues. Unusual behaviour triggers abnormal responses in the system, and AI systems, trained by already-discovered payloads for existing vulnerabilities, can suggest new payloads that would increase the chances of discovering new systems' exposures. AI can also help to exploit, not just discover, these newly discovered vulnerabilities by generating exploit variants and running them faster.[77]

Finally, it appears that such an increase of actors also impacts national and international security, particularly because of the inherent dual use of AI technology. According to the available literature, the repurposing of easily accessible AI systems is already having a significant effect on the development of lethal autonomous weapons systems (LAWS).[78] The availability of accessible AI solutions will also expand the possibility of warfare activities and tasks that will have a strategic impact being relayed to surrogates to conduct. Both state and non-state actors are increasingly relying on technological surrogates such as AI to be used as a force multiplier. An example of this is the alleged meddling in the 2016 US election, when a disinformation campaign aimed to persuade targeted voters to support the winning candidate. Another example is the US offensive operation carried out in 2019 as part of the ongoing cyberwar against Iran. This disabled a critical database that Iran was using to plot attacks against US oil tankers.[79]

3.1.1 Characteristics of AI-powered attacks80

Three characteristics of AI are likely to affect the way in which AI-powered attacks are carried out:

1. *Evasiveness*: AI is helping to modify the way in which attacks are detected. An AI-powered malware is much more difficult to detect by an anti-malware. The case in point

[76] M. Brundage et al. (2018), op. cit., p. 18.

[77] I. Novikov (2018), "How AI Can Be Applied To Cyberattacks", *Forbes*, 22 March (www.forbes.com/sites/forbestechcouncil/2018/03/22/how-ai-can-be-applied-to-cyberattacks/#27ef6e9849e3).

[78] C. Czosseck, E. Tyugu and T. Wingfield (eds), (2011), "Artificial Intelligence in Cyber Defense", Cooperative Cyber Defense Center of Excellence (CCD COE) and Estonian Academy of Sciences, 3rd International Conference on Cyber Conflict, Tallinn, Estonia. According to Ferguson, "*The repurposing of this technology will undoubtedly start at the nation-state level, and just like everything else it will trickle down into general availability. It is already past time for defenders to take the concept of autonomous cyber weapons seriously*." Ferguson, R (2019), "Autonomous Cyber Weapons - The Future of Crime?" *Forbes*, 10 September (www.forbes.com/sites/rikferguson1/2019/09/10/autonomous-cyber-weapons-the-future-of-crime/#549591f85b1a).

[79] See J. E. Barnes (2019), "Hurt Iran's Ability to Target Oil Tankers, Official Says", *New York Times*, 28 August (www.nytimes.com/2019/08/28/us/politics/us-iran-cyber-attack.html).

[80] This section draws from the contribution of Marc Ph. Stoecklin, IBM Research Centre Zurich and member of the Task Force Advisory Board, in the kick-off meeting of the CEPS Task Force.

is represented by IBM's Deeplocker malware. This is a new class of highly targeted malware that uses AI to hide its nature in benign applications, such as video conferencing applications, and identifies its target through face recognition, voice recognition or geo-localisation. The malware can conceal its intent until it reaches the defined target, which makes it fundamentally different from the classic 'spray-and-pray' attacks.

2. *Pervasiveness*: On 13 March 2004 driverless cars competed in the DARPA Grand Challenge in the Mojave Desert. Although this was deemed a failure because no vehicle achieved anything close to the goal, the improvements in driverless car technology have since been enormous. In 2016 DARPA launched the Cyber Grand Challenge in which competitors were asked to bring bots able to compete against each other without human instructions. As with the self-driving vehicles, the future pervasive potential of these new technologies is clear. This era of pervasive intelligence will be marked by a proliferation of AI-powered smart devices able to recognise and react to sights, sounds, and other patterns. Machines will increasingly learn from experience, adapt to changing situations, and predict outcomes. The global artificial intelligence market size was valued at $39.9 billion in 2019 and is expected to grow at a CAGR of 42.2% from 2020 to 2027.[81]

3. *Adaptiveness*: AI is adaptive, meaning that it can learn and to some extent become creative, and come up with ideas that attackers would not necessarily have thought of. During the DEF CON Hacking conference in 2017, a group of researchers showed how they successfully attacked a web application through an AI that found its way in using the Structured Query Language (SQL) database injection attack. The distinctiveness of this attack was that the AI figured out by itself how the SQL injection worked.

3.2 Introduction of new threats

As well as existing threats expanding in scale and scope, progress in AI means completely new threats could be introduced. The AI characteristics of being unbounded by human capabilities could allow actors to execute attacks that would not otherwise be feasible.

3.2.1 Deepfakes[82]

The use of 'deepfakes' has been steadily rising since a Reddit user first coined the term in 2017. Deepfakes are a developing technology that use deep learning to make images, videos, or texts of fake events. There are two main methods to make deepfakes. The first is usually adopted for 'face-swapping' (i.e., placing one person's face onto someone else's), and requires thousands

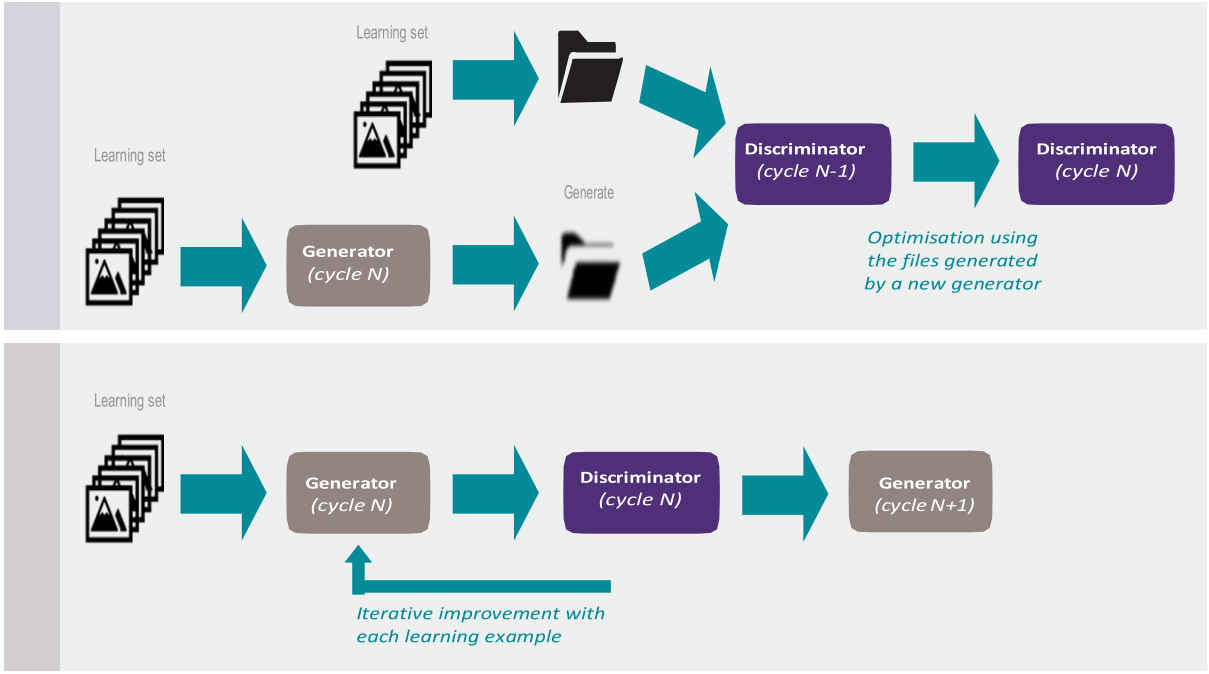
[81] Grand View Research (2020), "Artificial Intelligence Market Size, Share & Trends Analysis Report By Solution (Hardware, Software, Services), By Technology (Deep Learning, Machine Learning), By End Use, By Region, And Segment Forecasts, 2020 – 2027", July.

[82] This section of the report was contributed by Jean-Marc Rickli from the Geneva Centre for Security Policy (GCSP) and member of the Advisory Board of the Task Force, with the help of Alexander Jahns.

of face shots of the two people to be run through an AI algorithm called an encoder. The encoder then finds and learns similarities between the two faces, and reduces them to their shared common features, compressing the images in the process. A second AI algorithm called a decoder is then taught to recover the faces from the compressed images: one decoder recovers the first person's face, and another recovers the second person's face. Then, by giving encoded images to the 'wrong' decoder, the face-swap is performed on as many frames of a video as possible to make a convincing deepfake.[83]

The second and very important method to make deepfakes is called a generative adversarial network (GAN). A GAN pits two AI algorithms against each other to create brand new images (see Figure 4). One algorithm, the generator, is fed with random data and generates a new image. The second algorithm, the discriminator, checks the image and data to see if it corresponds with known data (i.e. known images or faces). This battle between the two algorithms essentially winds up forcing the generator into creating extremely realistic images (e.g. of celebrities) that attempt to fool the discriminator.[84]

Figure 4. The functioning of a generative adversarial network



Source: C. Meziat and L. Guille (2019), "Artificial Intelligence and Cybersecurity", Wavestone, 5 December.

[83] I. Sample (2020), "What are deepfakes and how can you spot them" *The Guardian*, 13 January (www.theguardian.com/technology/2020/jan/13/what-are-deepfakes-and-how-can-you-spot-them).

[84] K. Vyas (2019), "Generative Adversarial Networks: The Tech Behind DeepFake and FaceApp", *Interesting Engineering*, 12 August (https://interestingengineering.com/generative-adversarial-networks-the-tech-behind-deepfake-and-faceapp).

These images have been used to create fake yet realistic images of people, with often harmful consequences. For example, a McAfee team used a GAN to fool a facial recognition system like those currently in use for passport verification at airports. McAfee relied on state-of-the-art, open-source facial-recognition algorithms, usually quite similar to one another, thereby raising important concerns about the security of facial-recognition systems.[85]

Deepfake applications also include text and voice manipulation as well as videos. As far as voice manipulation is concerned, Lyrebird claims that, using AI, it was able to recreate any voice using just one minute of sample audio, while Baidu's Deep Voice clones speech with less than four seconds of training. In March 2019, AI-based software was used to impersonate a chief executive's voice and demand a fraudulent transfer of €220,000 ($243,000). In this case, the CEO thought he was talking to the chief executive of the firm's German parent company, who demanded the payment be made to a Hungarian subsidiary.

Deepfakes used for text manipulation is also increasingly concerning. With GPT-3 generative writing it is possible to synthetically reproduce human-sounding sentences that are potentially even more difficult to distinguish from human-generated ones than video content. Even with state-of-the art technology, it is still possible to tell video content has been synthetically produced, for example from a person's facial movements being slightly off. But with GPT-3 output there is no unaltered original that could be used for comparison or as evidence for a fact check.[86] Text manipulation has been used extensively for AI-generated comments and tweets. Diresta highlights how *"seeing a lot of people express the same point of view, often at the same time or in the same place, can convince observers that everyone feels a certain way, regardless of whether the people speaking are truly representative – or even real. In psychology, this is called the majority illusion."* As such, by potentially manufacturing a majority opinion, text manipulation is and will increasingly be applied to campaigns aiming to influence public opinion. The strategic and political consequences are clear.[87]

The malicious use of deepfakes is trending in many areas, as discussed below.

Pornographic

The number of deepfake videos online amounted to 14,678 in September 2019, according to Deeptrace Labs, an 100% increase since December 2018. The majority of these (96%) are pornographic in content, although other forms have also gained popularity.[88] Deepfake technology can put women or men in a sex act without their consent, while also removing the original actor, creating a powerful weapon for harm or abuse. According to a Data and Society

[85] K. Hao and P. Howell O'Neill (2020), "The hack that could make face recognition think someone else is you", *MIT Technology Review*, 5 August (www.technologyreview.com/2020/08/05/1006008/ai-face-recognition-hack-misidentifies-person/).

[86] R. Diresta (2020), "AI-Generated Text Is the Scariest Deepfake of All", *Wired*, 31 July (www.wired.com/story/ai-generated-text-is-the-scariest-deepfake-of-all/).

[87] Ibid.

[88] H. Ajde, G. Patrini, F. Cavalli and L. Cullen (2019), "The State of Deepfakes: Landscape, Threats, and Impact", DeepTrace, September, p. 1 (https://regmedia.co.uk/2019/10/08/deepfake_report.pdf).

report,[89] deepfakes and other audio and visual manipulation can be used with pornography to enact vendettas, blackmail people or trick them into participating in personalised financial scams. The increasing accessibility of this technology makes this even more problematic .[90] One recent example is the conjunction of 3D-generated porn and deepfakes, which allow a user to put a real person's face on another person's body, and do whatever violent or sexual act they want with them.[91] Notably, audiovisual manipulation and other less sophisticated methods such as basic video and photo-editing software (part of what Paris and Donovan call 'Cheap Fakes' or 'Shallowfakes'),[92] can also change audiovisual manipulation in malicious ways, much more easily and cheaply.[93]

Political

Deepfakes and cheap fakes also have malicious uses in political settings. Videos of heads of states saying things contrary to common belief have emerged in the past couple of years because of these technologies, and they are less and less easily differentiated from authentic videos. A recent example was when the incumbent UK Prime Minister, Conservative Boris Johnson, appeared to endorse his Labour Party rival, Jeremy Corbyn, and vice versa.[94] While those aware of the political context will see through this hoax, people less aware might believe them completely, and confusion and disorder is created in an important democratic process. This effect can be further maliciously exploited in countries where people have less digital literacy, and even more so as these technologies become more widely usable. In such a context comes Facebook's announcement that the company will remove misleading manipulated media whenever those *"have been edited or synthesized in ways that aren't apparent to an average person and would likely mislead someone into thinking that a subject of the video said words that they did not actually say",* and they are *"the product of artificial intelligence or machine learning that merges, replaces or superimposes content onto a video, making it appear to be authentic."*[95]

While deepfake videos could be spotted as such by countering software, they can still proliferate across social media networks in very little time, changing the course of a democratic election or even just one person's career.[96] Importantly, deepfakes can also be used as a

[89] B. Paris and J. Donovan (2019), "DeepFakes and Cheap Fakes: The Manipulation of Audio and Visual Evidence", Data and Society, September, p. 41 (https://datasociety.net/wp-content/uploads/2019/09/DS_Deepfakes_Cheap_FakesFinal.pdf).

[90] B. Paris and J. Donovan (2019), op. cit., p. 41.

[91] S. Cole and E. Maiberg, (2019), "Deepfake Porn Is Evolving to Give People Total Control Over Women's Bodies", *VICE*, 6 December (www.vice.com/en_us/article/9keen8/deepfake-porn-is-evolving-to-give-people-total-control-over-womens-bodies).

[92] H. Ajde, G. Patrini, F. Cavalli and L. Cullen (2019), op. cit., p. 1.

[93] Ibid., pp. 5-6

[94] S. Cole (2019), "Deepfake of Boris Johnson Wants to Warn You About Deepfakes", *VICE*, 13 November (www.vice.com/en_uk/article/8xwjkp/deepfake-of-boris-johnson-wants-to-warn-you-about-deepfakes).

[95] M. Bickert (2020), "Enforcing Against Manipulated Media", Facebook, 6 January (https://about.fb.com/news/2020/01/enforcing-against-manipulated-media/).

[96] A. Ridgway (2021), "Deepfakes: the fight against this dangerous use of AI", Science Focus, 12 November (www.sciencefocus.com/future-technology/the-fight-against-deepfake/).

scapegoat, often in political contexts, with people claiming that harmful video evidence has been altered when it has not. An example of this arose in 2018 when a married Brazilian politician claimed that a video allegedly showing him at an orgy was a deepfake, yet no one has been able to prove it so.[97] Similarly, when the Gabonese president Ali Bongo appeared on camera in a New Year's address at the end of 2018 to end speculation about his health, his political rivals claimed it was a deepfake. Yet experts have been unable to prove this.[98] Consequently, everyday voters and consumers of media need to be aware of the political impact of deepfakes as much as experts and politicians because very convincing fake videos can undermine trust in real ones in the eyes of the public.

Crime and cybersecurity

In recent years criminals have also made malicious use of deepfake technology for financial gain. According to Deeptrace, "Deep fakes do pose a risk to politics in terms of fake media appearing to be real, but right now the more tangible threat is how the idea of deep fakes can be invoked to make the real appear fake. The hype and rather sensational coverage speculating on deep fakes' political impact has overshadowed the real cases where deep fakes have had an impact", such as cybercrime.[99] While internet and email scams have been around for decades, the advance of deepfake technology in sound and video has allowed for even more intricate and hard-to-spot fraudulent criminal activity.

These sorts of crimes could range from a basic level of hacktivists making false claims and statements to undermine and destabilise a company, to more serious efforts such senior executives confessing to financial crimes or other offences. Deepfakes can also use social engineering to make frauds more credible by using video or audio of, for instance, a member of the targeted organisation, increasing the chances of the attacks succeeding.[100] Market Research company Forrester has claimed that deepfakes could end up costing businesses as much as $250 million in 2020.[101] Software tools that can spot criminal deepfakes are being developed, but it only takes one individual in a company to believe in a modified audio or visual source for a large amount of damage to be done.

Military

Concern about deepfakes has also reached hard security, with many of the world's militaries now being very worried about them. In 2018, funding began for a US DARPA project that will try to determine whether AI-generated images and audio are distinguishable, using both technological

---

[97] D. Thomas (2020), "Deepfakes, a Threat to Democracy or Just a Bit of Fun?", *BBC News*, 23 January (www.bbc.com/news/business-51204954).

[98] H. Ajde, G. Patrini, F. Cavalli and L. Cullen (2019), op. cit., p. 10.

[99] Orange Business Services (2020), "Fake news: What could deepfakes and AI scams mean for cybersecurity?", Orange, 2 January (www.orange-business.com/en/magazine/fake-news-what-could-deepfakes-and-ai-scams-mean-cybersecurity).

[100] C. Meziat et al. (2020), "Deep Dive into Deepfake – how to face increasingly believable fake news? (1/2)", Wavestone, 5 May (www.riskinsight-wavestone.com/en/2020/05/deep-dive-into-deepfake-how-to-face-increasingly-believable-fake-news-1-2/).

[101] Orange Business Services (2020), op. cit.

and non-technological means.[102] Legal witnesses and AI experts told US lawmakers in June 2019 that they needed to act immediately to stay ahead of the threat of deepfakes and other AI-led propaganda, which could be deployed by adversaries such as Russia ahead of the next presidential election. These efforts are ongoing, with the US Congress greenlighting a $5 million programme to boost new technologies in detecting deepfakes. This reveals that the Pentagon views audiovisual manipulation as a key national security issue. Examples of potential problems include a national security leader giving false orders or acting unprofessionally, which could cause chaos.[103] Todd Myers, automation lead for the CIO-Technology Directorate at the National Geospatial-Intelligence Agency, believes that China is the main proactive user of deepfake technology for military reasons, specifically by creating fake bridges in satellite images. "From a tactical perspective or mission planning, you train your forces to go a certain route, toward a bridge, but it's not there. Then there's a big surprise waiting for you," he warns.[104]

Finally, as mentioned throughout this analysis, one of the greatest threats of deepfakes to both public and private life is not the technology itself but to its potential to converge with other technologies and bring about new and unexpected challenges. By compounding different technologies, state and non-state actors will be able to further propagate misleading or false narratives, targeting harmful and disruptive content at specific populations with deepfake, IoT, and AI capabilities.[105] It is difficult to predict exactly how this issue of convergence will pan out, but it leaves a lot of room for devastating malicious uses should governments, private companies, and individuals fail to educate and prepare themselves against such threats.

3.2.2 Breaking CAPTCHAs

Completely Automated Public Turing tests to tell Computers and Humans Apart (CAPTCHAs) were created to preclude automatised programs from being malicious on the world wide web (filling out online forms, accessing restricted files, accessing a website an incredible number of times, etc.) by confirming that the end-user is in fact 'human' and not a bot. Today, machine learning is able to break CAPTCHAs in 0.05 seconds, using GAN. Indeed, synthesised CAPTCHAs can be created, along a small dataset of real CAPTCHAs, to create an extremely fast and accurate CAPTCHA solver.[106]

---

[102] W. Knight (2018), "The US military is funding an effort to catch deepfakes and other AI trickery", *MIT Technology Review*, 23 May (www.technologyreview.com/2018/05/23/142770/the-us-military-is-funding-an-effort-to-catch-deepfakes-and-other-ai-trickery/).

[103] J. Keller (2020), "U.S. intelligence researchers eye $5 million program to encourage new technologies in detecting deepfakes", *Military and Aerospace Electronics*, 8 January.

[104] P. Tucker (2019), "The Newest AI-Enabled Weapon: 'Deep Faking' Photos of the Earth", Defense One, 31 March (www.defenseone.com/technology/2019/03/next-phase-ai-deep-faking-whole-world-and-china-ahead/155944/).

[105] M. Erfourth and A. Bazin (2020), "Extremism on the Horizon: The Challenges of VEO Innovation", Mad Scientist Laboratory, 19 March (https://madsciblog.tradoc.army.mil/220-extremism-on-the-horizon-the-challenges-of-veo-innovation/).

[106] R. Irioni (2018), "Breaking CAPTCHA using Machine Learning in 0,05 Seconds", Medium, 19 December (https://medium.com/towards-artificial-intelligence/breaking-captcha-using-machine-learning-in-0-05-seconds-9feefb997694) and E. Zouave et al. (2000), "Artificial Intelligence Cyberattacks", FOI, p. 24.

3.2.3 Swarming attacks

AI systems could be used to control robots and malware behaviour that would be impossible for humans to do manually. This could allow 'swarming attacks' by distributed networks of autonomous robotic systems cooperating at machine speed, such as autonomous swarms of drones with facial recognition.[107]

3.3 Changes to the typical character of threats and new forms of vulnerabilities on AI systems

Inherent AI vulnerabilities are creating new and uncertain security risks adding a layer of complexity in the mapping of the threat landscape.

The Cybersecurity@CEPS Initiative has delved into the maze of responsible disclosure of software vulnerabilities. A 2018 report evidenced the widespread and traditional pattern of cyberattack prevention, whereby researchers analyse the lines of code of a determined software program in order to find errors and patch them.[108] Such errors are mostly dependent on intentional or unintentional bad programming. This practice has been the longstanding baseline approach in cybersecurity.

The AI and Cybersecurity Task Force has instead brought to light an additional layer of complexity in the potential attacks targeting AI systems. This was also revealed by the most recent literature on the security of AI, which describes how AI attackers can exploit several different types of vulnerabilities. Figure 5 presents an AI architecture in which each stage, such as initial data inputs, data-conditioning processes, algorithms, and human-machine teaming, represents an attack surface that is potentially vulnerable to cyberattacks. In this architecture it is not just software or hardware that can be attacked as in traditional IT systems, but also the data that are a critical element of all AI systems. AI systems, particularly those deploying machine learning, are not solely embedding traditional forms of cyber vulnerabilities. The existing attack surface composed of coding errors is in fact complemented by additional, seemingly *unpatchable* ones, which are inherently dependent on the way AI functions and learns, and result from the sophistication of the underlying AI technology. It is for this reason that the whole information technology system is rendered more open to attacks.[109]

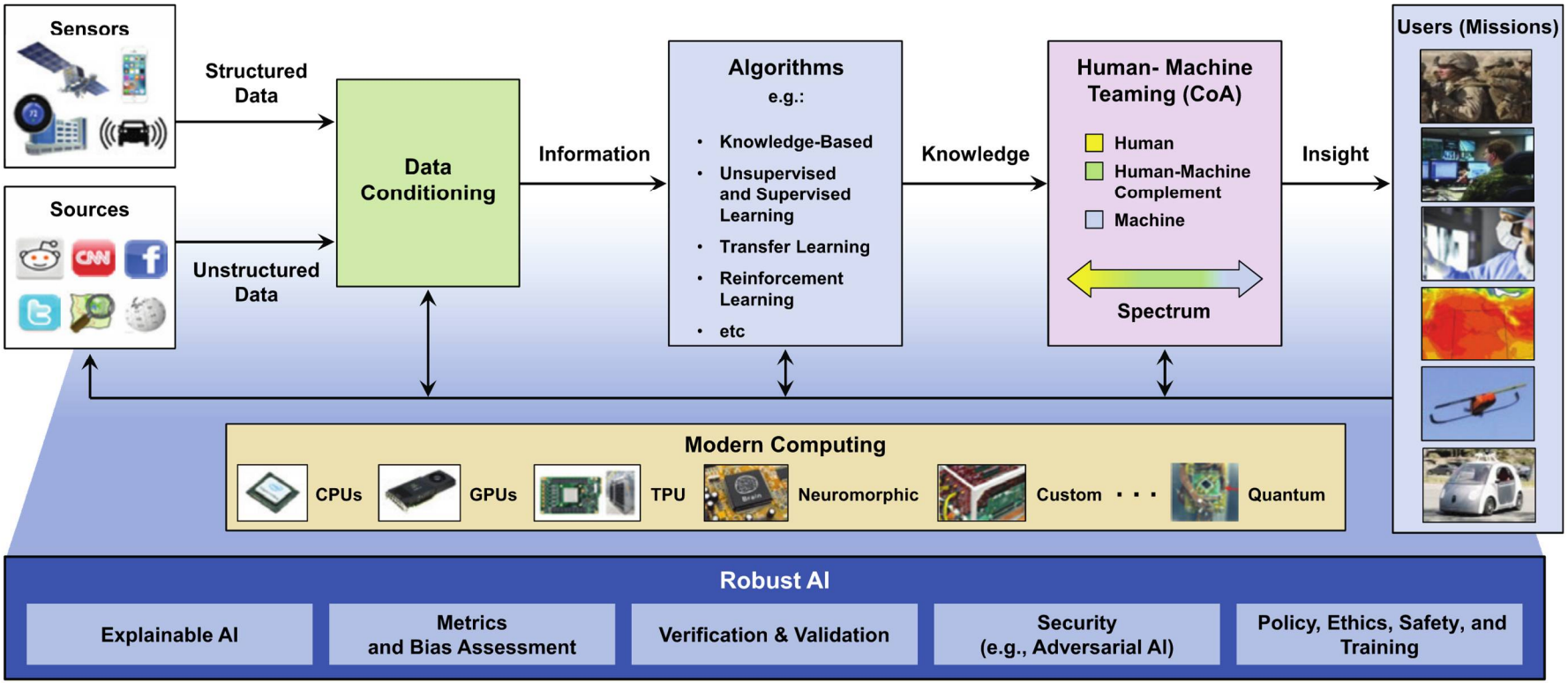
---

[107] See Andrea Renda (2019), op. cit., p. 22.

