Study of Research and Guidance on the Cyber Security of AI

Jesus Martinez del Rincon, Ehsan Nowroozi, Eleni Kamenou, Ihsen Alouani, Sandeep Gupta, and Paul Miller

Centre for Secure Information Technologies (CSIT), Queen's University Belfast (QUB), United Kingdom

EXECUTIVE SUMMARY

Groundbreaking innovations in Artificial Intelligence (AI) technology facilitate the automation of a wide range of complex tasks across diversified domains. These advancements can significantly contribute to the development of applications or systems for trivial to highly critical domains such as transportation and healthcare, thereby benefiting end-users. However, the indiscriminate use of AI, without due consideration of the implications of releasing AI models into the wild, creates aberrant opportunities for malicious actors to exploit vulnerabilities and gain significant advantages.

DSIT commissioned Queen's University Belfast to compile a comprehensive meta-study of the existing research and guidance on the cyber security of AI, including academic and industrial research, standards and regulations. The study reviewed a total of ≈18,000 publications in the field, including a thorough analysis and reporting on more than 415 publications. A Rapid Evaluation Assessment (REA) approach was applied to systematically collect the necessary information through keyword searching of the bibliometric databases.

The goal of this report is twofold. First, to collect the existing research on the security and privacy of AI published by both industry and academia. Second, to report on publications that may support AI developers and engineers in the design of secure AI models and systems. This includes publications by academia, governments, industry (particularly AI companies) and technical authorities. This study also aims to identify the primary actors and stakeholders engaged in the AI Security field. Consequently, our objective is to furnish a comprehensive review encompassing the latest advancements on AI security and to pinpoint gaps in their practical application.

The key findings of this study are highlighted below:

- 415 documents on the cybersecurity of AI were found, including 323 academic papers, 31 industrial reports and white papers, 41 standards organisations documents and 20 governmental documents.
- Research focuses on two main themes, ''Attacks'', which includes risks, vulnerabilities and threat modeling; and ''Defences'', including technical solutions, recommendations and guidance.
- The methodology in academic venues is usually validated quantitatively through experimentation in a laboratory setting with unrealistic threat models, while the remaining stakehoders' studies are mostly based on non-empirical analysis.
- Most research and guidance focusses on the design of Secure AI solutions, with a significantly smaller amount for development, deployment and monitoring of AI models.

=

1 INTRODUCTION

Artificial intelligence (AI) is a transformative technology that has the potential to change and improve society and people's lives. Thus, AI and Machine Learning (ML) technologies can help improve health, the environment, cyber security or transportation, to name a few. It can drive economic growth for all and increase the productivity of companies and businesses, as well as support scientific advancements. To reach these goals, AI/ML technologies must be designed, developed and deployed in a secure manner.

The Department for Science, Innovation and Technology (DSIT) commissioned Queen's University Belfast to compile a comprehensive meta-study of the existing research and guidance on the cyber security of AI. Any primary research, subsequent findings, or recommendations do not represent Government views or policy and are produced according to academic ethics, quality assurance and independence.

1.1 Objectives

Our systematic study encompasses:

- Existing research on the security of AI models and systems published by both industry (including AI companies in particular) and academia.
- Publications that can support AI developers and engineers to design secure models and systems. This includes publications by governments, industry (particularly, AI companies) and technical authorities.

This study focusses on the topic of Security of AI and adopts a strict definition of Security of AI as per the CIA Triad -Confidentiality, Integrity, and Availability-. It therefore considers only those risks and vulnerabilities where there is a malicious intent and a willing attacker - adversary. Thus, these are out of the scope of this study:

- Other AI-related risks such as safety, reliability, trustworthiness, biases or explainability.
- The topic of AI for cyber security, i.e. those studies that aim to develop AI-based solutions for cyber security.

1.2 Report Structure

This work is structured as follows: Section 2 addresses the background of AI and the security of AI, describing the AI lifecycle and the main concepts of how an adversary can employ adversarial attacks during the different AI model phases. Section 3 discusses the methodology used in this study. This section addresses the main search methodology used as well as the keyword and the bibliometric database that we considered for finding related papers at top conferences and journal venues. The core of the report is contained in Section 4, which describes our findings and provides a quantitative and qualitative analysis of the existing research and guidance from the different stakeholders involved, i.e. academia, industry, government, and standards organizations. Finally, Section 5 concludes the study.