[108] L. Pupillo, A. Ferreira and G. Varisco (2018), *Software Vulnerability Disclosure in Europe: Technology, Policies and Legal Challenges*, Report of a CEPS Task Force, CEPS, Brussels.

[109] *"One example is the training data which can be manipulated by attackers to compromise the machine learning model. This is an attack vector that does not exist in conventional software as it does not leverage training data to learn. Additionally, a substantial amount of this attack surface might be beyond the reach of the company or government agency using and protecting the system and its adjacent IT infrastructure. It requires training data potentially acquired from third parties which, as mentioned, can already be manipulated."*, S. Herping (2019), "Securing Artificial Intelligence", Stiftung Neue Verantwortung, October, p. 2.

Figure 5. Schematic representation of the AI architecture and its attack surface



Source: National Academies of Sciences, Engineering, and Medicine, (2019), "Implications of Artificial Intelligence for Cybersecurity: Proceedings of a Workshop", Washington, DC: The National Academies Press, p. 201.

Classic cybersecurity attacks mainly aim to steal data (extraction) and disrupt a system. Attacks against AI systems also often aim to steal information or disrupt the system but are crafted in a more subtle form and for longer-term orientation. They try to acquire control of the targeted system for a given intent or get the model to reveal its inner workings by intrusion into the system and then change its behaviour.[110] This goal can be achieved through mainly, but not exclusively, four types of attacks: data poisoning, tempering of categorisation models, backdoors, and reverse engineering of the AI model. Table 2 provides a more detailed overview of possible AI attacks.

Table 2. Intentionally motivated ML failure modes

Attack	Overview
Perturbation attack	Attacker modifies the query to get appropriate response
Poisoning attack	Attacker contaminates the training phase of ML systems to get intended result
Model Inversion	Attacker recovers the secret features used in the model through careful queries
Membership Inference	Attacker can infer whether a given data record was part of the model's training dataset
Model Stealing	Attacker is able to recover the model through carefully crafted queries
Reprogramming ML system	Repurpose the ML system to perform an activity it was not programmed for
Adversarial Example in Physical Domain	Attacker brings adversarial examples into the physical domain to subvert ML system e.g., 3D printing special eyewear to fool facial recognition system
Malicious ML provider recovering training data	Malicious ML provider can query the model used by customer and recover customer's training data
Attacking the ML supply chain	Attacker compromises the ML models as it is being downloaded for use
Backdoor ML	Malicious ML provider backdoors algorithm to activate with a specific trigger
Exploit Software Dependencies	Attacker uses traditional software exploits like buffer overflow to confuse/control ML systems

Source: R. Shankar, S. Kumar, D. O'Brien, J. Snover, K. Albert and S. Viljoen (2019), "Failure Modes in Machine Learning", Microsoft, November (https://docs.microsoft.com/en-us/security/engineering/failure-modes-in-machine-learning#intentionally-motivated-failures-summary).

[110] This section draws from M. Taddeo, T. Cutcheon and L. Floridi (2019), op. cit.

1. *Data poisoning*: Attackers may bring carefully crafted flawed data into the legitimate dataset used to train the system to modify its behaviour. It has been shown that by adding 8% of erroneous data to an AI system for drug dosage, attackers could generate a 75.06% change in the dosage of half of the patients using the system for their treatment.[111]

2. *Tampering of categorisation models:* By manipulating the categorisation models of e.g. neural networks, attackers could modify the final outcome of AI system applications. For instance, researchers using pictures of 3D printed turtles, obtained using a specific algorithm, were able to deceive the learning method of an AI system and classify turtles as rifles.[112]

3. *Backdoors*: Adversaries can also compromise AI systems through backdoor injection attacks. To perform such attacks, the adversary creates a *"customized perturbation mask applied to selected images"* to override correct classifications. *"The backdoor is injected into the victim model via data poisoning of the training set, with a small poisoning fraction, and thus does not undermine the normal functioning of the learned deep neural net"*. Hence, such attacks, once triggered, *"can exploit the vulnerability of a deep learning system in a stealthy fashion, and potentially cause great mayhem in many realistic applications – such as sabotaging an autonomous vehicle or impersonating another person to gain unauthorized access."*[113] This is the case where, for instance, a No Entry sign is instead perceived as an Ahead Only sign.

4. *Reverse engineering the AI model:* By gaining access to the AI model through reverse engineering, attackers are able to perform more targeted and successful adversarial attacks. For example, according to a study published by the Institute of Electrical and Electronics Engineers (IEEE), following the Differential Power Analysis methodology, an adversary can target the ML inference, assuming the training phase is trusted, and learn the secret model parameters.[114] As such, the adversary is able to build knockoffs of the system and put security and intellectual property at risk.

---

[111] See M. Jagielski et al. (2018) "Manipulating Machine Learning: Poisoning Attacks and Countermeasures for Regression Learning", 39th IEEE Symposium on Security and Privacy, San Francisco, CA, USA, 21-23 May, 2018, April (https://arxiv.org/abs/1804.00308).

[112] A. Athalye, L. Engstrom, A. Ilyas and K. Kwok (2018), "Synthesizing Robust Adversarial Examples", *International conference on machine learning*, July, pp. 284-293.

[113] C. Liao et al. (2018), "Backdoor Embedding in Convolutional Neural Network Models via Invisible Perturbation", August (http://arxiv.org/abs/1808.10307).

[114] In particular, according to the study, *"the adversary can and is able to make hypotheses on the 4 bits of a neural network weight. For all these 16 possible weight values, the adversary can compute the corresponding power activity on an intermediate computation, which depends on the known input and the secret weight. This process is repeated multiple times using random, known inputs. The correlation plots between the calculated power activities for the 16 guesses and the obtained power measurements reveal the value of the secret weight."*, A. Dubey, R. Cammarota and A. Aysu (2020), "MaskedNet: The First Hardware Inference Engine Aiming Power Side-Channel Protection", 2020 IEEE International Symposium on Hardware Oriented Security and Trust (HOST), IEEE, pp. 197-208.

Unlike traditional cyberattacks that exploit bugs or intentional and unintentional human mistakes in code, it's clear that these AI attacks fundamentally expand the set of entities, including physical objects, that can be used to execute cyberattacks.

4. Ethical considerations related to AI in cybersecurity[115]

The role that AI could play for system robustness, response, and resilience brings ethical challenges with it that could hamper its adoption in cybersecurity. Furthermore, if the issues are not properly addressed through governmental processes and policies, it could create significant problems for our societies. Table 3 presents the main ethical challenges.

Table 3. AI ethical challenges

	Ethical challenges
System robustness	*Control* Who is controlling the AI system? How do we make sure that the system is behaving according to our expectations?
System response	*Responsibilities* How do we ascribe responsibilities for autonomous response systems?
System resilience	*Skills* If we delegate threat detection to machines, how do we make sure our analysts will still be able to do it?

Source: contribution of M. Taddeo to the second meeting of the CEPS Task Force, as a member of its Advisory Board.

System robustness. As mentioned at the beginning of this chapter, system robustness is improved by using AI for software testing and for designing software that is capable of self-testing and self-healing. Self-testing refers to "the ability of a system or component to monitor its dynamically adaptive behaviour and perform runtime testing prior to, or as part of the adaptation process."[116] In this context, therefore, AI can enable continuous systems verification and optimisation of their state and respond quickly to changing conditions, making the proper correction. But who is controlling the AI system? In fact, although Article 22 of the GDPR states that no major decision regarding an individual, such as profiling, must be taken solely by an autonomous system,[117] it remains unclear, and it is left to organisations to decide where the human control ends and automation begins. Furthermore, it remains unclear how to make sure that the system is behaving according to the expectations.

[115] See M. Taddeo (2019), "The Ethical Challenges of AI in Cybersecurity", *Minds and Machines*, Vol. 29, No. 2. This section draws from this paper and from Mariarosaria Taddeo's contribution, as a member of the Advisory Board of the Task Force, to the second meeting of the CEPS Task Force.

[116] See T.M. King et. al. (2019), "AI for testing today and tomorrow: Industry Perspective", IEEE International Conference on Artificial Intelligence Testing, IEEE, p. 81-88.

[117] European Commission, Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), OJ 2016 L 119/1, hereinafter referred to as the GDPR.

Applying principal-agent theories in the context of AI is therefore more challenging. Agency theory defines a distinction between the owners of an organisation and the management of the organisation, whereby the management (the agent) has different objectives and goals than the owner (the principal). As a result, the owner receives a lesser return on investment since they do not manage the company themselves. This has been defined as the principal-agent problem and is a dilemma where the agent acts in their own best interest, which may be contrary to those of the principal. Ways to overcome the principal-agent problem have been envisaged, such as providing strategic and financial control methods, succession planning or monetary rewards, and coaching or mentoring. Those methods, nonetheless, are hardly more applicable, or not applicable at all, when the agent is artificial. Delegating control to the AI system may also lead to errors and increase the risks for unforeseen consequences and must be balanced appropriately, while envisaging some form of human oversight in any case.

System response. As aforementioned, AI can greatly improve systems' response, e.g. by automatically patching vulnerabilities. In the same fashion, it can also afford offensive options to threat response. There are autonomous and semi-autonomous cybersecurity systems that offer a set of predetermined responses to a specified activity allowing the deployment of specific offensive responses. *"AI can refine strategies and launch more aggressive counter operations, which could easily escape the control of its users and the intentions of the designer. This could snowball into an intensification of cyber-attacks and counter responses, which, in turn pose a serious risk of escalation into kinetic conflict,"* threatening key infrastructures of our societies.[118] While adding a human layer will inevitably cause delay in such responses, from a societal point of view this situation nonetheless raises the issue of responsibility: how do we ascribe responsibilities for autonomous response systems? How do we promote responsible behaviour in this context? Do we need to enforce and ensure proportionality of response, clear definition of legitimate actors and targets by regulation?

System resilience. AI is heavily used for TAD. AI systems are very good at finding vulnerabilities and identifying malware and anomalous behaviours. Indeed, they do that in less time and in a more effective way than security analysts could.[119] However, delegating threat detection completely to the AI systems would be a mistake since it could lead to a widespread deskilling of experts. Even at the state of the art of the technology, AI systems are still not able to fully understand complex attacks and threats. Human interaction is required to assess AI outcomes, to combine alerts, to reconstruct the attack that took place, to identify options for responses and to assess and select the best response. In this context, cybersecurity experts should keep finding vulnerabilities and detecting threats in the same way that radiologists need to keep reading X-rays or pilots landing aeroplanes, so that they are still able to do it if AI fails or gets it wrong.[120] It is interesting to note that, in the past few years, the US Navy has started to teach

[118] See M. Taddeo (2019), op. cit.

[119] Ibid.

[120] See G. Yang et al. (2018), "The Grand Challenge of Science Robotics", *Science Robotics*, Vol. 3, No. 14 (www.researchgate.net/publication/322859319_The_grand_challenges_of_Science_Robotics).

sailors to navigate by the stars again amid growing fears of cyberattacks on electronic navigation systems.[121]

5. Asymmetries in the interplay of AI and cybersecurity[122]

The AI and Cybersecurity Task Force highlighted three types of societal and geopolitical asymmetries in the interplay of AI and cybersecurity:

5.1 Asymmetry of cognition

Not everybody is equipped in the same educational and normative way to think critically about AI, and so asymmetry of cognition happens. This is much more than a matter of personal understanding. The asymmetry of cognition also derives from a lack of general public awareness and the lack of consensus in terminology and definitions, which is caused by governance mismanagement in governments, technology providers and international organisations. In this context, there is a growing danger of anthropomorphism of AI: it's a natural tendency to attribute human-like capabilities to a machine that mimics us. AI anthropomorphism could also be a deliberate strategy by some actors (technology companies, academia, or governments) to divert attention from more granular issues, for example, talking about General Intelligence to divert the discussion from accountability. Another example is Saudi Arabia granting citizenship to Sophia, a robot resembling a human being.[123] AI anthropomorphism can create situations where expectations about the machines' efficiency and outcomes are overblown; guards are lowered and new loopholes in cybersecurity are created.

AI anthropomorphism is also related to trust, defined as a form of delegation with no control. As more people trust AI systems, there is the fear that more people could lose their critical thinking about the recommendations and even decisions taken by AI. For instance, there have been countless car accidents where drivers followed their Global Positioning System (GPS) instructions without critically questioning the instructions they were given.[124] These asymmetries of cognition raise important questions about, for example, how prepared future generations will be to cope with critical perspectives on digital devices, and whether there is a risk of them anthropomorphising these technologies even more. Indeed, the generation born in the non-digital world maintains a point of reference that will be missing for future generations. As such, the education system should be better equipped than it is today to cope

[121] Seaman S. Apprentice and Jordan Ripley (2019), "Navigating by the Stars", July (www.dvidshub.net/news/309792/navigating-stars).

[122] This section draws on the contribution of Jean-Marc Rickli, Geneva Centre for Security Policy (GCSP) and member of the Advisory Board of the Task Force in the CEPS Task Force meeting, 29.06.20.

[123] A. Griffin (2017), "Saudi Arabia Grant Citizenship to a Robot for the first time ever." *The Independent*, 26 October, (www.independent.co.uk/life-style/gadgets-and-tech/news/saudi-arabia-robot-sophia-citizenship-android-riyadh-citizen-passport-future-a8021601.html).

[124] J. Thornhill (2020), "Trusting AI too much can turn out to be fatal," *Financial Times*, 2 March, (www.ft.com/content/0e086832-5c5c-11ea-8033-fa40a0d65a98).

with the social problems deriving from these risks. The implications of AI automorphisms and of delegating tasks to AI with no control will be further explored in this report.

5.2 Asymmetry in AI ethical standards development

Initiatives dealing with AI ethical standards have proliferated over the past few years.[125] About 85 principles on AI ethics have been released by governments and companies since 2016.[126] This is because the new challenges emerging from autonomous actions made possible by AI has left a regulatory vacuum. According to Jobin et al., while there is an overall convergence of five ethical principles (transparency, justice and fairness, non-maleficence, responsibility, and privacy), there are nonetheless substantive divergences in relation to *"how these principles are interpreted; why they are deemed important; what issue, domain or actors they pertain to; and how they should be implemented."*[127]

Furthermore, because of the impact that these principles will have on the working of future systems and applications, the normative aspect is closely related to geoeconomic and geostrategic factors. The interpretative stances that will gather more traction will also help define the future of AI. This creates a geopolitical race to establish the main international ethical standards that will likely generate and increase power asymmetries as well as normative fragmentation. It follows that the international community should be aware of the extent to which actors who, while not at the forefront of AI development are yet having a say in its decision-making and governance. Hence, it would be pivotal for the European Union to define an ethical framework beforehand, prescribing how to foresee, prevent and ascribe accountability for unintended consequences of a system that makes decisions in an unsupervised way. One of the reasons self-driving cars are not massively populating the streets is because we do not know how to insure them. It is not clear how to ascribe responsibilities, given that our moral framework does not envisage distributed integrated systems. Eventually, as will be further explained, this should entail going beyond the current provisions for trustworthiness of the AI systems by establishing sufficient forms of control to mitigate against the risks AI poses.

5.3 Offence/defence asymmetry

On any given day, millions of cyberattacks are occurring worldwide. In 2019, more than 11,000 exploitable vulnerabilities in commonly used systems and software were identified, of which a third had no patches available.[128] According to F-Secure, the number of attack events measured during the six months between January and June 2019 was twelve times higher than a similar

[125] See the OECD AI Policy Initiative for an exhaustive selection of AI non-governmental and intergovernmental initiatives (https://oecd.ai/countries-and-initiatives/stakeholder-initiatives).

[126] K. Johnson (2020), "How Microsoft, Open AI, and OECD are Putting AI Ethics Principles into Practice," *VentureBeat*, 6 May (https://venturebeat.com/2020/05/06/how-microsoft-openai-and-oecd-are-putting-ai-ethics-principles-into-practice/).

[127] A. Jobin, M. Ienca and E. Vayena (2019) "Artificial Intelligence: the global landscape of ethics guidelines", *Nature Machine Intelligence,* Vol. 1, No. 9, pp. 389-399.

[128] J. Fruhlinger (2020), "Top Cybersecurity Facts, Figures and Statistics for 2020," CSOOnline, 9 March (www.csoonline.com/article/3153707/top-cybersecurity-facts-figures-and-statistics.html).

period the year before,[129] and these trends have even accelerated because of the Covid-19 crisis, as threat actors have managed to exploit the panic and discomfort caused by the pandemic to conduct specially crafted malware and phishing attacks worldwide.[130]

According to most scholars, the nature of the cyber world seems to favour the offensive.[131] As this report shows, it is expected that AI will be able to reduce the gap between the offensive advantage and defence in cybersecurity. The rationale is that AI will be able to patch vulnerabilities before they have even been identified by human beings. Nonetheless, the reliance on AI involves its own specific vulnerabilities. Moreover, AI itself could also be used to seek offensive advantages. Also, AI will enable technology to increasingly become a surrogate that will work in the attacker's interests, such as in the case of the 2019 US cyber offensive operation against Iran.[132] Similarly, the access to technology will also determine who will be able to develop technological surrogates and seek offensive advantage. Thus, the issue of asymmetry between the offensive and the defensive and whether the use of AI will, in the cybersecurity paradigm, favour offence or defence remains rather unclear. The offence-defence debate is further analysed below.

6. Trustworthy versus reliable AI[133]

As explained earlier, once launched, attacks on AI are hard to detect. While extensive research has been carried out to further the understanding of the AI decision-making process, the networked, dynamic, and learning nature of AI systems makes it problematic to explain their internal processes (this is known as lack of *transparency*). It is difficult to reverse-engineer their behaviour to understand what exactly has determined a given outcome, and whether this is due to an attack, and of which kind. Furthermore, it may be difficult to understand when the compromised system is showing 'wrong' behaviour, because a skilfully crafted attack may determine only a minimal divergence between the actual and the expected behaviour. For example, Comiter highlighted AI attacks against content filters: "*unlike many other cyberattacks in which a large-scale theft of information or system shutdown makes detection evident, attacks on content filters will not set off any alarms. The content will simply fall through the filter unnoticed*".[134]

This is why it is crucial to ensure the robustness of an AI system – so that it continues to behave as expected even when its inputs or model are perturbed by an attack. Unfortunately, assessing the robustness of a system requires testing for all possible input perturbations. This is almost impossible for AI, because the number of possible perturbations is often exorbitantly large. For

---

[129] M. Michael (2019), "Attack Landscape H1 2019: IoT, SMB Traffic Abound," *F-Secure*, 12 September (https://blog.f-secure.com/attack-landscape-h1-2019-iot-smb-traffic-abound/).

[130] See WebARX, COVID-19 Cyber Attacks (www.webarxsecurity.com/covid-19-cyber-attacks/).

[131] J.-M. Rickli (2018), *The impact of autonomy and artificial intelligence on strategic stability*, UN Special, July-August, pp. 32-33 (www.unspecial.org/2018/07/the-impact-of-autonomy-and-artificial-intelligence-on-strategic-stability/).

[132] A. Krieg and J.-M. Rickli (2019), *Surrogate Warfare: the Transformation of War in the 21st Century*, Washington: Georgetown University Press, Chapter 4.

[133] This section of the report was contributed by Mariarosaria Taddeo, from Oxford Internet Institute, University of Oxford, the Alan Turing Institute, London, and member of the Task Force Advisory Board.

[134] M. Comiter (2019), op. cit., p. 35.

instance, in the case of image classification, imperceptible perturbations at pixel level can lead the system to misclassify an object with high-level confidence.[135] Alongside this characteristic intrinsic to AI systems, attackers are usually specifically trying to get around the status quo. So, it turns out that assessing the robustness of AI is often a computationally intractable problem; it is not feasible to exhaustively foresee all possible erroneous inputs to an AI system, and then measure the divergence of the related outputs from the expected ones. If AI robustness cannot be assessed nor can its trustworthiness.

Philosophical analyses define trust as the decision to delegate a task without any form of control or supervision over the way the task is executed.[136] Successful instances of trust rest on an appropriate assessment of the trustworthiness of the agent to whom the task is delegated (the trustee). Hence, while trust is the confidence in some person or quality, trustworthiness is the state or quality of being trustworthy.[137] 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 'truster', 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.

Notably, in cybersecurity, a fundamental principle is that trustworthiness should be maximised over trust.[138] While this remains valid, even trustworthiness is challenging with AI as there are fewer chances to conduct a formal validation of an AI system.

The lack of transparency and the learning abilities of AI systems, along with 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.

And as long as the assessment of trustworthiness remains problematic, trust is unwarranted.

Records of past behaviour of AI systems are neither predictive of the system's robustness to future attacks, nor an indication that the system has not been corrupted by a dormant attack 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 is unwarranted.

---

[135] C. Szegedy et al. (2013), "Intriguing properties of neural networks", *arXiv:1312.6199 [cs]*, December, (http://arxiv.org/abs/1312.6199) and J. Uesato, B. O'Donoghue, A. van den Oord and P. Kohli (2018), "Adversarial Risk and the Dangers of Evaluating Against Weak Attacks", *arXiv:1802.05666 [cs, stat]*, February (http://arxiv.org/abs/1802.05666).

[136] M. Taddeo (2010), "Modelling Trust in Artificial Agents, A First Step Toward the Analysis of e-Trust", *Minds and Machines*, Vol. 20, No. 2, June, pp. 243–257 and G. Primiero and M. Taddeo (2012), "A modal type theory for formalizing trusted communications", *Journal of Applied Logic*, Vol. 10, No. 1, March, pp. 92–114.

[137] M. Becerra et al. (2008), "Trustworthiness, Risk, and the Transfer of Tacit and Explicit Knowledge Between Alliance Partners", *Journal of Management Studies*, Vol. 45, No. 4, p. 696.

[138] D. Basin et al. (2011), *Applied Information Security*: *A Hands-on Approach*, Springer Science & Business Media (www.springer.com/gp/book/9783642244735#).

Clearly, while delegation of 3R tasks can and should still occur, some forms of control are necessary to mitigate the risks linked to the lack of transparency of AI systems and the lack of predictability of their robustness. Hence, the following developing and monitoring practices to address these risks are suggested, and will be further elaborated when discussing the development and deployment of reliable AI in the policy chapter:

1. Companies' *in-house development* of AI applications models and testing of data
2. Improving AI systems' robustness through *adversarial training* between AI systems
3. *Parallel and dynamic monitoring or punctual checks* of AI systems through a clone system as control, to be used as baseline against which the behaviour of the original system should be assessed.[139]

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. Conceptually, the distinction lies in the opportunity for the trustee to act against the wishes of the truster and in the trustee's consideration of the value of the trust that has been placed in them by the truster. As such, aligning trustworthiness with reliability removes the virtue from being trustworthy.[140]

Supporting the reliability of AI 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 computational feasibility.

AI systems are autonomous, self-learning agents interacting with the environment.[141] 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 take the dynamic and self-learning nature of AI systems into account 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.[142] In view of this, defining standards for AI in cybersecurity that seek to elicit trust (and thus forgo monitoring and control of AI) or that focus on outcomes alone, is risky. The sooner we focus on standards and certification procedures on developing *reliable* AI, and the more we adopt an *in house*, *adversarial*, and *always on* strategy, the safer and more secure AI applications for 3R will be.

---

[139] These recommendations will be discussed in more detail in Chapter 4 of this report.

[140] S. Wright (2010), "Trust and Trustworthiness", *Philosophia*, Vol. 38, p. 615–627.

[141] G.-Z. Yang et al. (2018), "The grand challenges of Science Robotics", *Science Robotics*, Vol. 3, No. 14, January.

[142] OECD (2019b), "Recommendation of the Council on Artificial Intelligence", May (https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449).

7. Cybersecurity risks associated with anthropomorphising AI[143]

Assuming we confine ourselves to considering only AI that is implemented as digital technology, then like every other digital technology, AI is subject to security concerns at every step of its development, distribution, and use. Unfortunately, the inexcusable levels of sloppiness rife in the software development industry are in some cases even more pervasive in technology termed 'intelligent'. This is posited to be because our understanding of the term *intelligence* hinges largely on our human identity, and thus the more intelligent a system is perceived to be, the more likely it is to be expected to not require the ordinary considerations of engineering, but rather of psychology and even human rights. This series of misattributions is termed *anthropomorphism*, meaning it is attributed human qualities. Anthropomorphism is a substantial area of risk for AI.

In comparative psychology, one extremely well-established definition of intelligence (dating from 1883)[144] is *"the ability to do the right thing at the right time, to recognize an opportunity or a crisis in an environment and to then do something about it."* Hence, intelligence is a form of computation. Computation is the systemic transformation of information from one state to another; intelligence is the transformation of information describing a context into an action. Artificial intelligence is the same, only expressed by an artefact, that is, something designed, built, and operated by some human organisation for a purpose.

Such a definition of intelligence relates to the philosophical distinction between *strong* and *weak* AI, and to John Searle's Chinese Room counterargument to the Turing test,[145] that strong AI would amount to a genuinely self-conscious system, whereas AI systems that exist today can actually only exhibit intelligent *behaviours*.

To the extent that AI is a digital artefact, every stage of that process is potentially vulnerable to attack. Design and development, and records of the decisions and processes that produced the artefact, are all subject to deliberate or careless interference. Although AI development relies heavily on libraries, and access to software codes has become an increasingly easy task, software libraries sourced without sufficient care may, as examined in this chapter, introduce backdoors. Anyone from an individual developer up to a company director might choose to include features or attributes that compromise the system's integrity or are contrary to its stated purpose. If machine learning is used as part of the development process, then the

[143] This section of the report was contributed by Joanna Bryson, Professor of Ethics and Technology at the Hertie School of Governance, Berlin, and member of the Task Force Advisory Board.

[144] G. J. Romanes (1883), *Animal Intelligence*, New York: D. Appleton.

[145] Searle describes an experiment in which a person who doesn't know Chinese is locked in a room. Outside the room is a person who can slip notes written in Chinese into the room through a mail slot. The person inside the room is given a big manual where she can find detailed instructions for responding to the notes she receives from the outside. Searle argued that even if the person outside the room gets the impression that he is in a conversation with another Chinese-speaking person, the person inside the room does not understand Chinese. Likewise, his argument continues, even if a machine behaves in an intelligent manner, for example, by passing the Turing test, it doesn't follow that it is 'intelligent' or that it has a 'mind' in the way that a human has. The word 'intelligent' can also be replaced by the word 'conscious' and a similar argument can be made. See Philosophy of AI, Elements of AI (https://course.elementsofai.com/1/3).

libraries of data as well as the libraries of source code need to have secure provenance, and to be proofed for other potential problems, such as biased sampling.

To date, although of course it should be, software development as a process may not be widely understood by corporate executives, regulators or governors, at least to the degree that any other process of manufacturing products or utilities is understood. But even within the development community, it has been noted that while software engineering has improved its standard of development and operations ('DevOps') over the past few decades, AI has often somehow been left behind. Developers are time-constrained, or don't bother to record model parameters used to achieve core results at all. The difference may be cultural, as some AI developers come from other cognate disciplines such as the quantitative social sciences, but it seems such a pervasive phenomenon that it is likely instead to be psychological. If an entity is seen as intelligent, it is expected to somehow learn for itself like a human would; that the systems scaffolding this learning process are engineering is frequently overlooked.

If even those who engineer AI are diverted from ordinary good practice by overidentification with the artefacts they create, then we cannot be surprised that ordinary consumers are all the more so untroubled by the security issues of having an object with cameras and microphones in their intimate household spaces, whether their office, their dining room, or their children's bedroom. Anthropomorphism may also be a deliberate malfeasance, committed for example by corporations attempting to evade regulation. If a corporation claims that their machine-learning algorithm is unknowable just like a human brain, that seems feasible on first impressions.

However, unlike a brain, a neural network is designed: its model is selected, as is the data to train it; as are the parameters that seem to generate the best fit of the model given that data; as are the tests used to determine when training and parameter-setting are completed or at least good enough for release…

Certainly, there is no procedure for understanding the precise semantics of every weight in a neural network, any more than for understanding a single synapse in a human brain. However, unlike a brain, a neural network is designed: its model is selected, as is the data to train it; as are the parameters that seem to generate the best fit of the model given that data; as are the tests used to determine when training and parameter-setting are completed or at least good enough for release. All of those decisions are auditable; best and poor practice can be established and checked for at every stage. Yet hand waving and talk about brains or, worse still, consciousness, have long been deployed as means to evade regulation and oversight.[146]

[146] See for example J. J. Bryson, M. E. Diamantis and T. D. Grant (2017), "Of, for, and by the people: The legal lacuna of synthetic persons", *Artificial Intelligence and Law*, Vol. 25, No. 3, pp. 273–291.

7.1 Deanthropomorphising and demystifying AI

To maintain not only appropriate cybersecurity but also accountability (again) in the deployment of intelligent technology, it is important to make clear that the system is an artefact. Once again, the focus should be on drafting standards and certification procedures on developing *reliable* rather than *trustworthy* AI. Trust, to recall the definition, is the decision to delegate a task without any form of control or supervision over the way the task is executed. In other words, trust is a relationship between peers who cannot micromanage one another. In this respect, responsibility is held by moral and legal agents. AI systems cannot be held the responsible agent because no penalty against them would motivate them to improve their behaviour. The nature of responsibility is that it can only be enforced by humans against humans, where humans will be dissuaded by loss of resource, status, or liberty.[147]

AI is being presented by some consulting companies and others as a sort of employee that human employees will need to learn how to train and work with. This is a very poor model to communicate cybersecurity risks if we want human employees to be aware of and defend against them, not to mention defending their own employment rights.

Another myth of AI is that it is generated by machine learning from data. Data may be used to train aspects of an AI system, but the quality of the results produced is *not* necessarily strictly proportional to the amount of data used, unless the purpose of the system is surveillance. This does not mean that data are not pivotal for ML systems' development and functioning, but ML is a statistical method of programming, and the amount of data required by statistics is strictly dependent on the amount of variation in that data. Redundant data is an unnecessary exposure to cybersecurity attack. Maintaining data (or its equivalent) that could easily be regathered is also an unnecessary risk.

Notably, data for training the classifier are, according to Herping, a prime target for adversarial interferences. Adversaries can access the data collected outside the training environment and, depending on where the future training data is produced or stored, how it is secured and who has legitimate access to it, perform data-poisoning attacks. Unpoisoned data can also be extracted and exploited for malicious purposes or used for gaining insights into the model functioning. Finally, the integration of collected data from the outside world (e.g. from mobile apps and services or sensory data) also bear the risk of compromising the training environment if, for example, malicious code is injected into build systems (e.g. via Python libraries).[148]

Selection and adaptation of algorithms for specific-use cases require knowledge. Many off-the-shelf ML solutions can be deployed once basic questions about the nature of the problem are determined. Obtaining reliable results, however, requires a relatively deep mathematical understanding of the techniques in use, particularly as models become increasingly complex and intricate. In addition to domain knowledge about algorithms, a deep understanding of the

[147] Ibid.

[148] S. Herping (2019) op. cit.

cybersecurity domain and how attackers behave can greatly simplify and improve the performance of threat detection by algorithms.

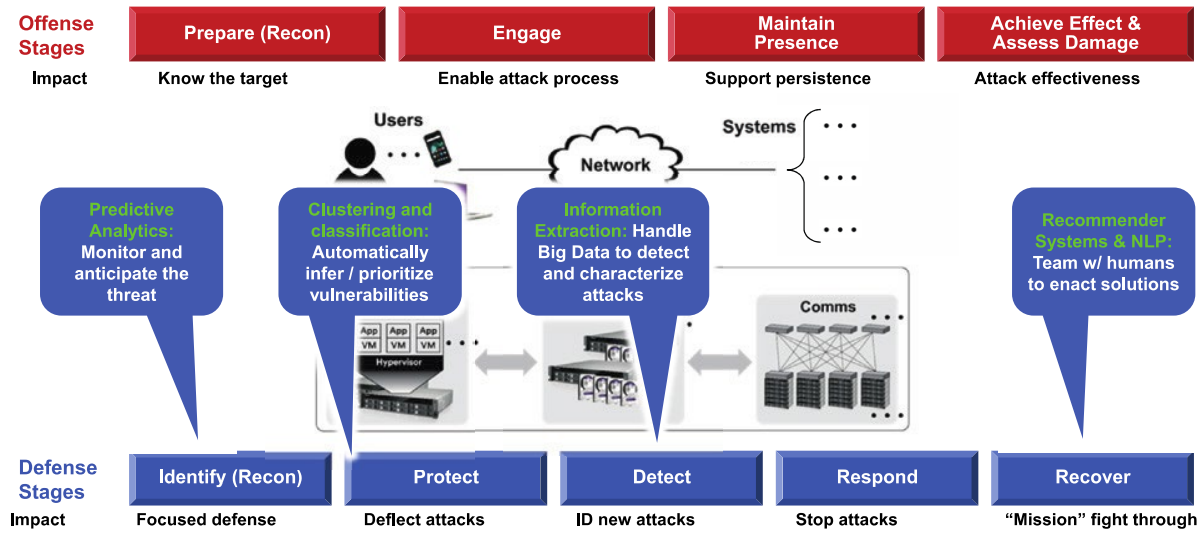
In conclusion, AI is a valuable tool that increases human productivity and gives broad access to the advantages of human knowledge and culture. But it is not a colleague, peer, or pet. It is not responsible or trustworthy: AI developers or users are, or deliberately are not. Intelligent systems are open to cybersecurity attacks at every stage of their development, deployment, and even in obsolescence if they retain memory. As such, they require good standards of engineering. Just as society has come to require the licensing of architects, the inspection of buildings, and applications for planning permission, so we may also need to move into a much more rigorous governance and regulation of software and its development process, whether or not it is conventionally considered intelligent or human-like.

8. Weaponisation and the offence versus defence debate

AI is and will increasingly be used to enhance attacks against information systems. Reliance on AI involves its own specific vulnerabilities and AI itself could also be used to seek offensive advantages. Yet whether this pattern in the cybersecurity paradigm will favour offence or defence requires further analysis.

Figure 6, from the MIT Lincoln Laboratory, presents applications of AI across the cyber kill chain.[149]

Figure 6. Application of AI across the cyber kill chain



Source: MIT Lincoln Laboratory, National Academies of Sciences, Engineering, and Medicine (2019), *Implications of artificial intelligence for cybersecurity: Proceedings of a workshop*, National Academies Press: Chicago, p. 33.

[149] National Academies of Sciences, Engineering, and Medicine (2019), Implications of artificial intelligence for cybersecurity: Proceedings of a workshop, National Academies Press: Chicago, p. 35.

This describes how AI in cybersecurity could have implications for both the offensive and the defensive positions. On the defensive side, as described earlier in the report, it points to many advantages across the multiple stages of identifying, preventing, responding to, and recovering from attacks. AI has also been used on the defensive side in support of logistics, military equipment maintenance, etc., as well as in more innovative sectors (e.g. the use of the IoT in military applications, or the use of AI to increase the effectiveness of cybersecurity for cloud computing applications, an area of great importance in today's trends for defence infrastructure).

For example, in 2019, the French Ministry of Defence published a report on the possible uses of AI support to the defence sector. One area of application is decision and planning support, where AI could help in filtering, exploiting, or sharing data and hence provide combats with informed choices to enable decisions to be taken more quickly and efficiently. Besides, AI could help in collaborative combating, "*whether [data mining] for the purposes of anticipation, immediate response or coordinated conduct of the action, or to the smart management of flows.*"[150]

AI could also offer support in logistics and operational readiness by, among other things, enhancing the efficiency of the supply chain or ameliorating the management of materials thanks to predictive management. Finally, as repeatedly mentioned, AI clearly offers support to cybersecurity. According to the French Ministry of Defence such AI applications include: the analysis of traces in a network to detect intrusion or malicious activity; the anticipation of threats based on available sources of information (open source); the measurement of system resistance levels; and the countering of digital influence.[151]

On the offensive side, Figure 6 describes several key stages: prepare, engage, maintain presence, and achieve effect and assess damage.

One prominent observation stemming from points raised earlier in this chapter is the tendency of AI systems to be used for malicious purposes even if there was benign intent in their design.[152] These types of technologies are generally referred to as dual use. Autonomous or unmanned aerial vehicles, for instance, could be reverted to endanger the physical security of individuals or infrastructures.[153] In the information security field, the use of AI for attacks is

---

[150] Ministère des Armées (2019), *Artificial Intelligence in Support of Defence: Report of the AI Task Force*, September.

[151] Ibid.

[152] See also E. Zouave et al. (2020), *Artificially intelligent cyberattacks*, Totalförsvarets forskningsinstitut FOI (Swedish Defence Research Agency), March (www.statsvet.uu.se/digitalAssets/769/c_769530-l_3-k_rapport-foi-vt20.pdf).

[153] M. Brundage et al. (2018), op. cit., p. 10 and p. 28. See also M. Cummings (2017), "Artificial Intelligence and the Future of Warfare", Research Paper, Chatham House, January: "*Another critical factor to consider in the debate over autonomous weapons is the increasing inability to disambiguate commercial drone autonomy from that of military UAVs. Indeed, with the rapidly expanding commercial market for both air and ground autonomous systems, there is evidence of some shifting in AI expertise from military to commercial enterprises. As a result, banning an autonomous technology for military use may not be practical given that derivative or superior technologies could well be available in the commercial sector*".

progressing at an increasingly fast pace, leading many experts to believe that "*the commoditization of AI is a reality now.*"[154]

Recent experiments showed how AI could be used by hackers to carry out spear-phishing attacks on social media such as Twitter.[155] During the 2017 Black Hat Conference, 62% of the surveyed attendees (mostly hackers and information security experts) "*believe artificial intelligence will be used for cyberattacks in the coming year.*"[156] On the other side of the spectrum, governments and institutional actors could make use of AI-enabled defensive systems for attacking adversaries.[157] For many, the lack of foreseeability of the risks that AI could bring is driving the belief that precautionary principles should be embraced to counter the widespread weaponisation of AI and prevent or limit unintended consequences for society.[158] Nonetheless, questions remain as to the effectiveness of such policy principles, particularly from those who consider the weaponisation of AI as something ongoing, inevitable,[159] and almost unstoppable. Regardless of the stance one may take on this, it seems vital that strategic reflection is required on how to most effectively regulate the adoption of AI systems. Choices will have to be made, particularly in sectors such as defence, law enforcement, and security, bearing in mind that "*constant attention will need to be given to the legal, ethical and strategic debates around human enhancement.*"[160]

This is particularly relevant given that scholars such as Slayton argue for challenging the common narrative that offence dominates cyberspace. According to Slayton, this narrative, by creating an undesirable belief in offence dominance, increases international tensions and makes states more ready to launch counter-offensive operations. No empirical evidence seems to justify the assumption that offence dominates, she asserts, and so rather than being a one-off, fixed assessment of cyberspace, the offence-defence balance should be understood as varying according to the specific cost-benefit of each operation. In this way, focusing on improving defence capabilities would allow states to acquire preparedness for cyber offence,

---

[154] J. Pandya (2019), "The Weaponization Of Artificial Intelligence", *Forbes*, 14 January (www.forbes.com/sites/cognitiveworld/2019/01/14/the-weaponization-of-artificial-intelligence/?sh=11f51dc43686).

[155] J. Seymour and P. Tully (2017), "Weaponizing Data Science for Social Engineering: Automated E2E Spear Phishing on Twitter", ZeroFox (www.blackhat.com/docs/us-16/materials/us-16-Seymour-Tully-Weaponizing-Data-Science-For-Social-Engineering-Automated-E2E-Spear-Phishing-On-Twitter.pdf).

[156] The Cylance Team (2017), "Black Hat Attendees See AI as Double-Edged Sword", BlackBerry (https://blogs.blackberry.com/en/2017/08/black-hat-attendees-see-ai-as-double-edged-sword).

[157] See for instance, G. Dvorsky (2017), "Hackers Have Already Started to Weaponize Artificial Intelligence", Gizmodo.

[158] D. Garcia (2018), "Lethal Artificial Intelligence and Change: The Future of International Peace and Security", *International Studies Review*, pp. 334-341. See also J.-M. Rickli (2020), "Containing Emerging Technologies' Impact on International Security," *Free World Forum*, Briefing no. 4 (https://frivarld.se/rapporter/containing-emerging-technologies-impact-on-international-security/).

[159] J. Pandya (2019), op. cit.

[160] J. Burton and S. R. Soare (2019), "Understanding the Strategic Implications of the Weaponization of Artificial Intelligence", 11th International Conference on Cyber Conflict: Silent Battle, CCDOE.

given the similarities between the two, *"without risking geopolitical instability or increasing vulnerability to attack."*[161]

Overall, when it comes to the offensive or defensive prominence of AI, the lack of reliable evidence and data means we cannot fully understand the impact of these systems on offensive capabilities vis-à-vis defensive ones, particularly when compared with one another. It remains unclear, for instance, whether using ML for detecting new cyber vulnerabilities will be more beneficial to a potential exploiter or for the system defender.[162] The quest for an answer has sparked an ongoing debate, with several eminent experts providing useful opinions on the matter while agreeing on their uncertainty because of the general unpredictability of the development of AI in cybersecurity.

In a 2018 essay published by the IEEE and on his personal blog, Bruce Schneier advances the hypothesis that AI could be of greater benefit harnessed for the defence of information systems than for attacks. His view derives from the argument that until now the human factor has kept cyber defence in a poorer position: "*present-day attacks pit the relative advantages of computers and humans against the relative weaknesses of computers and humans. Computers moving into what are traditionally human areas will rebalance that equation.*"[163] As such, improved cyber-defence techniques and the dual reinforcement of AI for cyber defence and cyber hygiene could somehow outweigh the growing availability and sophistication of AI attacks.

Based on the impact that AI has on robustness, resilience and response, Taddeo, McCutcheon and Floridi warned recently in *Nature* that the narrative should be seen from a multilevel perspective. From a tactical point of view, AI could enhance the protection of information systems, hence favouring defence over attacks.[164] However, from a strategic point of view, the situation turns in favour of the attacker, in that it would substantially alter the underlying dynamics of the game.

In terms of strategic stability, AI has the potential to escalate conflicts and to influence states' intentions to engage in conflicts. As underlined by Boulanin, given the difficulties in measuring the tangible evolution in military capabilities resulting from adopting AI systems, states might misperceive their opponents' capabilities. This might make them more inclined to trigger destabilising measures based only on the belief that their retaliatory capacity could be defeated

---

[161] See R. Slayton (2016), "What is the Cyber Offense-Defense Balance? Concepts, Causes and Assessment", *International Security*, Vol. 41, No. 3, pp. 72-109.

[162] M.C. Horowitz et al, (2018), op. cit, p. 4.

[163] B. Schneier (2018), "Artificial Intelligence and the Attack/Defence Balance", *IEEE Security & Privacy*, Vol. 16, No. 2, p. 96.

[164] "*For example, the use of AI to improve systems' robustness may have a knock-on effect and decrease the impact of zero-day 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.*", M. Taddeo, T. McCutcheon and L. Floridi (2019), op. cit., pp. 557–560. See also M. Taddeo and L. Floridi (2018), "Regulate artificial intelligence to avert cyber arms race", *Nature*, pp. 296-298.

by another state's AI capabilities.[165] Similarly destructive effect on the strategic stability might also be incurred in a case in which the opponents' ability is underestimated, especially in the case of poorly conceptualised, accident-prone autonomous systems.[166] Hence, Johnson argues, *"in today's multipolar geopolitical order, relatively low-risk and low-cost AI-augmented capability – with ambiguous rules of engagement and the absence of a robust normative and legal framework – will become an increasingly enticing asymmetric option to erode an advanced military's deterrence."*[167]

[165] V. Boulanin (2019), "The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk", Sipri, May.

[166] J. S. Johnson (2020), "Artificial Intelligence: A Threat to Strategic Stability", Air University Maxwell AFB, United States.

[167] Ibid., p. 29.

PART III.
CYBERSECURITY FOR ARTIFICIAL INTELLIGENCE

1. Introduction[168]

As mentioned, AI algorithms could be said to be of two types, symbolic and non-symbolic. In symbolic AI algorithms, knowledge, or if you prefer, intelligence, is coded as a normal computer programme, for example if <X divided by two is an integer> then <display "X is even"> else <display "X is odd">. One could say that this algorithm equates the knowledge of a seven-year old child.

Non-symbolic AI algorithms, of which ML and its derivatives are the best-known examples, create a network of nodes and weighted links between the nodes, which will be the *model* to be learned. During the training phase, the algorithm computes the weights of the links, based on a chosen training dataset. In our example, it could use a dataset representing two different categories, even and odd numbers, to train a model where each element of the training dataset is classified in one of two categories, A or B, according to their parity. Once the weights are such that the programme output matches the programmer requirements (e.g. when 97% of the test input is correctly classified), then the training phase is stopped and the weights are fixed in the programme, which is then ready to be deployed. The corresponding algorithm could be seen as if <X divided by 2 falls in category A after crossing the weighted network> then <display "X is even"> else <display "X is odd">. But since a training phase is now used, another, simpler way of having programmed it could be if <X falls in category A after crossing the weighted network> then <display "X is even"> else <display "X is odd">.

In this non-symbolic AI algorithm, the network and its weights – or the features model – defined at the end of the training phase will stop evolving once the programme is deployed, unless the programmer decides that the training phase could continue even after deployment. In this case, the weights would continuously evolve at runtime.

Thus, in the scope of this chapter, it is useful to separate AI systems into three categories.

- Symbolic AI systems that are never trained and do not change after being programmed but produce predefined outcomes that are based on a set of certain rules coded by humans.
- Non-symbolic AI systems that are trained only before deployment. These systems change their internal weights prior to deployment, but do not change at runtime. Let's call them static ML systems.
- Non-symbolic AI systems that continue to be trained after deployment and are said to 'evolve' at runtime because their internal statistical parameters may be changed according to new data. Let's call them evolving ML systems.

To ensure that the potentially large impact of AI systems that are not secure is well understood, we propose the following definition of attacks to AI systems: *An attack to an AI system is defined*

[168] This chapter includes valuable contributions from several Task Force members, in particular Professor J. Bryson, Calypso, Guardtime, and Wavestone. Other contributions are acknowledged elsewhere in the text.

as its purposeful manipulation to cause it to malfunction. And we note that the main point here is that *such attacks need not be conducted via cyberspace*, as will be explained below.

…all attacks that can be carried out on rules-based AI systems can also be carried out on static or evolving ML systems…Moreover, because the decisions taken by any *AI systems are increasingly based on inputs that are sensed from the environment …security concerns must also be addressed at usage level, even in the case of a perfectly correct and uncompromised system.*

It could thus be considered that evolving ML systems encompass the two others and that static ML systems encompass rules-based AI systems, in the sense that all attacks that can be carried out on rules-based AI systems can also be carried out on static or evolving ML systems. The attentive reader would have noticed that the inverse is not true, since symbolic AI systems do not have, for instance, a training phase. Moreover, because the decisions taken by *any* AI systems are increasingly based on inputs that are sensed from the environment – either obtained from online data sources or from the physical world in, for example the case of IoT – security concerns must also be addressed at usage level, even in the case of a perfectly correct and uncompromised system.

To illustrate a larger attack surface, take an example based on the Covid-19 pandemic. Contact tracing apps were being developed to warn users when they had been in close contact with persons that later turn out to test positive, because they would be at risk of having been infected. In the development of such a rules-based AI system, some intelligence was encoded in the apps to define the notion of 'close contact', e.g., less than five metres phone-to-phone, as measured over Bluetooth.

Then, the following hypothetic attack scenario was described in a paper written by French cybersecurity experts: "*Soccer player Ronought is set to play the next Champions League match. To prevent him from playing, it is enough for an opponent to leave his phone next to Ronought's without his knowledge, then to declare himself sick. Ronought will receive an alert because he is said to have been in contact with a person infected and will have to stay away from the fields in quarantine.*"[169]

This example shows that an AI system may be functioning perfectly from the information and communications technology (ICT) point of view, but at the same time it can be used as an attack because its decision-making capabilities can be tricked at usage time, stopping the system as a whole from functioning as intended.

The main message of this chapter is that AI systems are IT systems (software running on hardware) that nonetheless have specific internal and usage features. Accordingly, they must be cyber-protected as with any other ICT system, but additional protection must be designed for their special features, namely the training phase, the interaction with the environment (in

[169] X. Bonnetain et al. (2020), "Le traçage anonyme, dangereux oxymore: Analyse de risques à destination des non-spécialistes", 27 April (www.leclubdesjuristes.com/blog-du-coronavirus/que-dit-le-droit/le-tracage-anonyme-dangereux-oxymore-analyse-de-risques-a-destination-des-non-specialistes/).

particular in the case of Cyber-Physical Systems (CPS), because of safety issues), and the possibility of runtime systems' evolution. Therefore, AI systems must follow secure development life cycles, from ideation to deployment, including runtime monitoring and auditing. Ideally, this should be coupled, during the commercialisation phase of AI systems, with the proper conformity assessment and market surveillance mechanism to ensure AI security when the systems are placed in the market and during the whole life cycle of the products. While analysing the appropriate requirements for AI commercialisation is outside the scope of this report, it should nonetheless be acknowledged that envisioning appropriate provisions for such a phase should be regarded as essential as ensuring AI secure-development life cycle.

2. Machine learning systems do indeed have a larger attack surface

Since all AI systems are composed of software running on hardware, traditional cyberattacks in AI systems can use a traditional attack surface caused by software/hardware bugs, which usually stem from human mistakes made while writing the code or designing the hardware.

Since all AI systems are composed of software running on hardware, traditional cyberattacks in AI systems can use a traditional attack surface caused by software/hardware bugs, which usually stem from human mistakes made while writing the code or designing the hardware. In this scenario, adversaries will find those vulnerabilities and find ways to exploit them and get access to the AI system under attack. This is the usual scenario in cybersecurity. However, as previously mentioned, some AI systems, namely ML systems, present specific internal and usage features that can be attacked in manners that are different from traditional cyberattacks, raising new cybersecurity questions.

Some of the specific features include:

- the training or retraining phases,[170] which include a training dataset and a resulting feature model
- the interaction with the external environment, including sensing capabilities that will guide internal decisions and actuation of physical systems
- the ability, in some cases, to evolve during runtime.

Therefore, attacks on ML systems may leverage more than just software vulnerabilities. In particular, the training dataset – be it before or after deployment – may be compromised so that the resulting 'learning' of the system is not as intended. External objects sensed by the system can also be tampered with so that they are not recognisable as shown in the training dataset – a well-known example being Stop signs very slightly modified with some duct tape. Such attacks cannot be mitigated by software patches, because they belong either in the training dataset that has already been used to define the current 'behaviour' of the system, or

[170] The retraining environment is more difficult to hack because ideally it is more controlled. Attackers gaining access to the retraining environment can nonetheless still poison the data, extract intelligence, etc., with the difference being that in the retraining environment data are already labelled, thus possibly increasing their quality and value.

in external objects that have been tampered with or are tied closely with the functioning of the system, for example in the case of backdoors coded into the model.

Yet another kind of attack can be used to render the ML system unavailable. Because such systems are used in CPS,[171] the physical part of the system may be made to malfunction and hamper the availability of its cyber part. An example would be a smart ambulance that is made unavailable by puncturing its tyres. A consequence is that ML systems fundamentally expand the set of entities that could be used to execute attacks on the cyber system to include the training kit, but also physical objects.

The environment should now be considered as part of the attack surface in cybersecurity terms.

Data can be weaponised in new ways and objects and actors external to ICT systems can no longer be treated as something separate from the system. The environment should now be considered as part of the attack surface in cybersecurity terms. This is certainly an unexpected twist in the convergence between the physical and cyber realms, and one that makes it very difficult to secure AI systems that are deployed in CPS.

3. A high-level view of the threat landscape[172]

Attacks on ML systems can have very serious consequences when they are integrated into critical applications. The effects of AI attacks have been analysed earlier in this report, namely the expansion of existing threats, the introduction of new threats and the alteration of the typical characteristics of threats. As stated, an attack to a ML system is its purposeful manipulation with the end goal of causing it to malfunction.

AI attacks can also take forms that strike at different weaknesses in the underlying algorithms, or at different inputs sensed from the environment.

3.1 Input attacks

In an input attack there is a system behaving as it is expected and that would work properly under normal circumstances. The attackers target the input that feeds the ML system, for example the data that are acquired by the system. In the regular use of ML, the system takes input from the outside world and processes it. During an input attack, however, an attack pattern is added to the input such as tape on a Stop sign or a small change in the pixels of a photo uploaded to a social network. Based on how the pattern is chosen, the attack pattern will change the way the ML system processes the data and will eventually cause the ML system to fail.