2 BACKGROUND

This section seeks to provide the core definitions, context and processes in AI and the Security of AI to support the reader throughout the report.

2.1 Artificial intelligence

Artificial intelligence (AI) refers to machines that exhibit the ability to comprehend and learn tasks. Within AI, **machine learning (ML)** is the field of study concerned with the development and study of statistical algorithms that can effectively learn tasks from data rather than execute explicit instructions. Recently, **Deep Learning (DL)** approaches based on artificial neural networks (NN) have been able to surpass most previous ML approaches in performance

Training Process: Training, validation, and testing are essential parts of the AI model development pipeline. The training process includes preparing the data, determining the model architecture, and optimizing the parameters to minimize loss expectation over the training dataset. The validation sets boost the performance of the model and prevents overfitting. Finally, throughout the testing process (also known as inference), the model uses new data to evaluate its ability to generalize.

AI Lifecycle is the iterative process of solving a business problem by using an AI solution, that evolves from a concept to its effective implementation. The steps of the life cycle are the design, development, deployment and monitoring phases, through which the process iterates multiple times to achieve a refined solution.

2.2 Security of AI

Methodology Process. To consider the Security of AI, a range of methodologies have been applied across the different stakeholders. These can be divided into empirical analysis versus non-empirical analysis. **Empirical analyses** comprise those which rely on quantitative evaluations. This normally consist of selecting/creating a set of attacks under a predefined threat model, expose the AI model to those attacks, quantify the attack success, implement/design a defence technique and revaluate the attack(s) success under the new conditions. **Non-empirical analyses** are those which do not rely on quantitative data but using instead current events, literature reviews, personal observations, expert panels and subjectivity to draw conclusions.

Attacks: Substantial research has been devised to assess and challenge the security of AI and ML models. Numerous attacks have been designed to expose the vulnerability of AI systems. These attacks can impact all the phases of the ML lifecycle. For instance, some attacks have been designed to affect the training phase. This may include ‘poisoning attacks’ and ’backdoor attacks’ when the training data are intentionally altered to hinder the model’s learning process. Similarly, during the testing phase, the probability of ‘evasion attacks’ increases when the input data are modified to deceive the model during inference and alter the prediction of the AI system, usually by introducing minor and imperceptible alterations (adversarial attacks). Attacks can also be classified according to the adversary’s goal. Those that aim to diminish the effectiveness or detection performance of the ML model are described as questioning the model’s integrity. On the other hand, those that aim to recover private or confidential data embedded into the model or the training set are described as confidential or privacy attacks, e.g. model stealing, model inversion and membership inference.

Threat Model refers to the identification of the potential security threat of the ML system. It comprises the profile of the adversary, its motivations, knowledge, and level of access, as well as the amount of damage it can potentially produce. The threat model classifies adversarial attacks as white-box and black-box attacks. The concept of grey-box threat has also been used to describe intermediate but not complete knowledge of the targeted ML model. One key distinction between the two modes is the adversaries' knowledge.

- ***White-box threat.*** This setting assumes strong adversaries who possess complete knowledge of their target model, including its architecture and parameters, or if there are defences deployed. This facilitates the adversary samples being generated in the target model, simplifies the design of AI attacks and improves their success rate. They are also unrealistic and may not be viable or representative of real-world threats.
- ***Black-box threat.*** This setting assumes the attacker has access only to the input/output of the victim model and has no information about its internal architecture. As such, it only allows adversaries to build samples exclusively through query access to the largely unknown ML model. The adversary attempts to estimate the victim model's behavior to generate adversarial examples.

Adversarial training (AT) is among the strongest and most widespread defence methods in improving the robustness of AI models against adversarial attacks. The intuition behind adversarial training is to inject adversarial samples in the training data and expose the model to them during training to build resilience against malicious perturbations.

Certifiable Defences are mathematically proven methods that guarantee the ML model's robustness against a certain level of adversarial attack. This differs from **Empirical Defences**, such as adversarial training, pre-processing and gradient masking.

3 METHODOLOGY

This section delineates the research methodology applied for this survey. All references within this study exclusively pertain to the last decade, spanning from year 2013 -inclusive- up until December 2023 -inclusive-. The year 2013 is used as starting point for being the first academic paper on describing adversarial attacks in ML, starting de-facto the field of AI security [59]. Only English-language sources are included in the scope of this report. In order to classify all documents, a taxonomy was first created by analysing the survey papers found and comprehensively examined and compared with existing taxonomies. This resulted on dividing the papers into 'Attacks' and 'Defeces', similar to [65, 66] , as well as the corresponding subgroupings described in Section 4.1.