[171] Cyber-Physical Systems (CPS) are integrations of computation, networking, and physical processes. Embedded computers and networks monitor and control the physical processes, with feedback loops between these three components. Examples of CPS include smart grids, autonomous automobile systems, medical monitoring, industrial control systems, robotics systems, and automatic pilot avionics.

[172] This section of the report draws from M. Comiter (2019), op. cit.

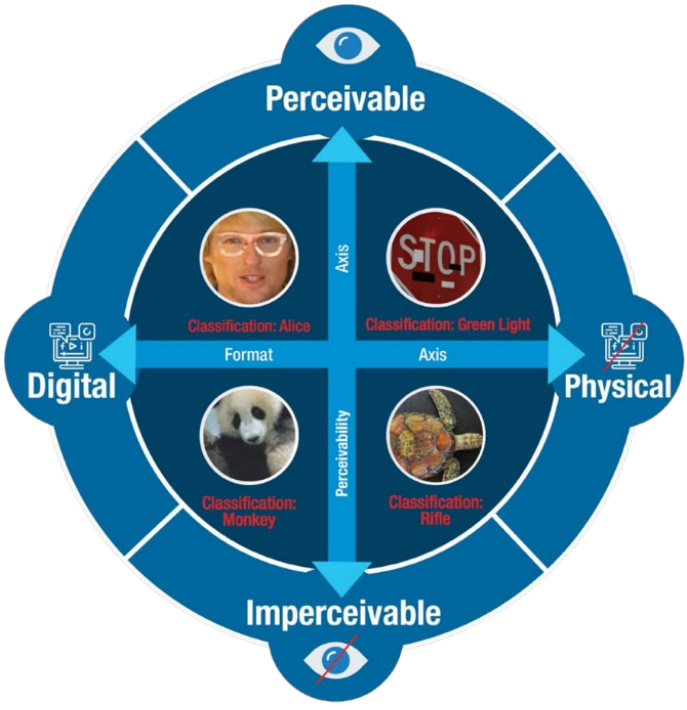
To perform an input attack, the attacker does not need to completely transform the input. In fact, an imperceptible change to the input, invisible to the human eye, can also cause the system to malfunction. In this respect, it has been noted that *"placing a five-centimetre piece of white tape on the upper corner of a stop sign would exploit a particular brittleness in the patterns learned by the model, turning it into a green light."*[173] Or, in the audio domain for example, attackers can introduce high-pitched sounds, imperceptible to human ears but able to be picked by microphones, to fool audio-based AI systems such as digital assistants.[174]

Those attacks can be crafted relatively easily by using optimisation methods whenever the attacker has access to the model. Notably, having access to the model is also fairly easy, because it is the model often distributed publicly as open source by the companies. Besides, even when attackers do not have access to the model used, input attacks can also be crafted by leveraging access to the output or the dataset.

The myriad variations of input attacks can be mostly characterised along two axes, depending on whether they are noticeable by humans (e.g., for attacks on physical entities, whether the attack is visible or invisible to the human eye) and what support is used for the attack vector (whether it is a physical object, like a Stop sign, or a digital construct, like an image file on a computer).

Figure 7 presents these two axes. The horizontal axis characterises the format of the attack and the vertical axis characterises the perceivability of the attack.

Figure 7. Input attacks



Source: M. Comiter (2019), op.cit., p. 19.

[173] M. Comiter (2019), op.cit, p. 18.

[174] Ibid., p. 22.

A different but similar classification of input attack is proposed by Gilmer et al.[175] According to these authors, input attack can be classified as:

1. Indistinguishable perturbation, in which any changes go completely undetectably by humans.
2. Content-preserving perturbation, in which attackers make changes in the data distribution while not altering the content.
3. Non-suspicious input, in which attackers can change the input as long as it does not look suspicious to a human.
4. Content-constrained input, in which attackers can change the input as long as it carries specific payload.
5. Unconstrained input, in which attackers can produce any input they want in order to induce the desired behaviour from the system.

3.2 Poisoning attacks

Unlike input attacks, the aim of poisoning attacks is to impede a proper ML system from being built. In poisoning attacks, the attacker seeks to damage the ML model itself by targeting the training phase so that once it is deployed, it is inherently flawed. Similarly, backdoors attacks target and jeopardise the model algorithm itself. Model poisoning attacks and backdoors attacks take place while the model is being defined, fundamentally compromising the ML system.

To poison the ML system, the attacker compromises its training and learning process so it can perform the tasks that are requested by attacker, such as failing on certain attacker-chosen input. An example could be an ML system trained to detect enemy aircraft poisoned in such a way to make certain aircraft no longer recognisable.

Data are major avenues through which a poisoning attack can be crafted, even if not the only ones. Most AI algorithms are powered by ML systems relying on data and extracting patterns from the dataset. Because information in the dataset is distilled into the ML system, any problems in the dataset will be inherited by the model. In this context, the attacker can either switch valid data with poisoned ones or they can attack the process through which data are acquired itself. In the latter, rather than changing an otherwise valid dataset, the attacker manipulates what is represented in the data in the first place.

…in an AI-dominated society data are not only a powerful resource but also a major source of vulnerabilities…

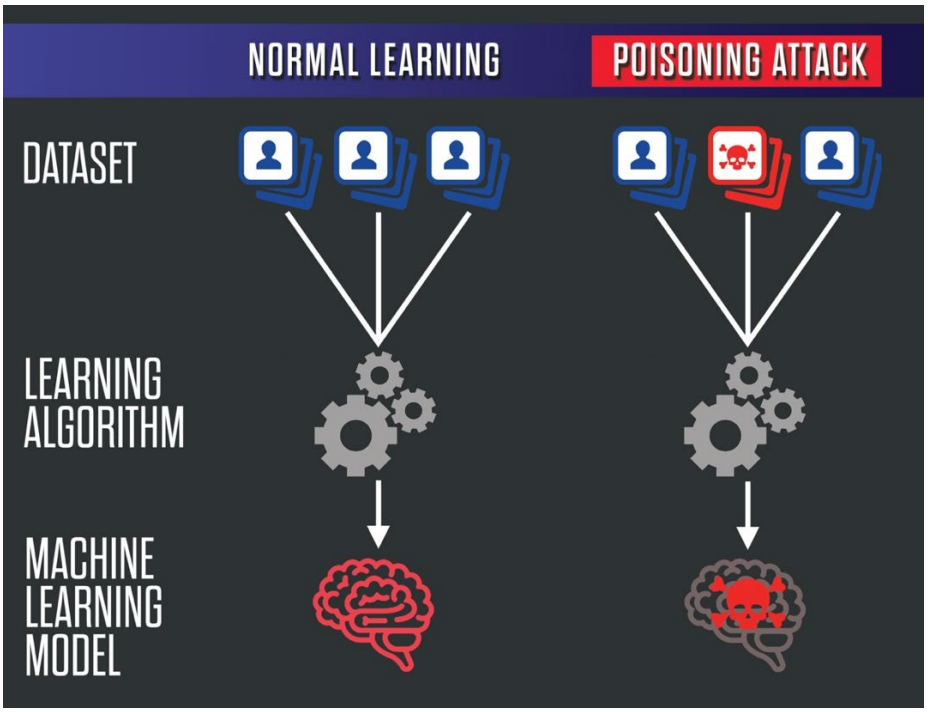
According to Comiter, unveiling such aspects is of pivotal importance for changing the societal perspective on data and reversing the conception of them as *'digital gold'*. In fact, in an AI-dominated society data are not only a powerful resource but also a major source of vulnerabilities,

[175] J. Gilmer, R. P. Adams, I. Goodfellow, D. Andersen and G. E. Dahl (2018), "Motivating the Rules of the Game for Adversarial Example Research", arXiv preprint arXiv:1807.06732.

in the way that Rome's powerful roads were turned against them by their enemies.

Figure 8 shows normal machine learning extracting pattern from the dataset, and a poisoning attack where the training data are poisoned to change the learned model.

Figure 8. Poisoning attacks



Source: M. Comiter (2019), op.cit., p. 29.

Another way to understand the threat landscape is to use a threat model, as explained in the following section.

4. An AI threat model[176]

Threat modelling is a structured approach that helps to identify possible threats to ICT systems. In the case of ML systems, the first thing to consider is the learning mode chosen by the system, because the way in which the control system is built differs according to the learning mode used by the ML system.

The AI attack surface must then be isolated. It is useful to think in minimal terms of the different layers here, with the intention of compiling the potential threats for each of those layers that need to be contained and controlled. The main layers are as follows.

[176] This section is based on Martin Dion's contribution to the third meeting of the CEPS Task Force.

Infrastructure layer

As more and more AI algorithms are implemented in silicon, specialised hardware takes an ever-increasing role in AI systems. Notwithstanding this, AI hardware security is still broadly overlooked.

Data layer

As discussed above, data is fundamental in the learning phase of ML systems. Data needs to be relevant, unbiased, and uncorrupted. In particular, it is crucial to devise the best way to encode the features of a question into the data that will be fed into learning modules of ML systems.

Model layer

As presented above, ML algorithms have a features model. According to the EU SHERPA project,[177] attacks against machine-learning features models can be divided into four main categories, based on the motive of the attacker.[178]

- Confidentiality attacks expose the data that was used to train the model. Confidentiality attacks can be used to determine whether a particular input was used during the training of the model.
- Integrity attacks cause a model to behave incorrectly because of tampering with the training data. These attacks include model skewing (subtly retraining an online model to recategorise input data), and supply chain attacks (tampering with training data while a model is being trained offline).
- Availability attacks refer to situations where the availability of a machine-learning model to output a correct verdict is compromised. Availability attacks work by subtly modifying an input such that, to a human, the input seems unchanged, but to the model, it looks completely different (and thus the model outputs an incorrect verdict). Availability attacks can be used to 'disguise' an input to evade proper classification. From the point of view of the attacker, availability attacks are similar to integrity ones, but the techniques are different: poisoning the model versus crafting the inputs.
- Replication attacks allow an adversary to copy or reverse-engineer a model. One common motivation for replication attacks is to create copy (or substitute) models that can then be used to craft attacks against the original system, or to steal intellectual property.

Algorithmic layer

In turn, the construction of the features model depends on many choices related to the design of the algorithm, such as the depth of the network or the learning rate. Protecting the integrity

[177] See "Shaping the Ethical Dimensions of Smart Information Systems (SHERPA) A European Perspective" (www.project-sherpa.eu/).

[178] A. Patel et al. (2020), D1.3 Cyberthreats and countermeasures, SHERPA, April (https://doi.org/10.21253/DMU.7951292).

of the features model includes safeguarding the choices in design as well as the resulting weights composing the model. For example, it has been demonstrated that, following the Differential Power Analysis methodology, attackers could discover the value of the secret weight of a neural network.[179] However, it is important to note that several attacks on ML systems may be averted by better algorithms. One example stems from the automation of image processing and the Stop sign attack seen above, which is possible because the attacked algorithms are optimised to rationalise the dataset required to produce meaningful outputs. However, if the algorithm is optimised to provide broader analytics of the underlying data instead, the noise that is injected over an input image may be detected as such.

Operational layer

Once the AI system is deployed, perhaps even autonomously, its environment plays an important part in its performance. For instance, it may be the case that input, upon which internal algorithmic decisions will be taken, is not properly protected or filtered, resulting in unexpected decisions by the system. If the AI system is embedded in an autonomous CPS, for example a smart car, tampering with its physical components can also modify the expected outputs. Another fundamental aspect that must be kept in mind is that AI systems operate in a time/space continuum that is very different from that of humans, since billions of operations per second are operated by a computer. Connected to the Internet, they know no geographical constraints. Therefore, it is more difficult to detect AI system failures in time to prevent malfunction.

4.1 Role of human operators

We note that most elements in this threat model are AI-specific and only a few, like attack on infrastructure, are generic and apply to every technological solution. Furthermore, the threat model of AI systems is very complex, and it is not something one single actor can tackle, be they governments or software vendors. A collective effort would be needed to make AI systems secure, where the technical expertise of human operators would match the complexity of the deployed systems.

If real-time monitoring objectives must be met, the following approach comes to mind.

- Another system is deployed to control/monitor the decisions made by the primary system. It would either trigger the fallback automatically or request the involvement of a human operator (human on the loop). The monitoring system itself will, by definition, have to exhibit an appropriate degree of sophistication and therefore complexity; otherwise, a simple decision-making system would have been deployed in the first place. The involvement of a second complex system raises the question of the detection of issues in the monitoring process, as the monitoring system could itself be biased or flawed.

[179] A. Dubey, R. Cammarota and A. Aysu (2020), op. cit.

If the real-time issue of detection constraint is lifted, the following approaches come to mind.

- A monitoring system that processes the logs and validates input/outcome expectations could be deployed. Most probably, the amount of data and detection of off patterns will require machine or deep learning techniques. It would either trigger the fallback automatically or request the involvement of a human operator (human on the loop). In all cases, the fallback plan would have to account for the time delay between the issue detection and the continued operation of the primary system.
- One or multiple human operators are kept in the loop at all times. Whether or not the human operator is assisted by a monitoring system, the primary system must operate at a pace that is commensurate with human operator checks/interventions. Here too, the fallback plan would have to account for the time delay between the issue detection and the continued operation of the primary system.

5. Safety and security of open, autonomous, AI-based IT infrastructure, and its runtime evolution[180]

In the emergence of AI-based systems there is one aspect that deserves extra attention: AI systems that interact with the physical world, or CPS. These systems are unique in that their agency is not limited to digital information – they have implications in a very concrete sense through, for instance, autonomous vehicles, smart homes, industrial robots, and even autonomous weapons systems. While the overall security and safety of any AI-based system is important, such systems do require some additional consideration because of their direct connection with truly unstructured and unpredictable real-world information, and their possibility to impact people's safety and wellbeing.

As with any critical system…we must consider the fundamental security aspects of those systems, and how they might be utilised for malicious purposes…Security must not be an afterthought: it should be a fundamental requirement for building such systems. And these considerations are not unique to AI-based systems.

As with any critical system, before even considering the additional implications that AI might bring, we must consider the fundamental security aspects of those systems, and how they might be utilised for malicious purposes. Starting from the hardware, interfaces, and software, all the layers of the solution should be designed for secure operation from the bottom up. As mentioned, AI systems are designed, and models are selected as the data, parameters and tests used to determine when training and parameter-setting are completed. All those decisions are auditable and best and poor practice can be established and checked at every stage. Security must not be an afterthought: it should be a fundamental requirement for building such systems. And these considerations are not unique to AI-based systems.

[180] This section was provided by F-Secure.

One concern that is especially relevant for CPS is the challenge of input data space. One core benefit of the human mind, over any current AI solution, is the ability to generalise. AI solutions are far behind in this ability. While people naturally generalise to new circumstances, this is often a very significant challenge for AI systems. For example, people know that driving a car in the dark impacts their visibility but does not change the fundamental physics of the situation. If driving is done on a forest road, the fact that a deer can run into the road might be anticipated. As such, the driver will look for movements at the side of the road, even if they have never actually had a close encounter with a deer. But current AI systems will not take special action unless they have been explicitly trained to do so. Although a simplistic example, this illustrates on the one hand the risks associated with anthropomorphising AI systems and, on the other hand and most importantly, the difficulty of factoring in something that has never been present in the data used to train an AI system.

Another similar example is the presence of fog or the alterations to Stop signs. What has not been accounted for in training can result in severe difficulties for AI systems. This basically stems from the simple fact that most AI systems are still effectively just very complicated curve fitting, which aim to find a decision boundary in a vast high-dimensional data space where the separation is often minimal. It is crucial to understand that examples like a panda being recognised as a gibbon in image recognition are real concerns that must be accounted for – not something that cannot be overcome, but something that must be considered.

Another key issue arises from the difficulty of assessing the correctness of actions taken by AI systems in real life. 'Doing the right thing' is often hard for humans, but much more so for machines. In a widely changing world, one of the biggest challenges is how to specify a high-level goal in a way that is understandable to an AI system. If it's true that, for example, reinforcement learning (via a delayed gratification learning approach of needing to specify the end outcome rather than a reaction on every input point) has provided great progress with respect to a simple supervised approach (where the system is trained with data that is tagged with predetermined categories) or an unsupervised approach (where the trained system itself creates categorises underlying similarities in the training data), this is still a very difficult problem to address for more complicated tasks.

There is even more complexity in situations where the AI system actually 'learns' online, namely evolving AI systems. One important issue is that the stability of training such systems is not something to be taken for granted. Combining this with the robustness (or lack thereof) of underlying models, the risks of decision boundaries moving, variations in the input data and models that change their behaviour over time based on the data they receive, can be both very powerful and potentially quite risky.

One, luckily quite benign, example of this is Microsoft's unfortunate Tay chatbot. Tay was an AI chatterbot that was originally released via Twitter in 2016, then immediately attacked online and taught how to post inflammatory and offensive tweets through its Twitter account. It was consequently shut down less than a day after its launch.

This is still just an interesting anecdote, but with enough knowledge of the AI system's underlying model and systematic manipulation of its decision surfaces, it is not impossible to

foresee autonomous cars changing their behaviour to have much less regard for safety than originally planned. Such a scenario is even more worrisome when considering the future potential of CPS applications. According to Cassandras, Smart Cities themselves should be understood as CPS, cyber-physical infrastructure combined with software for sensing, communicating, decision-making, and actuating – rather than simply collecting and sharing data.[181] As such, CPS are undoubtedly going to increasingly characterise our daily life.

The European Parliamentary Research Service published a Scientific Foresight study as long ago as 2016 to try to gain an understanding of what the impact of these technologies would be by 2050. This raised several ethical issues related to the adoption of CPS, including, for example, the fact that it *"will allow the gathering of large amounts of personal data on individuals for security surveillance purposes,"* or that *"CPS will result in far more information about consumers being collected, potentially intruding upon the privacy of individuals and communities."*[182] There is thus even more potential for undue appropriation or misuse of the collected data and manipulation of the underlying models, by either the owner of the data, or by attackers.

Given the current technological advances, the illustrated risks related to these systems must be very carefully considered before deploying an evolving cyber-physical ML system onto the physical world. And when such systems are really needed, for example for tasks that humans cannot do, extra care should be taken and safeguards applied.

Since all information-based systems rely on their input, many attacks against IT systems utilise unexpected or poorly handled inputs. While it is paramount for any deployed IT system to be resilient against data changes, there is usually a finite number of combinations (even if sometimes very large) that must be considered. Yet in the context of evolving ML systems, the number of such combinations explode, as the same inputs can cause a different output after real-time learning. So, to be in control of the ML system's reactions, one would need to test all possible combinations of all possible inputs where the order of the inputs does matter – which is just not feasible.

Similarly, attacks can also be carried out against AI systems when these are entirely deployed in the environment. Attackers can target the intersection with the outside world itself, for example by triggering the brakes of a self-driving car. Additionally, attackers can manipulate the data that are sent back to the system once deployed. Finally, attackers can manage to interfere with the model to derive the training data by using the knowledge of the model output using statistical hypothesis, for example.[183]

Hence, it must be admitted that an evolving ML system cannot be fully tested in advance, and an approach of constant monitoring of the system, its inputs, outputs, the model, and the overall reliability and robustness of the system, should be adopted instead. Of particular

[181] C. G. Cassandras (2016), "Smart cities as cyber-physical social systems", *Engineering*, Vol. 2, No. 2, pp. 156-158.

[182] European Parliamentary Research Service Scientific Foresight Unit (STOA) (2016), Ethical Aspects of Cyber-Physical Systems: Scientific Foresight Study, June (www.europarl.europa.eu/RegData/etudes/STUD/2016/563501/EPRS_STU%282016%29563501_EN.pdf).

[183] S. Herping (2019), op. cit.

importance would be the careful monitoring of all parameters of the system, with the following three main control points.

Understanding how well the input data maps to the known space of inputs. Basically, observing and understanding if the input data being served to the ML system at runtime, which are used by the system to learn and evolve, are a) within the range of values the system has previously been trained on and b) have been verified as consistent.

Even the most complicated ML models are not black boxes…

Understanding the internal workings of the model. Even the most complicated ML models are not black boxes – they are just very complicated statistical functions, with numerous parameters and connections.[184] As long as the ML system is running on controlled environments, its parameters can be observed if desired. And while these parameters may not be intuitively interpretable, changes in them can be followed, given the inputs and outputs of the system, and their inherent stability quantified, in order to observe and react to the emergence of potential risks.

Understanding the outcomes of the system. Monitoring the outputs of a system is at least as important as monitoring its inputs and internals, since the outputs are what matters, as it's these that impact the world around CPS. While this task is also far from trivial, one simply cannot afford not to put every effort possible into tracing the decisions made by a model, comparing them to the initial state (training before deployment and runtime learning), and understanding possible bias in its decisions (compared to what was to be expected).

To make these approaches actionable there must be something that can be done if there is reason for concern. For evolving ML CPS, the approach will largely depend on its functionalities. In some cases, a suitable outcome could be to return control to the human. In other situations, a rollback to recent updates or perhaps reverting to the initial state may be enough if malfunctions or unwanted behaviours are discovered.

It must be clear that if an evolving ML CPS is not carefully monitored during its interactions with the physical world, it is the cyber-physical equivalent of driving on the highway with your eyes closed. If things go wrong, there will be no time to implement corrective actions, since nobody was paying attention to what was happening in the first place. Risk-based approaches that consider the whole life cycle of AI systems, possibly guaranteeing the system operates at a pace that is commensurate with human operator checks/interventions, will be the best strategies if the benefits of such systems are to be reaped.

[184] New techniques such as LIME are also being developed. This helps explain predictions of the ML classifier and has been shown to increase human trust and understanding. Overall, through LIME, it is possible to learn a model's local behaviour by varying the input and seeing how the outputs (predictions) change. For more on this see M. T. Ribeiro, S. Singh and C. Guestrin (2016), "Why Should I Trust You? Explaining the Predictions of Any Classifier", *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining,* August, pp. 1135-1144.

6. Addressing the insecurity of the network as it relates to AI[185]

Data used for machine learning, like any other sort of data, is susceptible to theft. As such, common concerns about cybersecurity couple with those unique to the use of AI. Unauthorised manipulation and corruption of the data can lead to invalid outcomes from analysis, but this risk has always been present with any data processing. One way of mitigating this risk is to guarantee data integrity by ensuring the absence of unauthorised changes in the data, which can one day become generalised procedure, possibly based on blockchain technology. Note, however, that integrity does not mean correctness, as discussed below.

The concerns about cybersecurity become most relevant when the sources of the data that drives the analysis are highly distributed, and the analysis itself is highly distributed. A highly distributed system is at risk from disruptions in network availability and performance, especially if the goal is to use the outputs of the distributed system in a time-critical context.

Another risk from a highly distributed system arises when the identity and legitimacy of the data sources cannot be properly validated. In any system that has many participants, some of the participants may be malicious or incompetent, and the design of the system must somehow take this reality into account.

Control of autonomous vehicles is an example that illustrates many of these concerns. Today, each autonomous vehicle functions mostly independently and makes decisions based only on the data directly available to that vehicle. However, if other, indirect, sources of data could be incorporated into the algorithms that the vehicle uses for its decision-making, it would seem beneficial. A convoy of vehicles could communicate to synchronise their actions. A vehicle detecting an accident or other anomalous event could tell vehicles following it, which might allow a better response to the situation. There are protocols being designed today, such as Vehicle-to-Vehicle (V2V) that permit nearby vehicles to communicate.[186]

The question is then how to ensure that the messages from other vehicles are legitimate. For example, if a vehicle could forge a message, it could pretend to be an emergency vehicle that calls for priority access to the road, turning to green all traffic lights in its way. An actor intent on simply causing mischief (or a serious accident) could send a message saying that a road segment is clear when there is actually constriction. The risks become greater once vehicles interact with the infrastructure – traffic lights, traffic sensors, speed limit notifications and so on. Conversely, how does a vehicle know that a traffic light with which it is interacting is a legitimate traffic light?

Note that this is not the same as addressing confidentiality and integrity of data through encryption, because this does not resolve identification, and identity tools and key management infrastructure would be needed. For example, a user connecting to a website is given a certificate issued by a certificate authority, which guarantees its identity. If this is not

[185] This section was contributed by David Clark, Senior Research Scientist, MIT Computer Science & Artificial Intelligence Laboratory.

[186] See NHTSA, Vehicle-to-Vehicle Communication (www.nhtsa.gov/technology-innovation/vehicle-vehicle-communication).

implemented in the infrastructure, encrypted conversations may still happen, but with the wrong interlocutor. The bottom line is that encrypted lies are still lies.

There is research underway to understand and deal with these sorts of issues. But the design space is thorny. Should every vehicle and every element of the infrastructure have some sort of unique identity that can be verified? How would an identity system of that scale be built and made resistant to abuse? How could the resulting privacy issues be managed? If the system does not depend on robust identity to discipline bad behaviour, is there some way that technology can prevent the forging of messages?

Vehicle control is a highly time-critical system. Communication could be disrupted by jamming and other denial of service attacks. Somehow the system must protect itself from malicious messages. All of this must happen in real time. And while some of the communication may be local (as with the V2V scheme) some of it might happen through the cloud. Today we see cloud-based route-planning services that depend on real-time input from vehicles on the road. Route-planning services seem essential for autonomous vehicles – somehow the vehicle must be given (or compute) a route when given a destination. Collaborative route planning can conceivably have benefits in overall traffic management, but real-time sensing of traffic loads has already been disrupted by injection of false information – a man pulling a wagon with 99 cell phones down the street at walking speed tricked Google maps into thinking there was a traffic jam around which it should divert traffic.[187] Notably, while this holds true for the automotive sector, maintenance and monitoring of IoT systems, with wired and wireless real-time communication, can be regarded as an even bigger issue that will similarly need to be addressed in light of their massive commercialisation.

The high-level cybersecurity challenge for these sorts of distributed planning and control algorithms is to address the reality of malicious and manipulative input of data from some of the participants in the full CPS. More generally, it is still not clear exactly what features are required from communication networks and infrastructures for autonomous AI systems to be deployed in highly distributed settings, which are composed of many such autonomous and interacting AI systems. For instance, on the Internet, the issues of identity and key management had deliberately not been embedded in the network layer itself, but left to the discretion of application developers. This was a correct idea as different uses of the Internet have different needs for identification features.

7. An example of a secure development life cycle for AI systems[188]

From a software engineering point of view, although the debugging of ML systems is still a developing area, AI systems can ideally be patched like any other IT system, including stopping

[187] B. Barrett (2020), "An artist used 99 phones to fake a google maps traffic jam", *Wired*, 2 March (www.wired.com/story/99-phones-fake-google-maps-traffic-jam/).

[188] This section is based on the contribution by Martin Dion to the third meeting of the Task Force and also by Joanna Bryson, Professor of Ethics and Technology at the Hertie School of Governance, Berlin and member of the Advisory Board of the Task Force.

the runtime learning if needed. In a nutshell, problems with an AI system could happen at any of its main life-cycle stages.

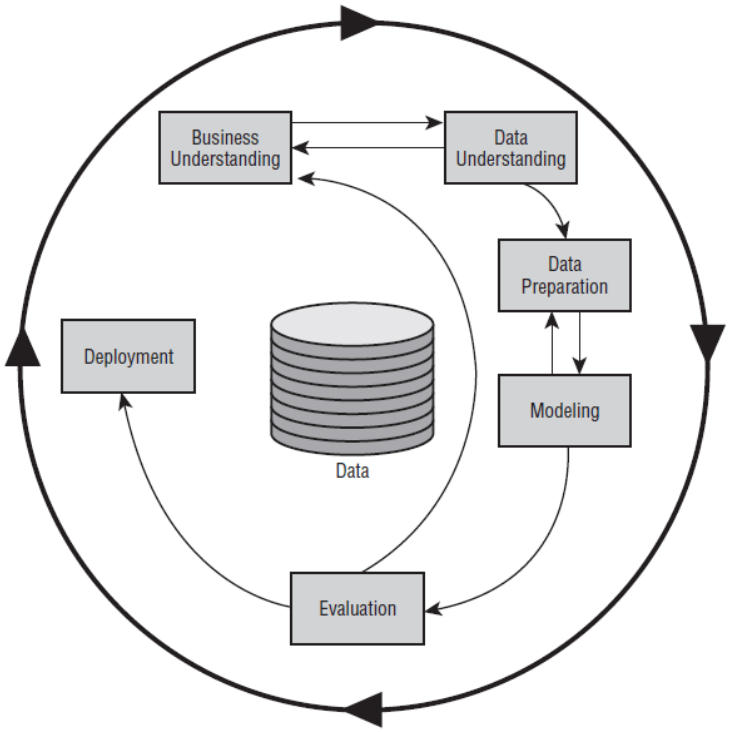
For the scope of this analysis, it is useful to frame the life cycle of AI systems on certain proven steps that has helped the deployment of other technologies in the past (see Figure 9).

Figure 9. AI systems life cycle

Source: authors' composition.

Because of the utmost importance of data in AI systems, the model that has been developed for illustrating Cross-Industry Standard Process for Data Mining (CRISP-DM) is one of the relevant methods of understanding of the life cycle of systems. The model is designed to be domain-agnostic and as such, is now widely used by industry and research communities. According to Plotnikova et al., *"the CRISP-DM is to be considered as 'de-facto' standard of data mining methodology."*[189] The model provides an overview of the life cycle of a data-mining project, as shown in Figure 10. It contains the project phases, their respective tasks, and the relationships between these tasks.

Figure 10. CRISP-DM phases



Source: P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer & R. Wirth (2000), "CRISP-DM 1.0: Step-by-step data mining guide" *SPSS inc.*, Vol. 9, No. 13.

[189] V. Plotnikova, M. Dumas and F. Milani (2020), "Adaptations of data mining methodologies: a systematic literature review", *PeerJ Computer Science 6*, p. e267.

Notably, the two models overlap insofar as the different phases of data understanding, preparation, modelling, evaluation, and deployment somewhat coincide with the steps illustrated in Figure 9, thus reinforcing the appreciation of the AI life-cycle model.

Accordingly, the security of AI systems must be considered at all their life-cycle stages, from their creation to their termination. Unlike the software development life cycle, which provides for software to perform as expected as long as the requirements are not changed, with AI the underlying characteristics of the data might change, so your models, which are built based on a specific dataset, may not be giving the right results.

We repeat here that whatever knowledge is encoded in an AI algorithm, program, or system, computers actions can reach huge distances at the speed of light. This also holds for evolving ML systems. Therefore, new mindsets are required to ensure that risks related to their ubiquitous utilisation remain manageable.

In this context, the assessment of the risk associated with the product is also key. A product's risk is considered to be the combination of impact and the probability of an attack under the condition of the product's intended use. Apart from addressing the whole product life cycle, security requirements should be established on a risk-based approach, and the AI systems' risk should be assessed on the product's intended use.

…addressing AI systems security is first and foremost related to maintaining accountability across systems, and across distributed sources of data and analysis, which is mandated by all contemporary adopted AI principles…

In terms of secure life-cycle development, addressing AI systems security is first and foremost related to maintaining accountability across systems, and across distributed sources of data and analysis, which is mandated by all contemporary adopted AI principles, notably those of the OECD/G20 adopted in 2019.[190] Security and accountability both require knowledge of the system' architecture, and having it specified in terms of the design and the documentation of its components and how they are integrated.

The following documentation is important to validate, explain, and secure AI systems.

- Have logs related to the development/coding/training of the system, recording who changed what, when, and why. These are standard for revision control systems used in developing software, which also preserve older versions of software so that differences and additions can be checked and reverted.
- Provide cybersecure pedigrees for all software libraries linked to that code.
- Provide cybersecure pedigrees for any data libraries used for training any ML algorithms used.
- Ensure that all data used is compliant with legal obligations and with corporate/institutional codes of conduct or ethical charters.
- Where ML is used, keep records of the model parameters and of the training procedure.

[190] OECD (2019b), op. cit.

- Keep records demonstrating due diligence has been followed when testing the technology, before releasing it, preferably including the actual test suites themselves so that these may be checked and then reused.
- For AI-based operating systems, maintain logs of inputs and outputs/decisions in files that are cybersecure and GDPR compliant. Depending on the capacities and the nature of the AI system, run logs can be kept just for the duration which is strictly necessary.
- Build resilience into the AI system through reset functions that can be used to bring the system back into a proper state, taking into account the timespan needed to implement corrective actions, in case of successful attacks or model drift. These could include fail-safe and kill switches.
- Where the operation of AI systems is outsourced to third parties, ensure that the overall security strategy and intellectual property remain consistent throughout the whole outsourcing chain.

Of course, all this documentation, as well as the system itself, must be cybersecure, not only to defend the system, but also to defend the systems' users and their environment.

Software engineering best practice can then be used to help develop AI systems that are more secure. AI systems then become part of a solutions life cycle in which solutions are crafted to address specific operational challenges. Such solutions are designed, operated, and maintained, while monitored and controlled. If the deployed solutions are not working properly and according to the established requirements, they are retired or, in the case of ML systems, retrained to better adapt their model to the operational challenges.

Cybersecurity risks clearly exist at all these stages and, again as per best practice, a control framework is developed to mitigate such risks. However, in the case of AI systems, the control framework is influenced by the degree of autonomy of the AI system and its purpose. There are differences, depending on whether the purpose is to, for instance:

- discover patterns
- provide answers
- provide human augmentation capabilities
- be in a non-lethal autonomy mode
- be in a lethal-autonomy weapon.

As such, once again, the AI system's intended use is central to understanding the level of risk of the system itself and the relative appropriate mitigating measures. Based on the life cycle of AI systems illustrated above, a control framework for AI can be designed. The main requirements of these stages are recalled in brief as follows:

Ideation stage

A. Clarify the level of the autonomy target, as this will define the control framework (see above).
B. Develop a clear understanding of dual-use and multi-use risks. If it wasn't understood that an autonomous car can be used both as a transportation means and as a mobile weapon, nobody would be worrying about securing pedestrians or urban landmarks.
C. Clarify the expected outcomes.

D. Analyse competing hypotheses.

Good DevOps are clearly needed at this stage. It is important that organisations do not change these elements while the system is running without reassessing the related risks. For instance, the organisation should not change the level of autonomy from automatic with manual validation to fully automated.

Planning stage

A. Understand what the interface being built is.
B. Understand what type of feedback the AI system needs to be efficient.
C. Understand what type of monitoring framework is required, not only on the AI algorithm but also on the AI system the algorithm is driving.
D. Gather requirements on security risk assessment, privacy risk assessment, risk-level acceptance, and those that are informational, functional, and behavioural.

Design stage

A. Define the dataset semantics, namely the language of reference upon which the AI system will rely.
B. Monitor the required data and assess the need (and possibility) to desensitise the data.
C. Define command and control interfaces, including the design of kill switches and security gates.
D. Define layers of control on the various interfaces for both inputs and outputs.
E. Analyse the attack surface and threat modelling.
F. Monitor the self-evolution of the algorithm, so that its new outputs are understood.
G. Define the target operating model, especially in cases where the AI system interfaces with the outside world and has influence on processes that impact life experience.

Building stage

A. Ensure that the infrastructure is resilient, which includes Telecom and Container aspects.
B. Build the training dataset.
C. Consider securing software development through automated computer-aided software engineering (CASE) tools and static analysis.
D. Build a safe learning repository to create the capacity to monitor the evolution of the ML system. This implies the ability to store the successive models at different learning stages, which in turn helps in the explainability of the ML system.
E. Ensure logging and auditing capabilities across the life cycle.
F. Ensure that the kill switch is functional.

Deployment stage

A. The target operating model must be reaffirmed.
B. Confirmation of the deployment containers' security.
C. Confirmation of the interface control effectiveness.
D. Confirmation of the resilience and monitoring.
E. Existence of a test incident response plan.
F. Ensure control for suspicious behaviours while the system is working.
G. Ensure the existence of effective control and auditing procedures, to avoid that any reactions in the case of unexpected behaviour by the ML system are not pure improvisations. Such systems may have potentially lethal consequences, and at a pace that no human being can cope with.

The scheme presented above is certainly full of complexities that need understanding and resources. However, from a technical perspective, it is only by starting to share good practice about the mapping of these complexities that vulnerabilities in AI systems can start to be addressed.

PART IV.
POLICY ISSUES AND RECOMMENDATIONS

1. Introduction

AI in cybersecurity has been presented throughout this report in terms of its great opportunities. But, as with any powerful general purpose, dual-use technology, it also brings great challenges. AI can indeed improve cybersecurity and defence measures, allowing for greater system robustness, resilience, and responsiveness. Yet AI in the form of machine learning and deep learning will allow sophisticated cyberattacks to escalate, making them faster, better targeted, and more destructive. The application of AI in cybersecurity also poses security and ethical concerns. Therefore, to avoid delays in the realisation of the beneficial applications of AI in the cyber domain, public policy should step in to avoid what some economists call the 'technology trap.'[191] This term refers to the fear that the use of AI in cybersecurity in the short run will make things worse for everyone in the long run by slowing the pace of automation and innovation.

This chapter assesses the major policy issues and regulatory frameworks related to the use of AI in cybersecurity and presents policy recommendations to ease the adoption of AI in cybersecurity in Europe.

2. Current and future AI laws: accountability, auditability, and regulatory enforcement

The report has introduced the legal frameworks where AI and cybersecurity intersect. As mentioned in the introduction, the European Commission's "Regulation on a European Approach for Artificial Intelligence"[192] fosters ad hoc protection for high-risk AI systems, based on a secure development life cycle. The Regulation provides that solutions aimed at ensuring the cybersecurity of high-risk AI shall encompass measures to prevent and control attacks trying to manipulate the training dataset inputs ('data poisoning') designed to cause the model to make a mistake ('adversarial examples'), or model flaws. The OECD includes the safety factor in the most important indicators for a context-based assessment of AI systems.[193] It is fair to say that these requirements represent a fundamental step forward in assuring the safety of AI systems.

Notably, issues of accountability and regulatory enforcement cannot be avoided by decision makers wanting to adopt a regulatory approach and this section now makes some preliminary observations in this regard.

Thinking about initiatives to enable appropriate regulatory enforcement is perhaps one of the most important elements to consider. Creating the precondition for the development of AI products and systems means that requirements for AI systems have to focus on post-compliant

[191] C.B. Frey (2019), *The technology trap: Capital, labor, and power in the age of automation*, Princeton University Press: Princeton.

[192] European Commission, Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts, COM(2021) 206 final, Brussels, 21.4.2021.

[193] OECD (2019a), op. cit., p. 99.

ethics. But there is also an important part of 'pre-compliant' that means legal-compliant. In this context, the work the European Commission has carried out in shaping the EU AI Regulation is noteworthy.

These efforts will nonetheless be severely hampered if regulators are not appropriately supported with adequate funding resources and expertise. The GDPR has meant a significant increase in workload for national data-protection authorities. Millions of euros in fines and sanctions have brought national watchdogs under the spotlight of mainstream media. However, the situation appears different from what could be inferred from the news. Unfortunately, not all authorities have found a commensurate amount of human and financial resources to be able to respond to such a growth in workload.[194] The consequences of this are that the citizenship's trust in the enforcement capabilities of national authorities declines, and there is potentially less adherence to the law by those actors that are supposed to be audited and scrutinised.

The fragmentation of enforcement powers and capabilities in the data protection sector across the EU member states is a lesson the EU should carefully consider. If there are no policy initiatives that aim to coordinate and support the use of appropriate skills and financial resources, there will be no effective pan-European AI regulation.

Furthermore, regulatory enforcement and accountability should be made complementary. Even with appropriate budget and coordination, we cannot expect that a regulatory authority will be able to closely follow each and every aspect of the security of AI systems in all firms. In organisations deploying AI (including for cybersecurity) with large-scale impacts, policy actions should enable both auditability by means of third-party authorities and interventions on their existing governance culture and decision-making processes. For instance, initiatives aiming to make the adherence to ethical and safety principles a prerequisite for the procurement of AI applications could be implemented. This would raise the discussions on AI and safety at organisations, including at board level.[195]

At the deployment level, record keeping and logging both system design and its life cycle will enhance accountability by augmenting human control over AI.[196] These aspects are reflected in the current AI Regulation. Initiatives to improve transparency in corporate governance will

[194] For instance, see Deloitte (2019), "Report on EU Data Protection Authorities", Deloitte Privacy Services – Privacy Response.

[195] See M. Veale (2020), "A Critical Take on the Policy Recommendations of the EU High-Level Expert Group on Artificial Intelligence", *European Journal of Risk Regulation*, *Faculty of Laws University College London Law Research Paper*, No. 8.

[196] Joanna Bryson, for instance, introduces these due diligence mechanisms as follows: "*Due diligence can be demonstrated by corporations despite the fact they employ people. People learn, have free will, and have incomprehensible systems of synapses making up their action selection mechanisms. Many humans are dishonest, careless, or sometimes just upset or incompetent. Nevertheless, we can construct systems that ensure that humans working together tend to succeed. These systems generally include records, such as financial accounts, access permissions, and meetings where executive decisions are agreed. They also include external regulatory bodies and law enforcement. Exactly the same kinds of procedures can be used for retaining control of AI*", J. Bryson and A. Theodorou (2019), *How Society Can Maintain Human-Centric Artificial Intelligence*, in Marja Toivonen and Eveliina Saari (eds, Human-Centered Digitalization and Services, Springer.

nonetheless have to complement the fact that the auditing of datasets, Application Programming Interface (API), or models, may not be made openly public, unlike normal praxis, for security reasons:[197] *"In already deployed systems that require both verified fairness and security, such as AI-based bond determination, it will be difficult to balance both simultaneously. New methods will be needed to allow for audits of systems without compromising security, such as restricting audits to a trusted third party."*[198]

Accountability and regulatory oversight are interconnected. Policies aiming to support regulatory enforcement efforts should complement the existing legislative initiatives to ensure an adequate level of supervision over the security risks of AI.

3. Existing legal frameworks: EU cybersecurity

The European cybersecurity legal landscape has been significantly boosted over the past five years by a couple of new legal acts, the Network and Information Security (NIS) Directive (2016)[199] and the Cybersecurity Act (CSA).[200] The NIS Directive introduced the requirements for member states to establish national Computer Security Incidents Response Teams (CSIRTs), representatives for a pan-European cooperation group, and national cybersecurity strategies. Furthermore, it has established a list of actors (divided into two groups of Operators of Essential Services (OESs) and Digital Service Providers (DSPs)) required to implement certain information security requirements.

This legislation has undoubtedly helped to harmonise national legal landscapes on the matter. However, lawmakers could have taken the opportunity of the recent revision of the Directive to consider mapping the use of AI systems in determined critical sectors comprising DSPs and OESs, while not leaving anyone out. While this would have required the effort of having to define AI in some way, thereby presenting a threshold therein, it could have led to a coherent overview of all those essential services which deploy AI to protect themselves. Many national critical infrastructures are coming under the definition of OESs, and the use of AI for the security of their network could bring benefits and drawbacks. In this respect, a complete mapping could help in understanding the level of (inter)dependency of our essential services, the vulnerabilities therein and the available technologies to reduce such risks. This exercise could lead to decisions such as mandating the insulation of certain services from determined technologies in times of cyber crisis. Similar exercises aiming to ensure the business continuity

[197] M. Comiter (2019), op. cit., p. 74.

[198] Ibid.

[199] European Commission (2016), Directive (EU) 2016/1148 of the European Parliament and of the Council of 6 July 2016 concerning measures for a high common level of security of network and information systems across the Union,OJ L 194, Brussels.

[200] European Commission (2019), 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 OJ L 151, Brussels.

of a determined number of critical infrastructures are already being discussed by lawmakers overseas.

The CSA[201] was issued in the form of a Regulation (i.e., directly applicable across all EU member states), and introduced two main items in fulfilment of the first package of EU cybersecurity initiatives.[202] On one side, the mandate of the EU Agency for Cybersecurity (ENISA), was expanded and made permanent. On the other side, a certification scheme for product, processes and services was laid down, though the non-binding nature of such a measure has left the determination of the relevance of this initiative to the market.

It is probably too early to assess the effectiveness of the AI cybersecurity certification measures introduced by the CSA. However, the certification schemes will be looked at more closely later in this report. This Task Force has observed how many commentators hope for a prominent role for ENISA.[203] The pervasiveness of AI technology in the future is driving the need for a unitary effort, which cannot be addressed without a European entity overseeing national initiatives.[204] An enhanced role for ENISA in this respect will also allow the agency to be better equipped to influence international developments. The CSA has provided promising resources and long-term vision to ENISA to undertake these challenges, creating a valuable source of support in the coordination of a European effort.

The prominent role of ENISA in EU cybersecurity policy could also add value in the monitoring and observation of the current research on both security of AI and AI for cybersecurity. Continuing (and possibly enhancing) the relationship with academic research centres and researchers in these fields could leverage the multitude of national initiatives to a more coherent and consistent vision to make the EU a global competitor in AI and security research activities.[205]

With respect to broader general principles, this Task Force has observed the importance of the security-by-design aspect, which implies thinking about security from the inception, design or implementation process, with a focus on the desired security capabilities and outcomes driving the selection of tools and technologies. This holds true for AI systems and for the data these

---

[201] European Commission (2019), 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 OJ L 151, Brussels.

[202] All EU cybersecurity initiatives were gathered in 2017 under the 'package', or 'framework'. See European Commission, Shaping Europe's digital future Cybersecurity Policies (https://ec.europa.eu/digital-single-market/en/cyber-security).

[203] S. Fantin (2019), "Weighting the EU Cybersecurity Act: progress or missed opportunity?", *CITIP Internet Publication, KU Leuven*, March.

[204] M. Taddeo et al. confirm this to be a needed agenda item for the Athens-based agency: "*It is crucial that ENISA will focus also on AI systems, otherwise the certification framework will at best only partially improve the security of digital technologies and services available on the EU market.*". ", M. Taddeo, T. McCutcheon and L. Floridi (2019), op. cit., pp. 557–560.

[205] It will be fundamental to follow up on the initiatives arising from the Horizon2020 SU-ICT-03-2018 (Establishing and operating a pilot for a Cybersecurity Competence Network to develop and implement a common Cybersecurity Research & Innovation Roadmap), whereby four projects were funded with the aim of enabling a European cybersecurity competence network.

systems rely on. In this context, AI developers or *deployers* are required to answer two basic questions. Firstly, what are the security risks in implementing AI (including in cybersecurity capabilities), and secondly, what is the actual and factual need or necessity for implementing AI in an existing system?

AI cybersecurity certification efforts should be coordinated by ENISA. The introduction of principles aiming to enhance security of our networks in light of AI deployment should follow a proactive approach and demand assessment actions be taken prior to any deployments, as well as during the whole cybersecurity-aware/-oriented life cycle of a product, service or process.

4. Major policy issues

4.1 Delegation of control

The European Commission High-Level Expert Group[206] has used the adjective *trustworthy AI* in its reports several times, so that it becomes almost a motif for the group's efforts. The White Paper[207] and the recently adopted EU AI Regulation[208] by the European Commission repeats the term, thereby making 'trust' the attribute that best represents the Union's vision on AI, and a precondition for people's uptake of this technology.

This Task Force, however, has collected opinions on the use of this terminology in relation to the concept of reliability and control. When considering the relationship between AI and cybersecurity, Taddeo, Floridi, and McCutcheon argue[209] that relying on the concept of trust is somewhat misplaced. They believe the presence of unforeseen vulnerabilities in AI systems could compromise and negate the benefits in terms of response, robustness and resilience in the protection of information systems.[210] For this reason, they say, the concept of reliance draws upon a higher notion of control and supervision over the machine than the concept of trust: *"(…)while trust is a form of delegation of a task with no (or a very minimal level of) control of the way the delegated task is performed, reliance envisages some form of control over the execution of a given task, including, most importantly, its termination."*[211] The concept of Reliable AI will be further discussed in this chapter.

Control has been long debated in cyber and technology policy and examined as a means of analysing the relationship between users or designers and the machine.[212] Notwithstanding the

[206] High-Level Expert Group on Artificial Intelligence (2019), *Ethics Guidelines for Trustworthy Artificial Intelligence*, April (https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai).

[207] European Commission, White Paper On Artificial Intelligence - A European approach to excellence and trust, COM(2020) 65 final, Brussels, 19 February 2020.

[208] European Commission (2021), *op. cit.*

[209] M. Taddeo, T. McCutcheon and L. Floridi (2019), op. cit.

[210] See more on this in the section of this report of the transformation of the threat landscape.

[211] M. Taddeo, T. McCutcheon and L. Floridi (2019), op. cit.

[212] See for instance, L. Lessig (2006), *Code and Other Laws of Cyberspace (Version 2.0)*, Creative Commons, Basic Book: New York. See also S. Fantin (2019), "Sixty Minutes with Steve Crocker", *SSRN 3477815.*

proven benefits of AI for protecting and securing existing systems, it was observed and deemed appropriate to recommend the adoption of policies that mandate a cautious evaluation of the level of control on an AI system based on the tasks against which it was designed. This behaviour would therefore be much closer to the concept of reliability than to trustworthiness. As a practical outcome, for instance, red teaming[213] or suitability tests could support the assessment of how an existing information system could rely on AI for security purposes, leading to different sorts of results, from full human control to total AI autonomy.[214]

4.2 Openness of research

The 2019 US National Security Artificial Intelligence Commission Report begins by introducing the background context as follows: "*AI is not hidden in a Top-Secret Manhattan Project. Tech luminaries and blue-ribbon panels have sounded alarm bells on AI's peril and have championed its promise. Big countries and big tech companies (…) are leading the way, but the algorithms that fuel the applications are publicly available. Open applications and development tools are diffusing at an accelerating pace.*"[215]

It was observed in the reviewed literature and during the meetings of this Task Force that the research on AI has always encouraged a great deal of openness. Nonetheless, for many of the reasons set out above (the weaponisation of AI, the transformation of the threat landscape and the multipurpose nature of AI, etc.), a full openness policy for research outputs might not always be the perfect choice. Limiting the dissemination and the sharing of data and codes could enable a more complete assessment of the security risks related to the technology and its vulgarisation. However, striking a balance between appropriate limitations and the fundamental interest of our society in pursuing innovation and openness is not a trivial task.

…models are often made 'open source' since their research has successfully led to AI applications serving a broad general interest. However, poor cybersecurity in the protection of these models may lead to hacking opportunities for actors seeking to steal such information.

The publication of AI components and its subsequent use for malicious purposes finds different examples. For instance, models are often made 'open source' since their research has successfully led to AI applications serving a broad general interest. However, poor cybersecurity in the protection of these models may lead to hacking opportunities for actors seeking to steal such information.[216] This does not necessarily imply that the only viable option would be to limit the publication of AI research; rather that researchers should verify as much as possible the level of cybersecurity of libraries and tools and perform a review against misuse/prevention measures prior to the publication of their research.

[213] M.C. Horowitz et al. (2018), op. cit., p. 13.

[214] More will be explored below. See also, M. Comiter (2019), op. cit. p. 59.

[215] US National Security Commission on Artificial Intelligence (2019), Final Report (www.nscai.gov).

[216] M. Comiter (2019), op. cit., p. 26.

In fact, the research environment has often been open about its successes; publication and sharing of results has proved the whole experts' community has more opportunity to optimise previous works as well as a chance to enable reproducibility.[217] In general terms, this practice has an immensely positive value for society as a whole. The tendency to publish research results is the very essence of academic work, instrumental for the completeness of scientific analysis.[218] We should bear in mind that this is not only advocated for reasons of broad, principles-based scientific integrity, or for the economic benefits of enabling collective innovation. In particular, openness in AI research has, according to many, several sector-specific benefits. It "*makes it easier to uncover and fix flaws in software and expose biases in data sets that are difficult for a researcher's own colleagues to spot,*"[219] while nonetheless improving interoperability and usability.[220] Therefore, any attempt to delimit the dissemination of results in AI research should be made with extreme caution, with particular attention paid to the balancing of secrecy due to security risks and dissemination interests.[221]

A growing number of experts at the intersection of AI and cybersecurity have expressed the need to re-evaluate and re-establish normative frameworks and behaviours to preserve the research outputs while mitigating the risk of AI components being misused. Any initiative in that sense should not be aiming to neutralise such risks *in toto*. Rather, measures could instead be oriented towards affording enough time after a result becomes scalable to carefully evaluate the security risks carried within. The Partnership on AI (PAI) is undertaking a multistakeholder project to facilitate the exploration and thoughtful development of publication practices for responsible AI, building on previous activities related to high-stakes research. By convening with the AI/ML research community, the PAI explores the challenges and trade-offs in responsible publication to shape best practice in AI research.[222]

There are numerous proposals to delimit the openness of research to mitigate AI security risks, and many of them invite us to look at long-established practices in other disciplines. In the cybersecurity community, for instance, researchers hunting down software flaws are normally engaged in a 'coordinated vulnerability disclosure', whereby the publication of such a vulnerability is being held in order to allow enough time for the manufacturer to release a patch. Similarly, AI research could be held from publication to enable an assessment over the inherent security risks. However, the implementation of such an approach seems to prompt several questions about the acceptable level of mitigation measures, the fields and applications

[217] K. M. Slayer (2020), "Artificial Intelligence and National Security", Congressional Research Service, November, p. 33.