3.1 Selection of Bibliometric Database

Scopus and Web of Science (WoS) are chosen as the primary bibliometric databases for our academic search, with Scopus being selected for this study due to its nearly 60% broader coverage compared to WoS according to [58].

Our study deliberately selected studies from the most prominent conferences and journals in the domains of ML and cyber security, such as CVPR, NeurIPS, ICCV/ECCV, IEEE S&P, USENIX Sec, NDSS, ACM CCS, ACM CODASPY, ICML, IEEE Transactions on Dependable and Secure Computing, Computers & Security, IEEE Transactions on Information Forensics and Security, IET Information Security, Journal of Cybersecurity, ACM Transactions on Privacy and Security, Journal of Information Security and Applications, IEEE Transactions on Neural Networks and Learning Systems, Pattern Recognition, and Journal of Machine Learning Research and other first quartile venues. This selection aims to draw conclusions from the most innovative and rigorous state-of-art research.

Moreover, in order to find other relevant documents for this study, related to guidance, regulations and recommendations, such as white papers, industrial reports and standards, Google (https://www.google.com) and Google Scholar (https://scholar.google.com/) were used as sources exclusively for documents beyond academic research publications.

3.2 Keyword Search

A Rapid Evaluation Assessment (REA) approach was designed to systematically collect the desired information within a constrained timeframe. By using keyword searches in academic databases (Scopus, WoS) as well as Google, a systematic literature review on the security of AI models and systems, and guidance and recommendation for their deployment was conducted.

To cover the relevant landscape, we searched the academic research literature using the following keywords: "AI", "security", "threat model", "training attacks", "machine learning", "neural networks", "causative attacks", "poisoning attacks", "backdoor attacks", "testing attacks", "exploratory attacks", "white-box attacks", "black-box attacks", "oracle attacks", "membership inference", "model inversion", "stealing", "model extraction", "robustness", "adversarial training", "pre-processing", "denoising", "gradient Masking", "distillation", "obfuscation", "quantisation", "certified defenc(s)e", "certifiable defenc(s)e", "homomorphic encryption", and "differential privacy" on WoS and Scopus.

To capture any other publications not considered in academic research databases, an additional search on Google and Google Scholar was conducted using keywords that restrain the search to documents different from research publications. Specifically, three searches were conducted:
"AI security" AND white papers OR reports OR manuals
"AI security" AND framework OR standards
"AI security" AND tools OR workshop
Please note that only documents providing some level of analysis (based on data or theoretical),

guidance and recommendations, are considered. Thus, marketing and investment reports are not considered in this document.

To address the vast body of work in the Security of AI while being exhaustive and systematic, we followed the procedure depicted in Figure 2:

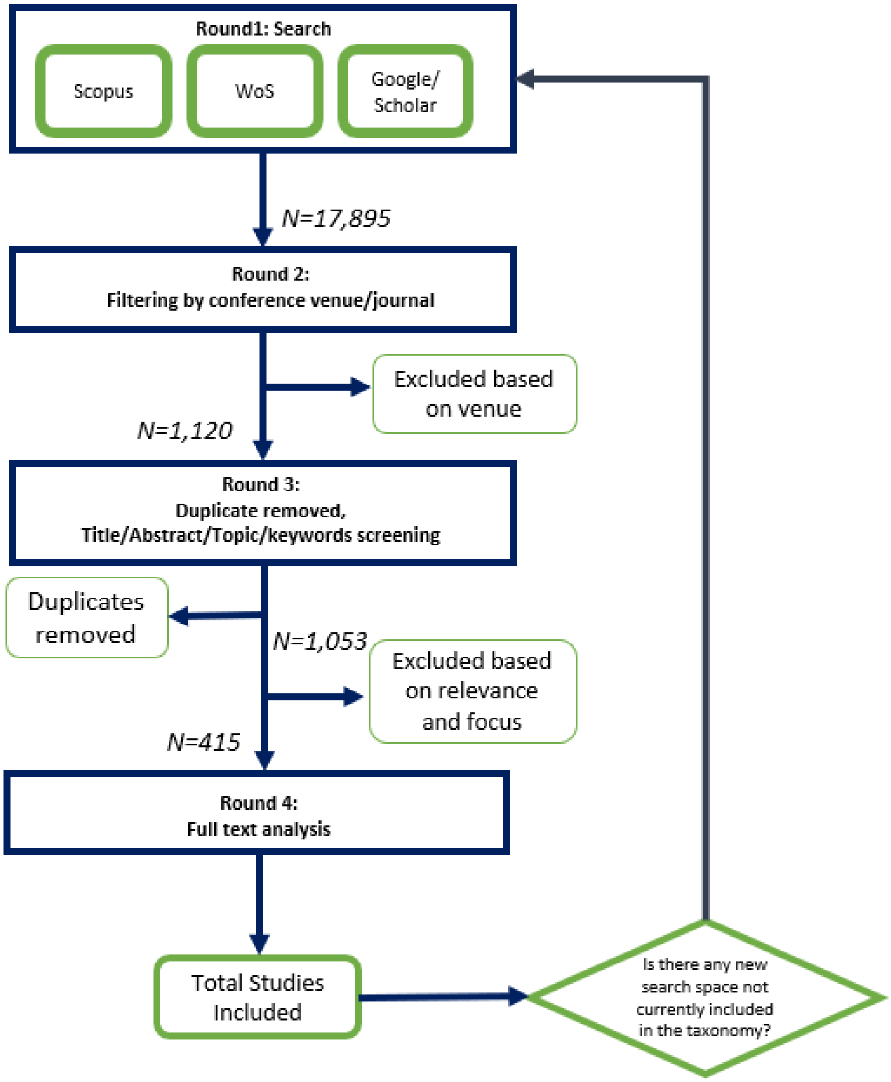
Round1: Search
Scopus
WoS
Google/ Scholar
N=17,895
Round 2: Filtering by conference venue/journal
Excluded based on venue
N=1,120
Round 3: Duplicate removed, Title/Abstract/Topic/keywords screening
Duplicates removed
N=1,053
Excluded based on relevance and focus
N=415
Round 4: Full text analysis
Total Studies Included
Is there any new search space not currently included in the taxonomy?

Figure 2. Methodology for document search. *N* indicates the resulting number of papers after each stage and filter.

Figure 3 shows the result of the non-academic guidance and recommendation documents found. Figures 4 and 5 depict the result (in logarithmic scale) of research publications before and after filtering desegregated on Attack and Defence themes and subgroups, respectively.

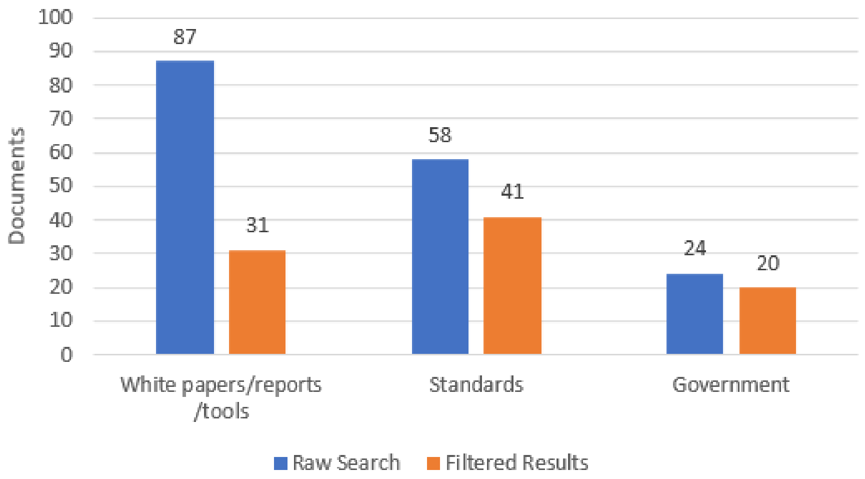

Figure 3. Number of non-academic publications per keyword (in linear scale), for the *Industry, Standards and Government* sections, on Google and Google Scholar search engines. *Raw search* refers to the total number of documents found in our initial search, while *Filtered results* refers to the refined number after removing duplicates and filtering them by their relevance to this study.

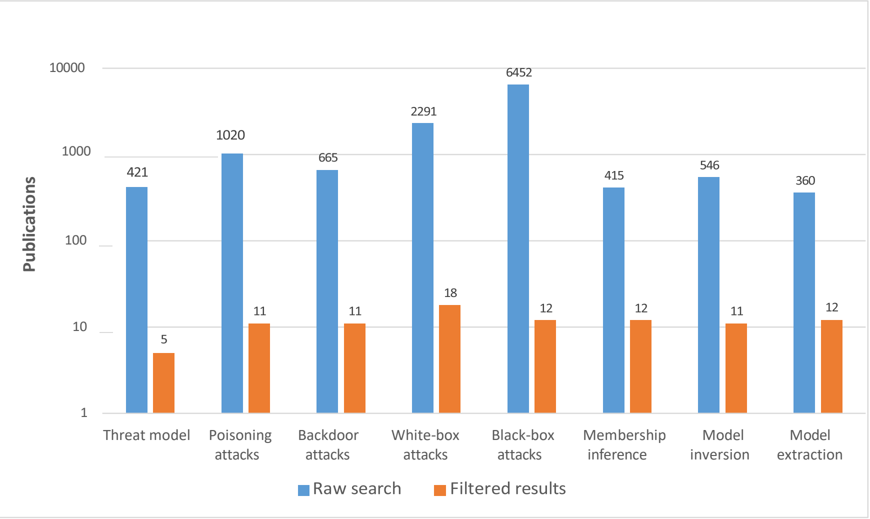

Figure 4. Number of publications per keyword (in logarithmic scale), for the *Attack* grouping, on both search engines (WoS and Scopus), over the last 10 years. *Raw search* refers to the total number of papers found in our initial search, while *Filtered results* refers to the number of papers after removing duplicates and filtering them by the quality of the venue and their relevance to this study.

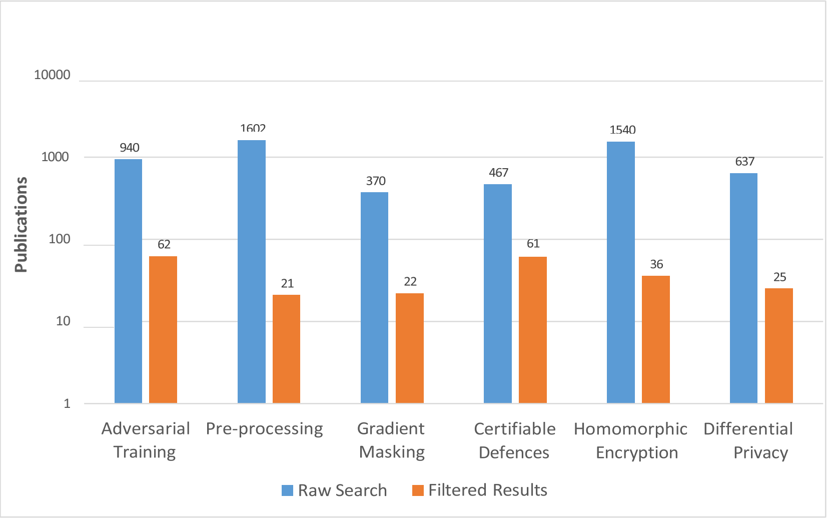

Figure 5. Number of publications per keyword (in logarithmic scale), for the *Defence* grouping, on both search engines (WoS and Scopus), over the last 10 years. *Raw search* refers to the total number of papers found in our initial search, while *Filtered results* refers to the number of papers after removing duplicates and filtering them by the quality of the venue and their relevance to this study.

4 FINDINGS

In this section, quantitative and qualitative analysis of the existing research and guidance are presented for the different stakeholders involved, i.e. academia, industry, government, and standards development organizations. Figure 6 summarses the number of analysed documents by stakeholder.