[218] For an account on secrecy vs. openness, see M. D. Resnik, (2012), *Openness versus Secrecy in Scientific Research*, Cambridge University Press, p. 2.

[219] J. Léveillé (2019), "Embrace Open-Source Military Research to Win The AI Competition", RAND Corporation and War on the Rocks.

[220] S. Sonnenburg et al. (2007), "The Need for Open Source Software in Machine Learning", *Journal of Machine Learning Research.*

[221] See also M. D. Resnik (1996), "Social epistemology and the ethics of research", *Studies in History and Philosophy of Science.*

[222] For more on this, see Partnership on AI, Publication Norms for Responsible AI (www.partnershiponai.org/case-study/publication-norms/).

where responsible disclosure could be adopted, and what exact types of mitigation measures should be implemented, given the uncertainties in the state-of-the-art AI technologies.[223]

Brundage et al. suggest that a proper form of pre-publication 'security risk assessment' could help the process of marking the level of classification of a given research output. This exercise would then support the determination of the amount of information to be disseminated. In combination with the coordinated vulnerability disclosure approach, this could also help the balancing of the equities at stake: "*In the case of AI, one can imagine coordinating institutions that will withhold publication until appropriate safety measures, or means of secure deployment, can be developed, while allowing the researchers to retain priority claims and gain credit for their work.*"[224] Of course, the implementation of such proposals in daily research might not be immediate, nor should it be. More open discussions on the feasibility of these initiatives are needed, accompanied by the rollout of adequate legal protections for all the actors involved in the process and adequate investments for raising the awareness towards a "*sustainable culture of AI security development.*"[225]

Having said that, two caveats regarding the adoption of measures for delimiting the openness of AI research are worth considering. The first refers to the potential adoption of coordinated vulnerability disclosure processes in AI vulnerabilities. It highlights a big conceptual difference between the approach behind traditional cybersecurity vulnerabilities and AI-related ones, and among other things, from a geopolitical angle. In the traditional scenario, the choice of somebody discovering a vulnerability is whether they communicate it to whoever may patch it. This behaviour holds true because the vulnerability is unknown, but the remedy is, whereas in new types of AI vulnerabilities the contrary holds true: the vulnerability/security risk is known, but the remedy is not. "*This potential situation poses significant ethical and policy questions. It is important for a country to realize that the disclosure of any protective techniques will have impacts beyond its own borders.*"[226] Likewise, these reverse dynamics may have an impact on AI-responsible disclosure programs that have not been explored appropriately.[227]

The second warning refers to the actual target of these policies when adopting measures restricting openness in AI, particularly in light of disclosing potential advantages or weaknesses to adversaries. This highlights a major difference from other sectors where the disclosure of research outputs is restricted by security, safety, or military concerns. While we have seen that such measures could potentially be aimed at the whole line of research (publications), some argue that this could become very challenging in practice. The AI research community already functions with a 'by default' openness policy, and the technical features of an AI system may not be the crucial element to protect. Furthermore, these challenges couple with the multi-use nature and adaptation of open-source models and material. This is why some suggest that the

[223] M. Brundage et al. (2018), op. cit.

[224] Ibid.

[225] J. Pandya (2019), "The Dual-Use Dilemma Of Artificial Intelligence", *Forbes*, 7 January (www.forbes.com/sites/cognitiveworld/2019/01/07/the-dual-use-dilemma-of-artificial-intelligence/?sh=487a66df6cf0).

[226] M. Comiter (2019), op. cit., p. 46.

[227] M. Brundage et al. (2018), op. cit., Annex 2.

focus should be on the applications where such features will be deployed. "*What needs to be hidden are the intended uses of a technology.*

What needs to be hidden are the intended uses of a technology.

Disclosing an AI technology that inadvertently exposes previously unexpected strategic or tactical objectives could indeed provide critical information to potential adversaries. […] This is unlike progress in most other military technologies, which is driven by research from a few well-funded state actors. This fundamental difference has important repercussions for sharing both algorithms and data."[228]

4.3 Risk-assessment policies and suitability testing

The focus of policy efforts must also be on enabling developers, deployers and end users to understand as much as possible the risks of implementing AI in running applications. As far as developers and deployers are concerned, the expansion of the vulnerabilities landscape confirms the need for a whole chain of measures targeting insecurity, including processes aimed at mapping the threat landscape. Similarly, the possible physical and online risks of using AI need to be better comprehended by end users.[229]

The adoption of an AI system may lead to several outcomes from the perspective of its inherent security risks. According to its purpose and use cases, such risks might be accepted and balanced differently. Organisational measures are therefore needed in each sector to carefully assess the feasibility of the deployment of AI in a determined context. Such measures often take the form of risk-assessment mechanisms. What is important to note here is that risk assessment should not be intended as aiming to neutralise risks, but rather at mapping and mitigating them in an accountable way. Indeed, risk may depend on the context and the end use of deployment.[230]

Within such an approach, several exercises have been proposed that would support the assessment of risks in a determined AI-deployment context. For instance, Comiter suggests the introduction of mechanisms to deploy so-called *security compliance programs*.[231] One of the main drivers here is the full cycle of the proposal, i.e., the fact that such measures should be in place during planning, implementation, and mitigation stages. The most interesting one is in the first category (planning), whereby it is recommended, to evaluate the related security risks, that a *suitability testing* exercise could be mandated before the adoption of a determined AI system.

---

[228] J. Léveillé (2019), op. cit.

[229] K. M. Slayer (2020), op. cit.

[230] "*The use of risk management in AI systems lifecycle and the documentation of the decisions at each lifecycle phase can help improve an AI system's transparency and an organization's accountability for the system.*" Furthermore, the OECD summarises the six main steps for AI-systems risk management as follows: *"1. Objectives; 2. Stakeholders and actors; 3. Risk Assessment; 4. Risk mitigation; 5. Implementation; 6. Monitoring, evaluation and feedback."* See OECD (2019a), op. cit., p. 96.

[231] M. Comiter (2019), op. cit., p. 56.

Such tests should be performed by all stakeholders involved in a deployment project. Under such premises, suitability testing would mean weighting several stakes, namely value, ease of attack, damage, opportunity cost and alternatives. More specifically, the value added has to be examined with respect to the societal and economic benefits associated with its potential implementation. The ease of the attack is determined according to the nature of the dataset or model used, such as their public availability or the easiness of their reproducibility, among other things. The damages are calculated according to both the likelihood and the possible ramifications of an attack. The suitability test also has to include the opportunity costs of not implementing the system, meaning loss in societal benefits. Finally, the test has to take into due consideration whether valuable alternatives to AI that could deliver the same societal benefits exist. This holds especially true when considering the current tendency for private firms to race to implement AI systems in the name of innovation and progress. Boosted by the conviction that adopting AI systems is compelling in reputational and competitive terms, many companies fail to thoroughly evaluate equally valuable alternatives.[232]

Such a suitability test could be combined with a review of the data-collection and sharing practices. AI developers and users must validate their data-collection practices so as to identify all potential weaknesses that could facilitate an attack: namely, how data have been collected; does the attacker have access to the same dataset; can s/he recreate it? Restriction in data-sharing practices also has to be considered, particularly for critical AI applications, to reduce the ease of crafting an attack. The balancing of these interests should lead to an "*implementation decision.*"[233]

Three factors are important in this measure's benefits. First, the result of the implementation decision would support the understanding of the acceptable use of AI in an existing information system, having considered the unknown risks and vulnerabilities that new deployments bring along. It is not a binary choice (implement/do not implement), but rather a gradual decision, measured against the benefits and shortcomings (including societal, ethical, legal, and economic factors) that comprise the spectrum of full AI autonomy, and the different degree of human oversight.

Second, the adoption of suitability testing would force the stakeholders to question the necessity for adoption of a certain AI system. If, after such an exercise, the security shortcomings would largely overcome the benefits of goals such as automation, scalability, or effectiveness of a task, then a clear indication that no actual need would justify an investment on AI.

Third, suitability testing would support accountability and auditability, in that it would force stakeholders to demonstrably map the link between adoption of an AI system and its related purpose. This is to avoid the subsequent and (un)intended reuse of AI for other finalities in the same information system, which could enhance the associated security risks. In certain sectors deemed critical for our society (for instance, law enforcement or healthcare), the use of AI systems might have to be strictly related to the purposes for which they were deployed. Suitability testing might therefore support auditing the intended purpose, its optimisation and the uses therein.

---

[232] Ibid.

[233] Ibid.

4.4 Oversight

Ensuring human control on AI calls for a level of expertise by humans to exercise 'true' control or oversight.

Ensuring human control on AI calls for a level of expertise by humans to exercise 'true' control or oversight. In particular, it is necessary to protect against what can be called the 'half-automation problem.' What is meant by this is the phenomenon in which, when tasks are highly but not fully automated, the human operator tends to rely on the AI system as if it were fully automated.

Semi-autonomous vehicles provide multiple examples of the 'half-automation problem'. These vehicles are not fully automated; they require full human attention and the ability for humans to intervene at split-second notice. Another example is the automation of industrial processes that similarly require humans to pay full attention in case the AI system malfunctions. However, use patterns show that humans do not maintain this level of control despite requirements to do so; they tend to shift their attention or lose it completely. This should be explicitly considered, as broad directives to maintain human control may not be followed in practice.

The need to have stop buttons or kill switches to maintain human control over AI processes is also relevant, especially in the field of an automated response. But are kill switches an option in all cases? Do humans in control have to take back full control? In some cases, like the self-driving vehicles or auto-pilot on modern aircraft, the issue of when to return control is an open question that still needs to be answered. According to studies carried out on handover scenarios (the act of transitioning control from a vehicle to driver), a sudden handover request needs to be generally avoided, as it is highly unlikely that a driver will be able to keep the car on the road if s/he has only a few seconds to react. Simulator studies with professional truck drivers showed that they could respond well to timely handover requests in non-critical situations, however.[234]

Hence, while the ability to introduce kill switches for AI systems might be desirable, their feasibility is still debatable. Besides, there are different levels of kill switches. They can indeed be autonomous, namely complex electrical systems, or they can be non-autonomous. In this context, the question would be whether a master kill switch is advisable, and who should be in charge of controlling this system. Therefore, while applying kill switches in AI systems is non-trivial, some form of control over those systems, such as security gates, must nonetheless be established.

[234] B. Zhang et al. (2019), "Transition to manual control from highly automated driving in non-critical truck platooning scenarios", *Transportation Research Part F, Traffic Psychology Behaviour*, pp. 84-97.

4.5 Privacy and data governance [235]

It has been made clear in this report that data represent an essential element for the functioning of AI systems; to use a metaphor, data is the fuel for the running of AI engines. In certain cases (but not all AI-related), such data are personal data, thus falling under the data protection legal framework. The EU legal system on data protection regarding AI and cybersecurity shall therefore be analysed in more granular detail.

The EU personal data protection framework is composed of various levels of legal sources. The right to personal data protection, beyond being embedded in Article 16 Treaty on the Functioning of the European Union (TFEU), is enshrined in the Charter of Fundamental Rights of the European Union, at Article 8, which reads: "*1. Everyone has the right to the protection of personal data concerning him or her. 2. Such data must be processed fairly for specified purposes and on the basis of the consent of the person concerned or some other legitimate basis laid down by law. Everyone has the right of access to data which has been collected concerning him or her, and the right to have it rectified. 3. Compliance with these rules shall be subject to control by an independent authority.*"[236]

In recent years, the European Court of Justice (CJEU) interpreted the essence of this right as including principles of data security,[237] thus giving to its core a rather 'technical' character. While this approach was heavily criticised by EU law experts and constitutionalists,[238] for the purposes of this section it will be sufficient to highlight that the European jurisprudence sees principles of data security, such as confidentiality, integrity, and availability, as inextricably linked to the enjoyment of the essence of the right to data protection.

Having said that, secondary law on personal data protection is composed predominantly of the GDPR. A wide variety of principles therein are highly relevant for the deployment of AI that makes use of personal data, for instance with respect to the stringent vetting on the datasets used to train and learn AI systems in order to comply with the data protection, data quality and data accuracy principles,[239] or the uncertainties behind techniques and measures in use to enable data-sharing practices among cybersecurity firms. It is worth mentioning that the GDPR not only uses the classic confidentiality, integrity, and availability (CIA) triad mentioned above, but also introduces resilience in Article 32 as the fourth constituent part of secured processing

[235] The authors wish to extend their gratitude to Eleftherios Chelioudakis (Homo Digitalis) for the precious feedback and valuable contribution to this section on data protection.

[236] European Commission (2012), Consolidated versions of the Treaty on European Union and the Treaty on the Functioning of the European Union - Protocols - Annexes - Declarations annexed to the Final Act of the Intergovernmental Conference which adopted the Treaty of Lisbon, signed on 13 December 2007 - Tables of equivalences, Official Journal C 326, 26/10/2012 P. 0001 – 0390, Brussels.

[237] See for instance, Joined Cases C-293/12 and C-594/12, Digital Rights Ireland, ECLI:EU:C:2014:238., para. 40.

[238] For example, M. Brkan (2018), "The Essence of the Fundamental Rights to Privacy and Data Protection: Finding the Way Through the Maze of the CJEU's Constitutional Reasoning", The Essence of Fundamental Rights in EU Law, Leuven, May 17.

[239] European Union Agency for Fundamental Rights (2019), "Data quality and artificial intelligence – mitigating bias and error to protect fundamental rights", 11 June (https://fra.europa.eu/en/publication/2019/data-quality-and-artificial-intelligence-mitigating-bias-and-error-protect).

systems. Such an addition recognises that for their proper functioning, these systems need also to withstand and recover from disruptions. Therefore, the reliability of information systems processing personal data (such as the systems' fault tolerance and/or absence of single points of failure) is important for the development of the digital single market and the provision of services within the EU.[240] Finally, the addition of the accountability principle in the GDPR is also closely related to new obligations on security, since data controllers and processors need to not only apply security measures, but also mandatorily document them (for example through privacy policies, records of processing activities, data-processing agreements, etc).

4.5.1 Application of GDPR in securing AI and in using AI for cybersecurity

It is not within the scope of this chapter to understand the impact of AI and cybersecurity on the GDPR, and vice versa. However, open questions that could help policymakers and regulators supporting the application of the GDPR in the two use cases we hold at stake in this report (i.e., securing AI and using AI for cybersecurity) will be analysed. In this context, it is relevant to point out some main provisions of the GDPR with respect to security and AI.

Article 32 GDPR requires controllers and processors to implement technical and organisational measures to ensure the security of the processing of personal data appropriate to the security risks therein. Measures such as encryption, pseudonymisation or anonymisation are given as an example of mechanisms to protect the CIA of personal data. Furthermore, Article 35 requires that a data protection impact assessment is carried out prior to the processing of personal data any time a new technology that is intended to be deployed in an existing information system brings along high risks for the rights of individuals.[241] Moreover, paragraph 3 makes this exercise mandatory when "*a systematic and extensive evaluation of personal aspects relating to natural persons which is based on automated processing, including profiling, and on which decisions are based that produce legal effects concerning the natural person or similarly significantly affect the natural person.*" Seeing these provisions in the context of AI-based tools, the GDPR requires both appropriate security measures and a proper data protection impact assessment to accompany similar implementations. This holds true for both AI used for cybersecurity and for securing AI systems.

A number of provisions leave unanswered questions or interesting policy inputs about the impact of the GDPR on the two use cases mentioned above. In both these end uses, AI systems often enable practices of profiling or the automation of decision-making processes. Here, it is useful to note that Article 22 prohibits decision-making based solely on automated means or profiling, which produces significant effects on the data subjects. An exception to this general rule is provided by Article 22.2, whereby automated decision-making is considered not prohibited if provided by national or EU law.

---

[240] Stock taking of security requirements set by different legal frameworks on Operators of Essential Services and Digital Service Providers: The NISD and the GDPR, ENISA.

[241] According to some scholars, the traditional DPIA approach has a strong focus on data security. See A. Mantelero (2018), "AI and Big Data: A Blueprint for a Human Rights, Social and Ethical Impact Assessment", *Computer Law & Security Review*, Vol. 34, No. 4, 12 May, pp. 754-772.

According to Recital 71 of the GDPR, ensuring the security of data processing could be one of the cases where EU or domestic law may authorise automation: "…*decision-making based on such processing, including profiling, should be allowed where expressly authorised by Union or member state law to which the controller is subject, including for fraud and tax-evasion monitoring and prevention purposes (…) and to ensure the security and reliability of a service provided by the controller"* as long as the legislation lays down appropriate data subject safeguards.[242]

It follows from the interpretation of Article 22 GDPR made through the reading of Recital 71,[243] that under this exception a national or European law may regulate the use of automated decision-making deployed for the purposes of ensuring the security of a service. Based on a literal interpretation of the sources analysed above, this would be the case for both cybersecurity services using AI and cybersecurity services for AI deployment. Nonetheless, specific guidance in this respect has not yet been consolidated. Furthermore, what remains unclear is the level of granularity that laws enabling the Article 22.2 GDPR exceptions should have, particularly with respect to the limitations set on controllers, and the relationships such frameworks have against European and domestic cybersecurity laws (e.g., NIS Directive and national transpositions).

It should be remembered that one of the most important data protection principles is the so-called purpose limitation concept. Under this requirement, enshrined in Article 5.1(b) GDPR, personal data should be "*collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes*". The purpose limitation principle could be divided into two sub-categories, namely purpose specification (data to be collected for specified, legitimate and explicit purposes) and compatibility of use (further processing of data must respect the compatibility between the intended purpose for collection and the first processing goal). In an era of extreme (and ever increasing) interconnectedness, where big data and IoT represent respectively two baseline technologies upon which AI is and will be deployed, many argue that the principle of purpose limitation is becoming harder and harder to meet; personal data is no longer a by-product of a service, but the service has become a by-product of the collection of personal data.[244]

---

[242] GDPR, Recital 71: "*Suitable safeguards include the rights for the data subject to exercise a number of actions against the processing of their personal data: access, rectification, suspension, erasure and explanation."* The latter has been long debated in doctrine and partially addressed elsewhere in this report. For an account of the academic discussion on explainability in AI, see for instance Goodman, B. and S. Flaxman (2016), "EU Regulations on Algorithmic Decision-Making and a "Right to Explanation", *ICML workshop on human interpretability in machine learning (WHI 2016)*, New York, NY (http://arxiv. org/abs/1606.08813 v1) and Wachter, S., B. Mittelstadt and L. Floridi (2017), "Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation", *International Data Privacy Law.*

[243] And confirmed by the guidance provided by Article 29 Working Party (the main EU privacy regulators forum, now called European Data Protection Board – EDPB).

[244] Moerel and Prins, by emphasising that the modern objective of the purpose limitation should be looking at the legitimate interest rather than the initial collection purposes, sum up this fallacy in these terms: "*In today's data-driven society, however, the purpose limitation test has become outdated as a separate test. Imagine a mobile app that on a real-time basis records our health data, predicting how we'll feel the next day and where to avoid getting*

Academics and others are long debating about the current shortcomings of the purpose limitation principle in an AI context. Moreover, various stakeholders seek concrete guidance from the EDPB on the implementation of the purpose limitation principle in AI-driven businesses, while the use of regulatory sandboxes is suggested as a useful approach to test the purpose limitation principle in innovative areas.[245] Nonetheless, a useful characteristic that should be looked at is the declaratory nature of this principle. In a nutshell, by means of the purpose limitation principle, controllers are forced to declare upfront the exact purposes personal data will be processed for. In an AI context, this would mean that the specific objectives shall be made clear prior to the training or development, and the assessment be reconducted should the system produce unexpected results.[246]

Notwithstanding the aptness of this principle for the future of data protection, in the context of the security risk assessments suggested above, and particularly in the suitability testing exercise, these characteristics could be thought of as offering a particularly relevant blueprint. Indeed, the declaratory nature of the purpose limitation could, if re-adapted to a whole AI system architecture (rather than just the use of data therein), serve as a mechanism to understand the optimisation of the AI system (i.e., measure its results against the initial goals),[247] and by consequence, the exercise of comparing the purpose for AI implementation with the related (certain or uncertain) security risk or vulnerabilities. This would help auditability of the AI application, given that the purposes for implementation had been previously declared.[248]

---

the flu. Perhaps pushing the bounds of creepy, this app, however, may be of great value for the World Health Organization (WHO) to protect civilians from life-threatening infectious diseases. These two apps collect and use data for the same purpose, namely mapping and predicting health and illness, but our assessment of the two apps is totally different. The commercial application will not automatically gain societal acceptance, while most of us would see the value of the second application. Whether personal data may be collected and used by such a mobile app is not so much based on the purpose for which the data is collected and processed but on the interest that is served.", L. Moerel and C. Prins (2015), "On the Death of Purpose Limitation", *Intl. Association of Privacy Professionals* (https://iapp.org/news/a/on-the-death-of-purpose-limitation/).

[245] Multistakeholder Expert Group to support the application of Regulation (EU) 2016/679 (2019), "Contribution from the multistakeholder expert group to the stock-taking exercise of June 2019 on one year of GDPR application", 13 June (https://ec.europa.eu/info/sites/info/files/report_from_multistakeholder_expert_group_on_gdpr_application.pdf).

[246] Centre for Information Policy Leadership (CIPL) (2020), "Artificial Intelligence and Data Protection - How the GDPR Regulates AI", March (www.huntonprivacyblog.com/wp-content/uploads/sites/28/2020/03/cipl-hunton_andrews_kurth_legal_note_-_how_gdpr_regulates_ai__12_march_2020_-1.pdf).

[247] On the debate optimisation vis-à-vis *explainability*, see D. Weinberger (2018), "Optimization over Explanation - Maximizing the benefits of machine learning without sacrificing its intelligence", Berkman Klein Center, Medium.

[248] This idea was also suggested by the author in an abstract submitted to UNICRI on purpose limitation and AI: I. Emanuilov, S. Fantin, T. Marquenie and P.Vogiatzoglou (2020), "Purpose Limitation For AI: Mitigating Adverse Impact On Human Rights And Cybersecurity. Abstract for UNICRI Special Collection on Artificial Intelligence".

…the GDPR provides that the data subjects shall have the right to obtain confirmation regarding the existence of automated decision-making, including profiling, and to acquire meaningful information about the logic involved, as well as the significance and the envisaged consequences of this automated decision-making activity on themselves.

The GDPR also introduces a set of rights that relate to the explainability of AI systems. More precisely, the GDPR provides[249] that the data subjects shall have the right to obtain confirmation regarding the existence of automated decision-making, including profiling, and to acquire meaningful information about the logic involved, as well as the significance and the envisaged consequences of this automated decision-making activity on themselves. Article 22 GDPR also includes provisions that support data subjects so that they can vindicate their rights and hold controllers accountable for the processing of their personal data.

As mentioned, Article 22(1) GDPR contains a general prohibition on fully automated decision-making, while Article 22(2) lays down several exceptions to this prohibition. When one of these exceptions applies, Article 22(3) provides that the data controllers shall implement measures to safeguard data subjects' rights and freedoms. Nevertheless, as noted above, decision-making based solely on automated processing, including profiling, is allowed when Union or member state law authorises its use to ensure security and reliability.[250]

It is important to mention that some categories of AI systems appear to have limited capacity to provide the 'reasoning' principles behind an automated decision, mainly because the logic is automatically inferred from vast amounts of data and is embedded in complex mathematical structures that could be considerably opaque for humans. While AI systems are not black box, and even complicated ML models can be regarded just as very complicated statistical functions, the opaqueness of the latest generation of AI systems nonetheless raises the issue of how such an explainability of AI systems could actually be implemented in practice. As the European Commission's Joint Research Centre underlines, the explainability of AI systems plays a key role in their auditing, which could enhance the proposed safeguards under the requirements of Article 22 GDPR. In concrete terms, the audit of AI systems could help the data controller to ensure that such systems are robust (i.e. unbiased and resilient against edge cases and malicious input), as well as to demonstrate compliance with the GDPR requirements. Also, it is suggested that such audit measures could be provided by a third party and possibly evolve into certification mechanisms. Thus, it is understood[251] that the explainability of models is not only

[249] Articles 13(2)(f), 14(2)(g) and 15(1)(h).

[250] *Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679, Article* 29 Data Protection Working Party, 2017.

[251] R. Hamon, H. Junklewitz and I. Sanchez (2020), Robustness and Explainability of Artificial Intelligence - From technical to policy solutions, EUR 30040, Publications Office of the European Union, Luxembourg.

a key point for their transparency, but also a concept of paramount importance in assessing the reliability of a model and its exposure to failures.[252]

As noted, many cybersecurity firms and professionals make use of personal data processing techniques in their daily activities. According to the observations made in the Task Force meetings, the implementation and use of AI in their practices will not change this trend (and possibly even increase it), as AI-based tools seem to play a vital role in the scalability and efficiency of existing cybersecurity tools, for instance in the area of network security. In this specific sector, the GDPR has been interpreted as having a rather broad concept of personal data, and the same holds with respect to the evolution of the current jurisprudence of the CJEU.[253]

For such reasons, cybersecurity professionals have long struggled to find an adequate legal basis for conducting their daily practices without being found in breach of the GDPR. It is often the case that the collection of personal data cannot rely on the consent or the performance of a contract executed with the data subject, and so other legal grounds ex Article 6 GDPR have to be sought. In this respect, Recital 49 acts as a useful interpretative aid, bringing forward the idea that cybersecurity practices could rely on the controllers' legitimate interest.[254] While this specification is to be welcomed as possibly providing the cybersecurity industry some room to continue operating in respect of the legal grounds ex Article 6 GDPR, questions remain on

[252] Several data protection authorities in Europe have issued guidelines that set out good practices for explaining decisions to individuals that have been made using AI systems. For example, in the context of the project "ExplAIn", the UK Information Commissioner's Office (ICO) and the Alan Turing Institute offer some practical guidance on this matter. More precisely, it is suggested that explanation on the safety and performance indicators of AI systems helps data subjects to understand the measures that data controllers have put in place to maximise the accuracy, reliability, security, and robustness of the decisions that are generated from such AI models. On the one hand, such explanations could be process-based, by providing information on the measures taken to ensure the overall safety and technical performance (security, accuracy, reliability, and robustness) of the AI model – including information about the testing, verification, and validation done to certify these measures. On the other hand, the explanation could be also outcome-based, meaning that it revolves around information on the safety and technical performance (security, accuracy, reliability, and robustness) of the AI model in its actual operation, e.g. information confirming that the model operated securely and according to its intended design in the specific data subject's case. When it comes to audits and testing of AI systems, the Council of Europe provides a set of useful guidelines as well. Specifically, with its latest recommendations, the Council of Europe suggests that regular testing and auditing against state-of-the-art standards related to privacy, data protection and security before, during and after production and deployment of AI systems should form an integral part of testing efforts, particularly where such AI systems are used in live environments and produce real-time effects.