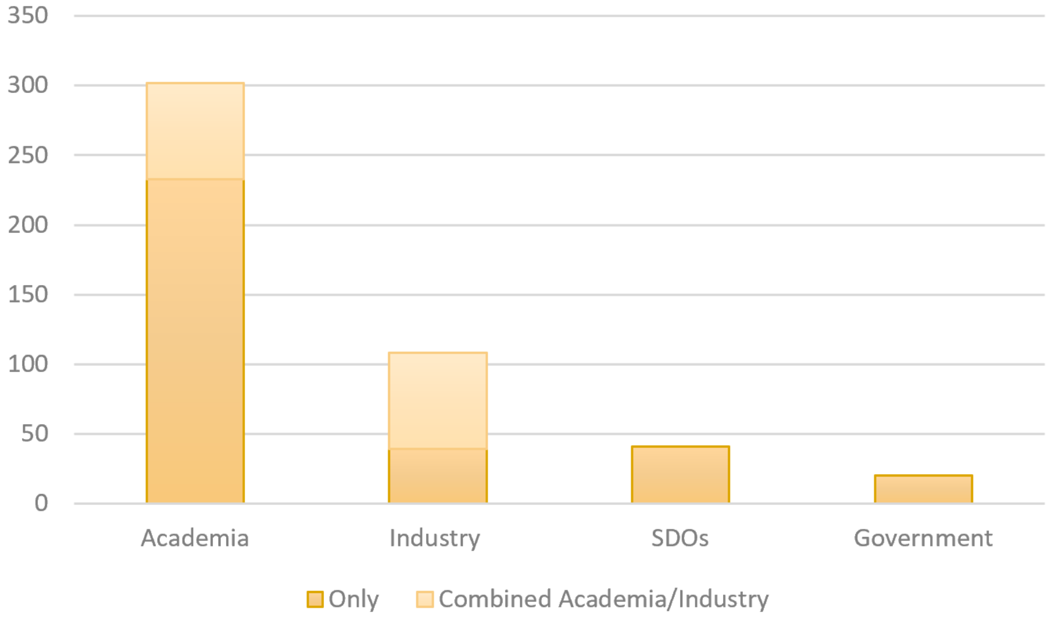

Figure 6. Number of publications by stakeholder, either in isolation or in collaboration with each other.

4.1 Academia

Academia is the biggest current contributor to the field of Security of AI. Figure 7 shows the temporal trend of academic research on AI security, by placing all papers analysed in this study in Chronological order according to the publication date. It can be seen how the security of AI is a recent but booming research field, which has exploded in the last 5 years. Please note that 2023 statistics may be affected for the publication process and updating of the databases.

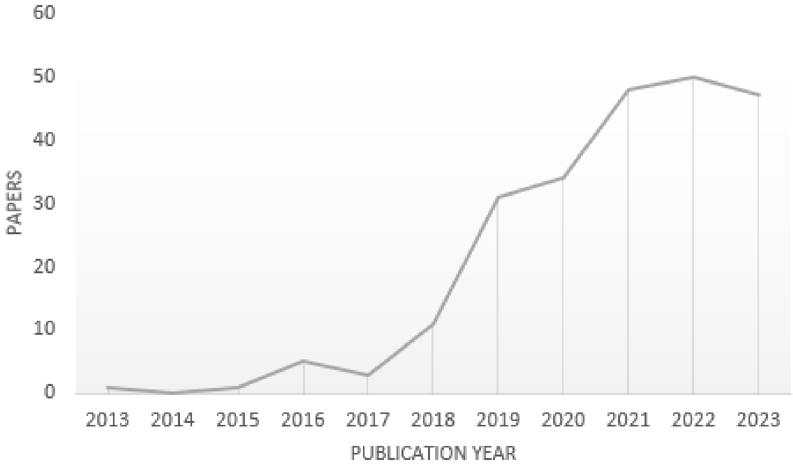

Figure 7. Number of publications per year in AI security.

Out of the 323 academic papers fully analysed in this study, 70% of them were performed solely by academia, and 91% of them involved at least one academic institution. These publications focus on 2 main themes, Attack and Defence, and 11 subgroups, including Poisoning/Backdoor Attacks, Adversarial Attacks, Membership Inference Attacks, Model Inversion Attacks, Model Extraction Attacks, Adversarial Learning, Preprocessing Defences, Gradient Masking, Homomorphic Encryption and Differential Privacy.

Research conducted in the topic of vulnerability assessment and threat modelling is largely led by Academia as the primary stakeholder, comprising hundreds of papers. This academic research is conducted experimentally using multiple academic public datasets and benchmarks and well-known AI models. However, it uses unrealistic threat models under a lab setting, such as the attacker's ability to easily interfere with the training process (25% of the papers) or presenting white-box attacks (60% of the papers). Most academic research also focuses exclusively on image recognition applications due to the simplicity of generating attacks, which does not translate to other applications.

A similar body of knowledge has been investigated by academia on the defensive mechanism to preserve model integrity. 64% of the analysed papers focus on empirical methods, with adversarial training being the strongest and most widespread defence method in improving the robustness of AI models against adversarial attacks. Despite having demonstrated promising results, they have only been evaluated under lab settings using academic datasets, and under threat models that range from a pure academic exercise -still useful to find the security boundaries- to a more realistic black-box setting. No evaluation in the wild has been found. The best approximation being a handful of studies where realistic testbeds or simulators are employed.