[253] It is exemplary in the case of IP addresses, whereby, at times and under certain circumstances, even dynamic IP addresses are considered as personal data. See CJEU, Case C-582/14 *Patrick Breyer v Bundesrepublik Deutschland*, ECLI:EU:C:2016:779.

[254] The processing of personal data to the extent strictly necessary and proportionate for the purposes of ensuring network and information security, i.e. the ability of a network or an information system to resist, at a given level of confidence, accidental events or unlawful or malicious actions that compromise the availability, authenticity, integrity and confidentiality of stored or transmitted personal data, and the security of the related services offered by, or accessible via, those networks and systems, by public authorities, by computer emergency response teams (CERTs), computer security incident response teams (CSIRTs), by providers of electronic communications networks and services and by providers of security technologies and services, constitutes a legitimate interest of the data controller concerned. This could include, for example, preventing unauthorised access to electronic communications networks and malicious code distribution and stopping 'denial of service' attacks and damage to computer and electronic communication systems.

several points. For instance, and given that AI could potentially increase the amount of (personal) data processed for security purposes, such a use could prove to add some value in terms of the efficiency of such tools. In this respect, therefore, and looking at Articles 6 GDPR and its Recital 49, the 'extent strictly necessary and proportionate' which, seemingly, the GDPR would allow in order to rely on the legitimate interest ground for processing is unclear.

Another important piece of legislation of the EU personal data protection framework is the ePrivacy Directive, which is currently under negotiations for revision (the proposal for the ePrivacy Regulation).[255] The revised provisions aim to address new risks to the protection of personal data and private life in the context of electronic communications in the era of big data analytics and the IoT. The proposed text is still being discussed by the Council, after which it will enter the trialogue negotiations. Thus, the exact scope and meaning of some core notions are yet to be delineated. Nevertheless, the reform of the ePrivacy framework is necessary to deliver effective confidentiality and security of modern online communications, to ensure clarity of the legal framework, and to enhance public trust in the EU digital economy.

Finally, having examined the EU personal data protection framework in detail, it is also important to highlight the existing EU legal provisions on the processing of non-personal data. More precisely, in many real-life situations in the field of cybersecurity, the datasets used are likely to be composed of both personal and non-personal data. Such datasets are often referred to as 'mixed datasets'. At the same time, the rapid development of emerging technologies such as AI, IoT, technologies enabling big data analytics and 5G, is raising questions about the access and reuse of such datasets. One could argue that this should not create confusion for the stakeholders involved, since there are no contradictory obligations under the GDPR and the Regulation on the Free Flow of Non-personal Data (FFD Regulation).[256] However, the personal data protection framework is much stricter than the non-personal data framework, while the boundaries between non-personal data and personal data are too fluid to act as a regulatory anchor. In this way, datasets that could fall under the notion of personal data might be treated as non-personal data by the stakeholders involved in their processing activities. Furthermore, even non-personal data could be used to attribute identity or other personal characteristics. Thus, two separate regimes applicable to opaque datasets might lead to challenges related to the adequate enforcement of personal data protection rules.[257]

[255] Council of the European Union (2020), Proposal for a Regulation of the European Parliament and of the Council concerning the respect for private life and the protection of personal data in electronic communications and repealing Directive 2002/58/EC (Regulation on Privacy and Electronic Communications), 6 March, Brussels (www.parlament.gv.at/PAKT/EU/XXVII/EU/01/51/EU_15125/imfname_10966469.pdf).

[256] European Parliament and The Council of the European Union (2018), Regulation (EU) 2018/1807 of the European Parliament and of the Council of 14 November 2018 on a Framework for the free flow of non-personal data in the European Union, 28 November, Brussels.

[257] I. Graef, R. Gellert and M. Husovec (2018), "Towards a Holistic Regulatory Approach for the European Data Economy: Why the Illusive Notion of Non-Personal Data is Counterproductive to Data Innovation", *TILEC Discussion Paper No. 2018-029*, September.

The GDPR offers a modern baseline for the protection of personal data in an AI and cybersecurity context. But the GDPR should not be regarded as resolving all security issues arising from AI. Nor should AI be a driver for demanding GDPR amendments, considering, inter alia, that the law was drafted on the basis of the technological neutrality principle. This Task Force showed how the GDPR lays down a number of provisions that might require further study and regulatory interpretation, even at national level (for instance, with respect to Recitals 49 and 71, or with reference to data-sharing practices for information security aims), in the context of both AI for cybersecurity and applications aimed at securing AI. Nonetheless, the GDPR confirms that this is representative of a data protection framework whose principles are arguably and potentially relevant for broader information security AI regulatory schemes (for instance, requiring a delimitation of purpose of AI deployment to serve broad information security accountability).

5. Develop and deploy reliable AI[258]

As previously noted in this chapter and elsewhere in this report, it would be more appropriate for nascent standards and certification methods for AI in cybersecurity to focus on supporting the *reliability* of AI, rather than its trustworthiness. As highlighted in Chapter 2, such a distinction carries both conceptual and operational differences. In this context, the following three requirements can, among other things, be beneficial from a policy perspective to mitigate the vulnerabilities of AI systems and improve their reliability:

1. *Companies' in-house development*. Attackers of AI systems can exploit the use of commercial services that support the development and training of AI systems, such as the cloud. Breaches in the cloud can indeed provide them with easy access to the model and to the training data. It is suggested, especially in the context of applications in national critical infrastructure, that it is best to rely on trusted suppliers' design, to develop the model in-house and to have the system providers directly collect, curate, and validate the testing dataset and secure its storage.
2. *Adversarial training*. This could help to improve the robustness of the AI system and to identify vulnerabilities, given that using feedback loops can improve AI performance. Taddeo et al., however, underline that the effectiveness of adversarial training is conditioned on the refinement of the adversarial mode. Appropriate metrics for level of refinement, corresponding with the expectation on the outcome of that kind of system, should thus be mandated together with adversarial training.
3. *Parallel and dynamic monitoring or punctual checks*. The deceptive nature of AI attacks and the limits in assessing AI robustness *"require form of constant (not merely regular, i.e., at time intervals, but continuous, twenty-four hours a day, seven days a week) monitoring during deployment."*[259] It is suggested that a clone system be created to be deployed in a controlled environment, to effectively carry out such constant monitoring. Such a clone system would serve as a benchmark to assess that the model deployed in the external environment is behaving as it is expected to, thus allowing any divergence from it to be promptly captured and addressed.

[258] This section of the report is drawn from M. Taddeo, T. McCutcheon and L. Floridi (2019), op. cit.

[259] M. Taddeo, T. Cutcheon and L. Floridi (2019), *op. cit.*

6. The role of AI standards activity and cybersecurity[260]

Standards are one tool for industry to implement or demonstrate adherence to policy and regulatory requirements, along with voluntary labelling, operational guidelines, codes of conduct, and open-source software. Standards activities must therefore take place in the broader context of law and regulation and through transparent processes. AI systems are already governed by many national, regional, and sector-specific laws and regulations, such as GDPR requirements. With respect to the cybersecurity of AI, the new guidance being considered globally intends to address new security concerns characteristic of AI systems and their operation.

Standards generally fall into three categories: foundational, technical interoperability, and management. Foundational standards establish globally shared basic concepts to generate a common understanding for AI policy and practices, and may define terms, use cases, and reference architectures. Technical interoperability standards establish mechanisms such as protocols to enable disparate systems to communicate. Because of the rapid innovation in AI technologies, interoperability solutions, at least in the short term, may gravitate to open-source collaborations versus standards. To outline the standards activity influencing the AI and cybersecurity relationship, the final category – management – is most relevant.

Management standards establish governance guidelines, form the criteria for responsible behaviour, and enable organisations to demonstrate conformity to best practices and regulation. At present, there is no clear answer as to how many new standards of any category will create cybersecurity requirements for AI systems, or whether and how ongoing projects will point to existing cybersecurity standards. AI management standards activity establishing requirements for organisational governance and system robustness could prove especially significant given the risk of cyberattack escalation, potential loss of control, anomaly detection and monitoring challenges, and the large attack surface of the digital environment. The understanding of AI technology and its risks will evolve in parallel with attack techniques.

To expand on how they will affect AI and cybersecurity and the global standards development taking place, management standards can be broken down into further categories. The efforts of several standards are highlighted within each category.

A) Establishment of organisational governance and principles, including definition of risk criteria and appetite and identification of concerns to stakeholders as related to stakeholder assets and values

AI cybersecurity interacts with cybersecurity governance architecture, and various standards bodies acknowledge pre-existing governance standards. Defining the new risks presented by AI systems and the risk criteria for different applications and sectors will enable organisations to adapt their risk management frameworks to include AI. For example, organisational governance standards can determine the functions to assess risk levels of system accuracy and robustness.

[260] This section of the report has been contributed by the Microsoft team participating in the CEPS Task Force.

- In the US, Executive Order 13859 called on several bodies such as the National Institute for Standards and Technology (NIST) and the White House Office of Management and Budget (OMB) to direct how the US government guides both the development and deployment of AI through technical standards. In its planning document for AI standards, NIST references existing standards that already apply to AI governance and risk management, such as the International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC) standard on Information Technology – Governance of IT for the organisation (ISO/IEC 38500:2015), while acknowledging that existing standards need to be reviewed and potentially augmented for new AI considerations.
- The EU Commission's White Paper on AI[261] suggests that a European governance structure for AI should leverage and complement existing governance and risk-assessment tools, for example through the ENISA.
- The National Information Security Standardization Technical Committee (TC260) in China named *responsibility* as a part of its principle of AI security in its White Paper on Standardization of AI Security.[262] It elaborated that AI security standards and policy will "set up the mechanism" of AI responsibility, and identify what entities might be responsible for AI system performance and how they will be audited. The China Academy of Information and Communication Technology (CAICT) is a coalition of over 200 companies that collaborates with the government, representing an industry perspective on the development of AI applications for a variety of uses including cybersecurity, and input into AI security considerations. Its White Paper on AI Security focuses heavily on cybersecurity.[263] The paper highlights AI cybersecurity as well as different types of AI applications to cybersecurity and the particular security risks resulting from these uses through a proposed security framework.

B) Risk assessment of control objectives (i.e., threats and vulnerabilities) and risk treatment or mitigation through controls

Once an organisation has determined its risk appetite, standards can establish a process for risk assessment and management. Cybersecurity of AI systems can appear as a part of risk assessment standards, or the focus can be on ensuring system robustness and resilience to threats including cyberattacks.

- The ISO/IEC are taking an *ecosystem* approach to AI. Given that AI applications will operate in cyber, physical, and hybrid environments, the ecosystem approach may be used to support the cybersecurity of AI, with tools to evaluate how robust systems are to

[261] European Commission, White Paper On Artificial Intelligence - A European approach to excellence and trust, COM(2020) 65 final, Brussels, 19 February 2020.

[262] See National Information Security (2019),《人工智能安全标准化白皮书（2019版》发布 (Artificial Intelligence Security Standardization White Paper (2019 Edition), (www.tc260.org.cn/front/postDetail.html?id=20191031151659).

[263] See CAICT (2018) 人工智能安全白皮书(2018), (Artificial Intelligence Security White Paper (2018) (www.caict.ac.cn/kxyj/qwfb/bps/201809/t20180918_185339.htm).

threats. The ongoing ISO/IEC NP TR 24029-1 – Assessment of the robustness of neural networks is an example of an activity that may support the security of AI to adversarial input and the use of AI in adversarial environments such as cybersecurity. The ISO has also stated that cybersecurity is a key threat to trustworthiness.[264]

- The EU Commission White Paper on AI suggests that policymakers might consider robustness-related standards that could mitigate cyberattacks, including *"requirements ensuring that AI systems are resilient against both overt attacks and more subtle attempts to manipulate data or algorithms themselves, and that mitigating measures are taken in such cases."*[265]

- The TC260 in China identified 'classification' as a part of its principles of AI security in its White Paper on Standardization of AI Security, linking the security requirements of the AI application to potential future classification criteria that may be linked to application, sector, risk, or a combination of these features. This approach indicates that China will develop AI management standards and that these will be linked to cybersecurity considerations and robustness.

C) Implementation guidelines

Implementation guidelines are tools to assist organisations apply security standards to their AI development life cycle, provide common terminology for implementation, and to evaluate implementation progress. These guidelines complement organisational governance. Examples of implementation guideline activity for AI and security include:

- The OMB may require US federal agencies to submit plans to *"achieve consistency"* with their guidance for regulation of AI applications once published.[266]
- The EU Commission White Paper proposes that a European governance structure for AI should issue guidance on implementation of AI regulation, and that stakeholders from the private sector, academia, and civil society should be consulted in developing this guidance.
- The CAICT White Paper on AI Security indicates that guidelines for risk management will be related to existing cybersecurity requirements and that security assessment will be part of future policy.

[264] C. Naden (2019) "It's All About Trust", ISO, 11 November (www.iso.org/news/ref2452.html).

[265] European Commission, White Paper On Artificial Intelligence - A European approach to excellence and trust, COM(2020) 65 final, Brussels, 19 February 2020.

[266] Office of Management and Budget (2020), Draft Memorandum for the Heads of Executive Departments and Agencies: Guidance for Regulation of Artificial Intelligence Applications (www.whitehouse.gov/wp-content/uploads/2020/01/Draft-OMB-Memo-on-Regulation-of-AI-1-7-19.pdf?utm_source=morning_brew).

D) Evaluation of characteristics using qualitative or other metrics

Key measures to ensure the security and reliability of AI systems include evaluating their robustness and resilience. Even if an AI system's operating environment or data has been compromised through a cyberattack, system robustness and resilience can minimise the negative impact on performance and are part of risk-mitigation measures. Robustness and resilience to adversarial input can improve reliability and safety, a major ethical concern of AI systems. While standardised metrics for robustness, resilience, and security are largely under development, standards activity reflects the need for tools and resources to evaluate AI systems.

- Proposed system evaluation standards, such as the ISO/IEC NP TR 24029-1 – Assessment of the robustness of neural networks, will give organisations methods to evaluate their AI systems against international benchmarks.
- The NIST AI standards plan names the need for metrics of robustness, resilience, and security as components of trustworthy AI.
- The TC260 White Paper on Standardization of AI Security gives a framework for standards development for, inter alia, basic security, data, algorithms, models, the surrounding system, and testing and evaluation.
- While not a public standards activity, the Guaranteeing AI Robustness against Deception (GARD) research program of the US DARPA is currently working on developing theoretical foundations for general defences against adversarial machine learning, which will be used to create metrics for robustness. The program considers the many different types of data and exposure that AI systems will have in a battlefield context, which could include robustness and resilience to adversarial examples introduced from cyber and physical domains or a hybrid of the two. Their national security implications means that these findings may not make it into the public domain, but other governments may have similar programs.

These early stages of standards development for AI cybersecurity are critical for ensuring that standards integrate into existing security frameworks and are compatible for diverse systems across the globe. Cybersecurity factors into the risk of an AI application, which will translate into the risk-based standards thresholds for different uses and sectors. Standardisation will be necessary for ensuring consistent global approaches to addressing risk, building and managing the technology, and ensuring the cybersecurity of AI.

In this context, companies at the forefront of standards implementation can support standardisation efforts by sharing practical experiences – a prerequisite for developing advanced technical standards – and by investing in research and development of tools for new technologies. More specifically, companies could develop and share:

- Business understanding tools, including guidelines and best practice to support customer and partner decision-making processes.
- Data acquisition and understanding tools addressing common challenges related to training data and ML models.
- Modelling tools addressing AI systems and ML intelligibility and exploitability.

For their part, policymakers could consider the following actions for improving AI and cybersecurity with respect to standards:

- Clarify how existing cybersecurity policies and frameworks apply to AI, including through regulations, guidance, standards, or best practices.
- Identify priority areas for clearer cybersecurity guidance, for example in sensitive use cases and applications where the potential impact of failures is significant.
- Excluding sensitive use cases and areas where new AI-oriented policies become mature, and consider voluntary consensus standards as a non-regulatory approach to addressing AI cybersecurity risks.
- Promote research, evidence, information sharing and engagement to inform best practice and standards-making activities for AI cybersecurity.
- Encourage AI practitioners to engage in developing cybersecurity best practice and standards, including adopting evolving guidance into their product development and support life cycles to better protect customers and improve their ability to contribute to standards.

Hence, a multistakeholder approach could be based on established best practice. For example, to improve collaboration among companies, the European Commission, in the context of the research programme Horizon 2020, is promoting projects such as Sharing and Automation for Privacy Preserving Attack Neutralization (SAPPAN).[267] SAPPAN aims to develop a platform to enable a privacy-preserving efficient response and recovery plan that utilises advanced data analysis and ML. It also aims to provide a cyber-threat intelligence system that decreases the effort required by a security analyst to find responses and ways to recover from an attack. Such an initiative can be understood in the context of the more general effort of the EU to foster data sharing, considered a prerequisite for the establishment of a European AI ecosystem.[268]

[267] For more on this see Sharing and Automation for Privacy Preserving Attack Neutralization (SAPPAN) (https://sappan-project.eu).

[268] In this respect, it can be noted that on 19 February 2020, the EU Commission published its 'EU Data Strategy', aiming to create a Common European Data Spaces organised in verticals such as Industrial Manufacturing, Health, Energy, Mobility, Finance, Agriculture and Science. An industrial package to further stimulate data sharing was also established in March 2020. In addition, the EU Commission has appointed an Expert Group to advise on Business-to-Government Data Sharing and the High-Level Expert Group on Artificial Intelligence devoted an entire section to fostering a European data economy, including data-sharing recommendations, data infrastructure and data trusts. Finally, the Ad Hoc Committee on Artificial Intelligence, established by the Council of Europe, is currently examining the possibility of a binding legal framework for the development, design and application of AI and data based on the Council's standards on human rights, democracy, and the rule of law.

7. Additional policy issues

As examined throughout the chapter, several policy issues stand at the intersection of AI and cybersecurity. Given the extensive and multifaceted nature of such an intersection, any mapping exercise, while pivotal for the understanding of the subject, is likely to be non-exhaustive in terms of all the aspects involved. In this context, the following additional issues are presented as important facets informing the subject of this report. However, the analysis of the following would require further enquiry, especially because of the number of open questions that these issues leave, at the state-of-the-art, still to be answered.

7.1 Dual use and export control

This report has frequently mentioned cases where AI could be used for malicious as well as benevolent purposes. What we called 'weaponisation' of AI will further blur the line between what are two sides of the same coin. It is the belief of many that several applications designed and developed by commercial actors could be used for malicious goals, or even turned into military applications, and vice versa.[269] Nonetheless, it would be wrong to believe that such a contraposition depicts the full picture; as some argue, AI is, to its fullest extent, a general-purpose technology.[270]

It would also be wrong to believe that such a transformative process (general purpose to malicious purpose) would follow the same pace and speed for all use cases. On the contrary, the repurposing of commercial AI technologies for cybersecurity could have different levels of pace against their adaptiveness for malicious uses.[271] As Brundage et al explain: "*Many tasks that it would be beneficial to automate are themselves dual-use. For example, systems that examine software for vulnerabilities have both offensive and defensive applications, and the difference between the capabilities of an autonomous drone used to deliver packages and the capabilities of an autonomous drone used to deliver explosives need not be very great. In addition, foundational research that aims to increase our understanding of AI, its capabilities and our degree of control over it, appears to be inherently dual-use in nature.*"[272]

To this end, and going a step further from the broad societal considerations, a more policy-oriented observation pertains to the impact of AI for the current EU dual-use and export control legislation. Export control is a particularly critical area, where many (and often diverging) interests collide. From an operational viewpoint, it often originates for producers' or traders'

---

[269] D. Azulay (2020), *Weaponized Artificial Intelligence – Critical Dual-Use Applications*, Emerj, 20 February (https://emerj.com/ai-future-outlook/weaponized-artificial-intelligence/).

[270] US National Security Commission on Artificial Intelligence (2019), Final Report (www.nscai.gov).

[271] "*A wide variance exists in the ease of adaptability of commercial AI technology for military purposes. In some cases, the transition is relatively seamless. For example, the aforementioned aircraft maintenance algorithms, many of which were initially developed by the commercial sector, will likely require only minor data adjustments to account for differences between aircraft types. In other circumstances, significant adjustments are required due to the differences between the structured civilian environments for which the technology was initially developed and more complex combat environments*", K. M. Slayer (2020), op. cit.

[272] M. Brundage et. al. (2018), op. cit., p. 16.

licensing obligations to be fulfilled prior to the sale. The dual-use export control framework governs the export, transit and brokering of goods and technologies (including software) that can be used for civil and military applications. At the EU level, it is formed by European Commission Regulation No 428/2009.[273] Since 2014, this Regulation has been under revision so as to take into account the evolution of technologies, and a Commission Proposal has been under discussion since 2018.[274] Both the EU and the Wassenaar Agreement community[275] have tried to cope with the fast-paced technological developments of the past decade by including in their respective lists (or planning to do so), tools enabling digital forensics or software intrusions.[276]

Over the past few years, the current export control scheme deriving from both the Wassenaar Agreement as well as the EU framework has been criticised from many sides, particularly with regards to the apparent failure to enforce these rules in the area of cybersecurity. First and foremost, human rights-based criticisms pointed out several scandals about the sale by EU-based companies of surveillance systems to authoritarian regimes. This provided an example of the inefficiency of the regulatory framework in enabling appropriate monitoring to prevent the transfer of certain tools to countries with a questionable approach to the rule of law.[277] But many others feared that the current system is counterproductive to the race to technological empowerment, which will eventually shape the new geopolitical dynamics.[278]

---

[273] International multilateral agreements on export control mostly refer to the so-called Wassenaar Agreement

[274] See European Commission, Proposal for a Regulation of the European Parliament and of the Council setting up a Union regime for the control of exports, transfer, brokering, technical assistance and transit of dual-use items (Recast) {Swd(2016) 314 Final} - {Swd(2016) 315 Final}, 28 September 2016. See also European Parliament and Council, Report on the Proposal for a Regulation of the European Parliament and of the Council setting up a Union regime for the control of exports, transfer, brokering, technical assistance and transit of dual-use items (Recast) (COM(2016)0616 – C8-0393/2016 – 2016/0295(COD)), 19 December 2018 and European Commission, Commission Delegated Regulation (EU) 2017/2268 of 26 September 2017 amending Council Regulation (EC) No 428/2009 setting up a Community regime for the control of exports, transfer, brokering and transit of dual-use items. C/2017/6321, 15 December 2017.

[275] The Wassenaar Agreement is a multilateral convention on dual use at the international level.

[276] "*In 2012 and 2013 members of the Wassenaar Arrangement added mobile telecommunications interception equipment, intrusion software, and internet protocol (IP) network surveillance to the organisation's list of controlled dual-use items. In December 2019 controls on monitoring centres and digital forensics were also added after several years of debate and discussion. The EU has included the 2012-2013 items in its own dual-use list and will add monitoring centres and digital forensics the next time it is updated in late 2020. Moreover, it has used EU sanctions to prohibit exports of a wide range of surveillance technologies to Iran, Myanmar, Syria, and Venezuela*.", M. Bromley (2020), "A search for common ground: export controls on surveillance technology and the role of the EU", About:Intel European Voice on Surveillance, 12 February (https://aboutintel.eu/surveillance-export-control-eu/).

[277] See for instance, S. Gjerding and L. Skou Andersen (2017), "How European spy technology falls into the wrong hands, *De Correspondent*, 23 February (https://thecorrespondent.com/6257/how-european-spy-technology-falls-into-the-wrong-hands/2168866237604-51234153); B. Wagner (2012), "Exporting Censorship and Surveillance Technology", Hivos People Unlimited, January; Report Without Borders (2012), "Position paper of Reporters without Borders on the export of European surveillance technology", 6 November (https://rsf.org/sites/default/files/2012_11_07_positionspapier_en_eu.pdf); Marietje Schaake – Member of European Parliament, Written submission to the public online consultation on the export control policy review (Regulation (EC) No 428/2009) (https://trade.ec.europa.eu/doclib/docs/2015/november/tradoc_154004.pdf).

[278] J. Leung et al. (2019), "Export Controls In The Age Of AI"; R. C. Thomsen II (2008), "Artificial Intelligence and Export Controls: Conceivable, But Counterproductive?", *Journal of Internet Law, DLA.*

Within the cybersecurity community it has been underlined that the inclusion of technologies such as intrusion software within the list of export control for dual-use items might have an impact on practitioners' information sharing as regards malicious software.[279] In fact, objections have been raised on the current (and developing) framework, given that the "*controls on intrusion software generated criticism from companies and researchers on the grounds that they inadvertently captured processes for reporting software vulnerabilities and tools for testing IT-security.*"[280]

It should be recalled here that the EU finds its foundations in a strong set of values inherently dependent on the rule of law and the human rights' traditions of its member states. For these reasons, export control and dual-use policies should not be regarded as solely addressing questions of trade and geopolitical power dynamics. A balanced approach shall take these values into account, in a human-security perspective.[281]

Nonetheless, it remains unclear how the use of AI systems in cybersecurity research and operations could be impacted by the current export control framework. There is still too little research on this subject, and a more informed and expert debate is required to feed decision makers with reliable and impartial opinions, as well as evidence-based conclusions.

Having said that, it is useful to shed light, as a last consideration, on one potential (indirect) implication that more stringent rules on export control may have on the development of AI systems for cybersecurity. In principle, stricter rules on export control should hold AI research and development more accountable, thus putting limits and barriers on the dissemination and sharing of project results that could become dangerously reproducible for non-accountable actors. Nonetheless, what has proved to be a shortcoming over the years has been the inability to clearly enforce such rules on the basis of the sole dual-use and export control regime, particularly with reference to the export of software technologies.

It should be underlined additionally that, while export controls might help in holding AI research and development accountable, if these efforts are not equally undertaken by other international actors the desired goal is unlikely to be accomplished. In fact, this could even undermine the EU internal market because of the unfair competition of third parties not equally applying export controls. In this respect, it should be noted that China, for example, at the forefront in the development of AI technologies, has not undersigned the Wassenaar Agreement.

With that in mind, an alternative for preventing AI systems ending up in the wrong hands should be envisaged through the experimentation of regulatory solutions by the information security community, setting up boundaries for the openness of AI research.

---

[279] Cristian Barbieri et al. (2018), "Non-proliferation Regime for Cyber Weapons. A Tentative Study", Istituto Affari Internazionali, March.

[280] M. Bromley (2020), op. cit., and N. Martin and T. Willis (2015), "Google, the Wassenaar Arrangement, and vulnerability research", Google Security Blog.

[281] See M. Schaake (2017), "AFET opinion dual-use regulation", 12 April (https://marietjeschaake.eu/en/afet-opinion-dual-use-regulation).

A dual-use technology transfer mechanism, specifically regulating the transfer of technologies from civilian to military applications, could also serve as a possible method to define and regulate the boundaries of a technological space landscape where military and civilian applications of a dual-use technology are present. Notably, if this mechanism is regulated at institutional level, through support of the industry, it could be the backbone of a controlled dual-use technology landscape, within the boundaries fixed by the Wassenaar Agreement. This mechanism does not de facto exist at this moment under a European coordinated framework. However, efforts are being made in this direction by the European Defence Agency (EDA). EDA launched a project called Analysis of Dual Use Synergies (ANDES) that has among its main objectives the definition of a possible dual-use technology transfer mechanism at European level and the creation of an avenue for developing a common approach among institutions dealing with dual-use technologies.[282]

The current (and future) dual-use and export control regulatory framework needs to be evaluated against the impact of AI systems in cybersecurity. Clear rules shall respect fundamental European values without creating hurdles to trade and openness.

7.2 Employment, jobs, and skills

Uncertainty underlies the impact of AI in the cybersecurity job market, and in particular the extent to which it will lead to greater job opportunities or cause job displacement.

The increasing automation of tasks brought about by AI in the cybersecurity field raises questions about skills and employment. A 2019 OECD report on AI argues that the increasingly far-reaching use of AI in digital security[283] has been driven, among others, by the shortage of cybersecurity skills, precisely because it improves productivity and accelerates tasks that can be automated.

Several surveys have found that the current cybersecurity job market is lacking adequate numbers and expertise. In 2014, Cisco estimated a global shortage of information security professionals of at least one million,[284] and the 2019 Cybersecurity Workforce Report forecasts a significant increase on this figure.[285] A CSO survey conducted in both the US and the Europe, Middle East, Africa (EMEA) regions finds that experts are facing s implementing high-level information security because of the shortages of personnel (24%) or of specific cybersecurity expertise (34.5%).[286]

[282] Contribution of the European Defence Agency to the work of the Task Force.

[283] OECD (2019a), op. cit., p. 67.

[284] Cisco (2014), Annual Security Report, Cisco (www.cisco.com/c/dam/assets/global/UK/pdfs/executive_security/sc-01_casr2014_cte_liq_en.pdf).

[285] (ISC)2 (2019), "Strategies for Building and Growing Strong Cybersecurity Teams", (ISC)2 Cybersecurity Workforce Study (www.isc2.org/-/media/ISC2/Research/2019-Cybersecurity-Workforce-Study/ISC2-Cybersecurity-Workforce-Study-2019.ashx).

[286] S. Morgan (2015), "Cybersecurity job market to suffer severe workforce shortage", CSO Online, 28 July (www.csoonline.com/article/2953258/it-careers/cybersecurity-job-market-figures-2015-to-2019-indicate-severe-workforce- shortage.html).

Others believe that the growing use of AI in cybersecurity will create new opportunities for the job market, particularly in positions labelled 'new collars': "*Essentially, AI helps to augment the analyst's daily activities by acting as an assistant. It would quickly research the new malware impacting the phones, identify the characteristics reported by others and provide a recommended remediation.*"[287] Indeed, it appears that a large part of the community believes that the increasing deployment of AI systems in existing information security practices will transform current jobs, rather than boost the unemployment rate.[288] There are examples of AI reducing the barriers to entry into the cybersecurity profession by augmenting and supporting less experienced workers to take on frontline operational roles. According to the executive director of the security operations centre at Texas A&M University, which has used AI to help overcome a cybersecurity skills shortage: *"student workers are effective as the front line of the Security Operations Centre and the selected AI tool enables them to make rapid, accurate decisions on whether to escalate a detected threat for further investigation."*[289]

Nonetheless, figures seem to diverge for the job market as a whole. For example, McKinsey predicts that as many as 400 million jobs will be displaced in the long term across the world because of the increase in automation of existing tasks and their corresponding jobs.[290] Forrester, however, gives a much more cautious figure of fewer than 25 million jobs lost to automation by 2027.[291]

These apparently contradictory figures reveal how difficult it is to accurately predict the impact of automation and AI on the broad job market and, narrowly, in the cybersecurity field.[292] According to the OECD, the impact of AI will depend on the capacity for the diffusion of the technology across different sectors.[293]

What seems to be a common denominator when predicting the future of the cybersecurity labour market is that professionals will have to adapt their skills, and that AI used for information security should not transcend a human presence in its life cycle. The former consideration holds true not only at the operational level, but also at the most senior and managerial levels of both private and public sectors. Initiatives aiming to close the knowledge gaps have already started in certain parts of the world. For instance, the Japanese Ministry of Economy, Trade and Industry (METI), via its Industrial Cyber Security Center of Excellence, is

---

[287] C. Barlow (2015), "Perspective: Artificial intelligence makes cybersecurity the ideal field for 'new collar' jobs", Duke Political Science Blog.

[288] M. Dsouza (2018), "How will AI impact job roles in Cybersecurity", Packt, 25 September (https://hub.packtpub.com/how-will-ai-impact-job-roles-in-cybersecurity/).

[289] Vectra (2020), Case Study: An academic and research powerhouse (https://content.vectra.ai/rs/748-MCE-447/images/CaseStudy_TexasAM.pdf).

[290] J. Manyika et al. (2017), "Jobs lost, jobs gained: What the future of work will mean for jobs, skills, and wages", McKinsey & Company, 28 November (www.mckinsey.com/featured-insights/future-of-work/jobs-lost-jobs-gained-what-the-future-of-work-will-mean-for-jobs-skills-and-wages#).

[291] E. Winick (2018), "Every study we could find on what automation will do to jobs, in one chart", *MIT Technology Review*.

[292] M. C. Horowitz et al. (2018), op. cit., p. 14.

[293] OECD (2019a), op. cit., pp. 106-107.

funding short- and long-term (one year) training to C-level professionals on cybersecurity.[294] Such initiatives not only have the long-term goal of creating a high-level labour force with cybersecurity knowledge, but also the short-term objective of training executives to be able to make conscious and informed decisions during a cybersecurity crisis.[295]

The Japanese example of knowledge transfer and capacity building prompts a further observation on a less-reported impact of AI in cybersecurity. Beyond the job displacement and skills issues lies the question of whether job numbers and expertise will be evenly distributed among players of different kinds. There is indeed a general expectation that in the long run, it is the big technology players that will attract more and more expert professionals at the expenses of SMEs and public institutions. A recent report from ENISA on the cybersecurity market for start-ups highlights how, one of the main challenges for SMEs operating in the cybersecurity field is the shortage of educational paths and lack of resources leading to employment scalability and therefore a lack of recruitable expert personnel. This situation is compounded by many valuable professionals being more attracted by the opportunities offered beyond the territories of the European Union.[296]

It is not just the SME sector that the personnel shortage affects. Governments and public institutions alike seem to have woken up to the fact that large technology companies are better at attracting promising talented cybersecurity and AI professionals, and not only from a financial perspective.[297] A 2019 report by the US Congressional Research Service posits numerous reasons for this preference: "[reports suggest that] *challenges when it comes to recruiting and retaining personnel with expertise in AI* [is] *due to research funding and salaries that significantly lag behind those of commercial companies. Other reports suggest that such challenges stem from quality-of-life factors, as well as from a belief among many technology workers that 'they can achieve large-scale change faster and better outside the government than within it.*"[298]

The deskilling of the workforce is a final aspect of the adoption of AI that needs to be analysed. On the one hand, as underlined in Chapter 2, AI systems used to support cybersecurity allow vulnerabilities, malware, and anomalous behaviours to be identified more effectively than a security analyst would be able to. On the other hand, a complete delegation of threat detection to AI systems might lead to widespread deskilling of cybersecurity experts. In this scenario,

---

[294] Information-Technology Promotion Japan, Industrial Cyber Security Center of Excellence.

[295] See EUNITY (2019), Minutes from the 3rd EUNITY Project Workshop, Kyoto (www.eunity-project.eu/en/workshops/3rd-eunity-workshop/).

[296] "*65% of the NIS start-ups panel do not find it easy to attract the best talents for their business. High cost (53%) and lack of suitable skills (47%) are cited as the most important factors*." ENISA (2019), "Challenges And Opportunities For Eu Cybersecurity Start-Ups", 15 May.

[297] C. Cordell (2019), "Talent and data top DOD's challenges for AI, chief data officer says", FedScoop.

[298] K. M. Slayer (2020), op. cit., p. 18, quoting Mary Cummings, "*Artificial Intelligence and the Future of Warfare*, Chatham House, 2017: "*the global defense industry is falling behind its commercial counterparts in terms of technology innovation, with the gap only widening as the best and brightest engineers move to the commercial sphere*", and A. Zegart and K. Childs (2018), "The Divide between Silicon Valley and Washington Is a National-Security Threat", *The Atlantic*, 13 December (https://medium.com/the-atlantic/the-divide-between-silicon-valley-and-washington-is-a-national-security-threat-4bf28276fca2).

without appropriate countermeasures in place, the whole organisation and work chain could be left vulnerable if the AI system failed.

Skills shortages and uneven distribution of talents and professionals among market players are two aspects of the impact of AI on the cybersecurity sector that decision makers often fail to sufficiently analyse. The public sector may not be ready to offer AI-related career paths aimed to train and to retain skills and talents.[299] If this is the case for all public sector departments, it is more critical for security-related agencies. Policies should aim to monitor the transformation of the sector and the skill shortages therein, while ensuring a smooth transition as AI is incorporated into existing cybersecurity professions.[300]

8. Overarching recommendations

On 21 April 2021 the European Commission published the "Regulation on a European Approach for Artificial Intelligence"[301] fostering ad hoc protection for high-risk AI systems, based on a secure development life cycle. An AI application is considered to be risky if the sector in which it is applied involves significant risks, or the application itself involves a significant risk. Where it is established that an AI application entails a high risk, a number of requirements apply, including those regarding the quality of training data, those related to transparency about AI systems, or those on the accuracy and reproducibility of outcomes.[302]

The Task Force on AI and Cybersecurity supports such an approach and, based on an extensive review of the existing literature and on the contributions from the participants, suggests the following recommendations to policymakers, the private sector, and the research community:

Recommendations – AI for cybersecurity

Specific EU policy measures that would ease the adoption of AI in cybersecurity in Europe include:

1. Enhancing collaboration between policymakers, the technical community and key corporate representatives to better investigate, prevent and mitigate potential malicious uses of AI in cybersecurity. This collaboration can be informed by the lessons learned in the regulation of cybersecurity, and from bioethics.

2. Enforcing and testing the security requirements for AI systems in public procurement policies. Adherence to ethical and safety principles should be regarded as a prerequisite for the procurement of AI applications in certain critical sectors. This would help to advance discussions on AI and safety in organisations, including at the board level.

[299] US National Security Commission on Artificial Intelligence (2019), Final Report, p. 26 (www.nscai.gov).

[300] OECD (2019a), op.cit., p. 110.

[301] European Commission, Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts, COM(2021) 206 final, Brussels, 21.4.2021.

[302] European Commission, White Paper On Artificial Intelligence - A European approach to excellence and trust, COM(2020) 65 final, Brussels, 19 February 2020.

3. Encouraging information sharing of cybersecurity-relevant data, for example data to train models according to established best practice. Private sector-driven, cross-border information sharing should also be supported by providing incentives for cooperation and ensuring a governance framework that would enable legal certainty when exchanging data.

4. Focusing on supporting the reliability of AI, rather than its trustworthiness, in standards and certification methods. The following developing and monitoring practices are suggested to ensure reliability and mitigate the risks linked to the lack of predictability of AI systems' robustness:

    - Companies' in-house development of AI applications models and testing of data
    - Improving AI systems' robustness through adversarial training between AI systems
    - Parallel and dynamic monitoring or punctual checks of AI systems through a clone system as control, which would be used as a baseline comparison to assess the behaviour of the original system.

5. Supporting and internationally promoting proactive AI cybersecurity certification efforts, to be coordinated by ENISA. These should demand that assessment actions be taken prior to deployments and during the whole life cycle of a product, service, or process.

6. Envisaging appropriate limitations to the full openness policy for research output, such as algorithms or model parameters,[303] to enable a more complete assessment of the security risks related to the technology and its dissemination, balanced with the EU policy objective of fostering innovation.

7. Promoting further study and regulatory interpretation of the General Data Protection Regulation (GRPR) provisions, even at the national level (for instance, with respect to Recitals 49 and 71, on data-sharing practices for information security aims), in the context of both AI for cybersecurity and applications aimed at making AI secure.

8. Addressing the challenges of adequately enforcing the personal data protection rules posed by datasets of mixed personal and non-personal data.

9. Evaluating how the use of AI systems in cybersecurity research and operations could be impacted by the current (and future) dual-use and export control regulatory framework;[304] drawing up clear rules that respect EU (treaty-based) values without hampering trade and sacrificing openness; establishing an EU-level regulated dual-use technology transfer mechanism, through the support of the industry and within the boundaries fixed by the Wassenaar Agreement, for defining a possible dual-use technology transfer mechanism and creating an avenue for developing a common approach among institutions dealing with dual-use technologies.

[303] Models are often made public and 'open source' having successfully led to AI applications performing tasks with a broad general interest.

[304] Wassenaar Agreement and European Commission Regulation No 428/2009.

10. Enhancing the cooperation between military and civilian entities in AI-based development topics by applying capability development concepts from the military sector (which reflect strong cybersecurity requirements) to civilian AI applications, or by defining a reference architecture for cybersecurity specifically for AI applications, to be used in both civilian and military domains.

11. Addressing the skills shortage and uneven distribution of talents and professionals among market players. The public sector, as well as security-related agencies, should be ready to offer AI-related career paths and to train and retain cybersecurity skills and talents. The transformation of the cybersecurity sector should be monitored while ensuring that AI tools and their use are incorporated into existing cybersecurity professional practice and architectures.

Recommendations – Cybersecurity for AI

Ways to make AI systems safe and reliable when developing and deploying them include:

12. Promoting suitability testing before an AI system is implemented in order to evaluate the related security risks. Such tests, to be performed by all stakeholders involved in a development and/or a deployment project, should gauge value, ease of attack, damage, opportunity cost and alternatives.[305]

13. Encouraging companies to address the risk of AI attacks once the AI system is implemented. General AI safety could also be strengthened by putting detection mechanisms in place. These would alert companies that adversarial attacks are occurring, that the system in question is no longer functioning within specified parameters in order to activate a fallback plan.[306]

14. Suggesting that AI systems follow a secure development life cycle, from ideation to deployment, including runtime monitoring and post-deployment control and auditing.

15. Strengthening AI security as it relates to maintaining accountability across intelligent systems, by requiring adequate documentation of the architecture of the system, including the design and documentation of its components and how they are integrated.[307] Strengthening measures include:

    a. Securing logs related to the development/coding/training of the system: who changed what, when, and why? These are standard procedures applied for revision control systems used in developing software, which also preserve older versions of software so that differences and additions can be checked and reversed.

[305] Some Task Force participants raised concerns about the feasibility of this requirement. A particular argument was that, given the fast pace of adoption of AI systems, innovation would be stifled if a suitability test were required for each and every AI system implemented.

[306] Some Task Force participants raised concerns about the maturity of AI technology, which at the current state of the art might not allow for effective detection mechanisms to be put in place.

[307] This should not be regarded as an exhaustive list of cybersecurity requirements for AI, for which further study will be required.

b. Providing cybersecure pedigrees for all software libraries linked to that code.

c. Providing cybersecure pedigrees for any data libraries used for training machine learning (ML) algorithms. This can also show compliance with privacy laws and other principles.

d. Keeping track of the data, model parameters, and training procedure where ML is used.

e. Requiring records that demonstrate due diligence when testing the technology, before releasing it. These would preferably include the test suites used so that they can be checked by the company itself or by third parties and then reused where possible.[308]

f. Maintaining logs of inputs and outputs for AI-powered operating systems, depending on the capacities of the system and when feasible, and assuming these are cybersecure and GDPR compliant.

g. Requiring in-depth logging of the AI system's processes and outcomes for life-critical applications such as automated aeroplanes, surgical robots, autonomous weapons systems, and facial recognition for surveillance purposes. For non-critical applications, the volume of input data should be evaluated before requiring an in-depth logging strategy. This is to avoid unfair competition between big and small players due to implementation costs.

h. Enhancing AI reliability and reproducibility by using techniques other than logging such as randomisation, noise prevention, defensive distillation, and ensemble learning.

16. Suggesting that organisations ensure models are fully auditable at time/point of failure, and to make the information available for subsequent analysis (e.g. analysis required by courts).[309] New methods of auditing systems should also be encouraged, such as restricting them to a trusted third party, rather than openly pushing datasets.

17. Suggesting that organisations develop an attack incident-response plan, and create a map showing how the compromise of one asset, dataset, or system affects other AI systems, for example how systems can exploit the same dataset or model once the attack has occurred. Policymakers should support the development and sharing of best practice. Validating data collection practices could guide companies in this process, for example in identifying potential weaknesses that could facilitate attacks or exacerbate the consequences of attacks.

[308] Some Task Force participants raised concerns about the proportionality and intrusiveness of this requirement, especially in terms of compliance with the GDPR provisions.

[309] Some Task Force participants raised concerns about the feasibility and economic burden of this requirement.

Annex I. Glossary[310]

Adversarial example: Inputs formed by applying small but intentional perturbations to examples from a dataset, such as that the perturbed input results in the model outputting an incorrect answer with high confidence.

Artificial Intelligence (AI): *"Machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments."*[311] It is important to distinguish between symbolic and non-symbolic AI. In symbolic (or traditional) AI, programmers make use of programming languages to generate explicit rules to be hard coded into the machine. Non-symbolic AI does not rely on the hard coding of explicit rules. Instead, machines are able to process an extensive set of data, deal with uncertainty and incompleteness, and autonomously extract patterns or make predictions.

Backdoors: A type of possible AI attack that manipulates the behaviour of AI algorithms. Backdoor attacks implant the adversarial vulnerability in the machine learning model during the training phase. They rely on data poisoning, or the manipulation of the examples used to train the target machine learning model. The adversary creates a customised perturbation mask applied to selected images to override correct classifications. The backdoor is injected into the victim model via data poisoning of the training set, with a small poisoning fraction, and thus does not undermine the normal functioning of the system.

Black box: Partial/total lack of knowledge or understanding about the model's functioning and decision-making process apart from the input fed into the system and the output.

Classifier: Algorithm that maps input data (e.g. pictures of animals) into specific categories (e.g. 'dog').

Cybersecurity: *"Cybersecurity refer to security of cyberspace, where cyberspace itself refers to the set of links and relationships between objects that are accessible through a generalised telecommunications network, and to the set of objects themselves where they present interfaces allowing their remote control, remote access to data, or their participation in control actions within that cyberspace."*[312]

Data poisoning: A type of possible AI attack where attackers bring crafted flawed data into the legitimate dataset used to train the system to modify its behaviour (e.g., increase the prediction error of the machine learning).

Development and operations (DevOps): Set of practices that combines software development and IT operations.

Dwell time: The length of time a cyber-attacker has free reign in an environment, from when they get in until they are eradicated.

[310] This section refers to both definitions provided throughout the report, and to S. Herping (2019), op. cit, p. 40-46.

[311] See OECD.AI Policy Observatory (www.oecd.ai/ai-principles).

[312] ENISA (2015), Definition of Cybersecurity – Gaps and overlaps in standardisation, December.

Generative adversarial network (GAN): A class of machine learning that enables the generation of fairly realistic synthetic images by forcing the generated images to be statistically almost indistinguishable from real ones.

Honeypot: Computer systems intended to mimic likely targets of cyberattacks and "intended to attract cyberattacks. Honeypots use attackers' intrusion attempts to gain information about cybercriminals and the way they are operating, or to distract them from other targets."[313]

Honeytoken: Chunks of data that look attractive to potential attackers.

Machine Learning (ML): Subset of non-symbolic AI. *"[A] set of techniques to allow machines to learn in an automated manner through patterns and inferences rather than through explicit instructions from a human. ML approaches often teach machines to reach an outcome by showing them many examples of correct outcomes. However, they can also define a set of rules and let the machine learn by trial and error."*[314]

Machine Learning Approaches

> **Deep Learning:** Models that involve feeding the training data through a network of artificial neurons to pull out distributional figures or other high-level information. Deep learning is a subset of neural networks. It is a particularly large neural network composed of hierarchical layers that increase the complexity of the relationship between input and output. Deep learning is an architecture able to implement supervised, unsupervised and reinforcement learning.
>
> **Reinforcement Learning:** Model that involves creating a system of rewards within an artificial environment to teach an artificial agent how to move through different states. It is commonly used in robotics for navigation and as a tool for solving complex strategy games.
>
> **Supervised Learning:** Most common form of Machine Learning, where the machine learns to map input data to known targets, given a set of examples, which are often annotated by humans. The system is trained with data that is tagged with predetermined categories.
>
> **Unsupervised Learning:** This consists of finding meaningful transformations of the input data without the help of any targets. The trained system itself creates categories underlying similarities in the training data.

(Machine Learning) Model: Artefact that is created by the training process. It represents the rules, numbers, and any other algorithm-specific data structures required to make predictions.

Neural network: Subcategory of Machine Learning characterised by layers that compute information in parallel. Neural networks are formed by interconnected nodes passing information to each other. Neural networks incrementally modify their own code to find and optimise links between inputs and outputs. Neural networks are loosely based on the biological concept of brains.

---

[313] See Kaspersky, "What is a honeypot?" (www.kaspersky.com/resource-center/threats/what-is-a-honeypot).

[314] OECD (2019°), op. cit.

Payload: Attack component responsible for executing an activity to harm the target. Examples of malicious payloads are worms or ransomware. Malicious payloads remain inactive until activated.

Perturbation: Small, hardly (or not at all) recognisable changes of an input that causes prediction errors (e.g. overlay of an input on an image that causes the image to be recognised as something else).

Reverse engineer the AI model: Act of copying the product by looking at how it is made. By gaining access to the AI model through reverse engineering attackers are able to perform a more targeted and successful adversarial attack.

Tamper: Interfere with (something) to cause damage or make unauthorised alterations.

Tampering of the categorization model: A type of possible AI attack. By manipulating the categorisation models of an AI system, attackers could modify the final outcome of AI system applications.

Technology trap: Fear that the use of AI in cybersecurity in the short run will make things worse for everyone in the long run by slowing the pace of automation and innovation.

Threat model: Structured approach that helps to identify possible threats to IT systems. Also referred to as a "defined set of assumptions about the capabilities and goals of the attacker wishing the system to misbehave."[315]

Training data: Set of data, describing a behaviour or problem, used to achieve a model that matches the data. The model learns to mimic the behaviour or to solve the problem from such a dataset.

[315] N. Papernot and I. Goodfellow (2016), "Breaking things is easy", Cleverhans-blog, 16 December (www.cleverhans.io/security/privacy/ml/2016/12/16/breaking-things-is-easy.html).

Annex II. List of Task Force members and invited speakers

Coordinator and rapporteur: Lorenzo Pupillo, CEPS
Rapporteurs: Stefano Fantin, KU Leuven, Afonso Ferreira, CNRS and Carolina Polito, CEPS

Advisory Board

Joanna Bryson, Hertie School for Governance, Berlin
Jean-Marc Rickli, Geneva Centre for Security Policy
Marc Ph. Stoecklin, Security Department, IBM Research Center, Zurich
Mariarosaria Taddeo, Digital Ethics Lab, University of Oxford

Companies and European organisations

Accenture, Barbara Wynne
Confederation of Danish Industry, Andreas Brunsgaard
Deloitte, Massimo Felici
ETNO, Paolo Grassia
F-Secure, Matti Aksela
FTI Consulting, William Dazy
Huawei, Sophie Batas, Mark K. Smitham
ICANN, Elena Plexida
JP Morgan, Renata Shepard
McAfee, Chris Hutchins
Microsoft, Florian Pennings, Rachel Azafrani, Rob Spiger
Palo Alto Networks, Sebastian Gerlach
Raiffsen Bank International AG, Martin Koeb
SAP, Corinna Schulze
Vectra AI, Inc., Matt Walmsley, Sohrob Kazerounian
VISA, Pedro Simoes
Wavestone, Gerome Billois, Aude Thirriot, Déborah Di Giacomo, Carole Méziat
Zurich Insurance Company, Ltd., Marc Radice

European institutions, agencies and intergovernmental organisations

Council of the European Union, General Secretariat, Monika Kopcheva*
EDA, Mario Beccia, Giuseppe Giovanni Daquino**
ENISA, Apostolos Malatras**
European Commission, Nineta Polemi
European Central Bank, Klaus Lober**
European Parliament, Adam Bowering

European Investment Bank, Harald Gruber

Financial Conduct Authority, Tim Machaiah

Hybrid CoE - The European Centre of Excellence for Countering Hybrid Threats in Helsinki (FI), Josef Schroefl**

NATO, Michal Polakow**

OECD, Laurent Bernat

* *In her own capacity*

** *Observer*

Academics/Think tanks

Centre for Economics and Foreign Policy Studies (EDAM), Usaal Sahbaz

Centre for Russia, Europe, Asia Studies (CREAS), Theresa Fallon

DeepIn Research Network/I-COM, Antonio Manganelli

TNO, Alex Sangers

University of Amsterdam, Federica Russo

Civil society

Homo Digitalis, Eleftherios Chelioudakis

Humanity of Things Agency, Marisa Monteiro

Invited speakers

Marcus Comiter, Harvard Kennedy School, Belfer Center

David Clark, MIT Computer Science & Artificial Intelligence Laboratory

David Gibian, Calypso Labs

Miguel Gonzalez-Sancho, European Commission

Martin Dion, Kudelski

Billy Hewlett, Palo Alto Networks

Jouni Kallunki, F-Secure

Andres Ojamaa, Guardtime

Andrea Renda, CEPS

Lucilla Sioli, European Commission

Tyler Sweatt, Calypso Labs