Table 1. Main actors, in academia and industry, in the research field of security of AI, ordered by number of academic research publications. In brackets, we indicate the number of academic papers found in this study.

Universities	Companies
University of Illinois (17)	Google (14)
Carnegie Melon University (15)	IBM (10)
University of California (14)	Microsoft (9)
Tsinghua University (14)	Alibaba (5)
Zhejiang University (14)	Bosch (4)
MIT (12)	OpenAI (3)
Chinese University of Hong Kong (10)	Tencent (3)
Chinese Academy of Sciences (9)	JD.com (3)
University of Texas (8)	Anazon (3)
University of Wisconsin–Madison (8)	Pluribus One (3)
Nanjing University (8)	NVIDIA (2)
ETH Zurich (8)	Ant Group (2)
University of Maryland (7)	Meta/Facebook (2)
Pekin University (7)	Samsung
Nanyang Technological University (7)	Airbus
University of Toronto (7)	Intuit Inc
Princeton University (7)	Baidu
Xidian University (7)	Sony
Shanghai Jiao Tong University (6)	Voleon Group
University of Science and Technology of China (6)	SAP
Vector Institute (6)	Zhongda Group
INRIA (5)	RealAI
GeorgiaTech (5)	Uber
University of Cagliari (5)	Adobe
University of Michigan (5)	Norton
...	Lumeros AI
Alan Turing Institute (4)	ByteDance AI Lab
University of Oxford (4)	NetApp
Imperial College London (3)	Foxstream

Moreover, adversarial training can only demonstrate their effectiveness experimentally without providing mathematical guarantees, and relies on the rigor of the evaluation to demonstrate its potential. Given the laboratory setting on most evaluations, this poses doubts on their transferability to real-world conditions. On the contrary, certifiable approaches have attracted a substantial recent interest by the academic community, particularly for critical applications. They offer theoretical proof of their effectiveness at the cost of lower clean accuracy (i.e. accuracy in normal conditions and under no attacks) compared to empirical defences. Furthermore, certifiable approaches are currently impractical and costly even on small ML models, requiring several orders of magnitude more time for training.

Methods developed by academia are mostly at prototype level, with many attacks and solutions being made public but through research repositories, rather than professional code. In the best case, toolboxes are being released to increase their use, but this still requires an expert level of knowledge in AI and may not be accessible for a broader range of practitioners, software developers and engineers. No evidence of deployment of these solutions has been found.

The vast majority, if not all, of the analysed academic research papers do not contain specific guidance or recommendations. This should not be understood as the total absence of academic interest on the matter, but rather as the main focus still being on identifying the risks and developing better AI security methodologies and theories, especially given the youth of the field (see Figure 6) and the existing gaps [17]. This is particularly the case on the top tier venues in cyber security and AI, where theoretical contributions are expected, rather than implementation and deployment considerations.

A small set of academic research focus on recommendations for business and policy makers, usually with input from disciplines beyond computer science such as management, policy or law. These papers are not based on data but on theoretical analysis [52, 14, 8, 50, 6, 7, 51]. As such, they are published in small, less recognised venues or in the shape of white papers. Among the relevant ones, a technical assessment of the EU AI Act[34] highlights the technical challenges that derive from the regulation gap between the proposed requirements and the available AI security countermeasures, and the necessity for an AI security evaluation framework.

4.2 Industry

Of the 415 sources identified, 28% were created by industry (see Figure 6). Figure 8 shows the split of analysed documents in this section, which includes research papers in academic venues, industrial reports and white papers and open-source tools. Those academic papers can be classified in the main themes of Attacks and Defences and the 11 subgroups described in the previous sections. They follow identical methodologies to those in academia and display the same caveats. Industrial reports and whitepapers focus on 2 main themes: threat modelling and recommendations to secure the full AI lifecycle. They are based on non-empirical analysis as the main methodology. Open-source tools allow practitioners to perform adversarial attacks and threat analysis of AI systems.

Formal research in academic venues have been conducted by industry, although in significantly smaller numbers than academia. Thus, big tech companies with AI lab divisions, such as Google Deepmind (50%), Microsoft (25%), OpenAI (13%), NVIDIA, Alibaba, Huawei and IBM are sole authors of 5% of the academia research papers analysed. This low figure increases to 26% when adding the collaborations between academia and industry and smaller AI SMEs. Table 2 presents the most prolific companies in academic research found in this study.

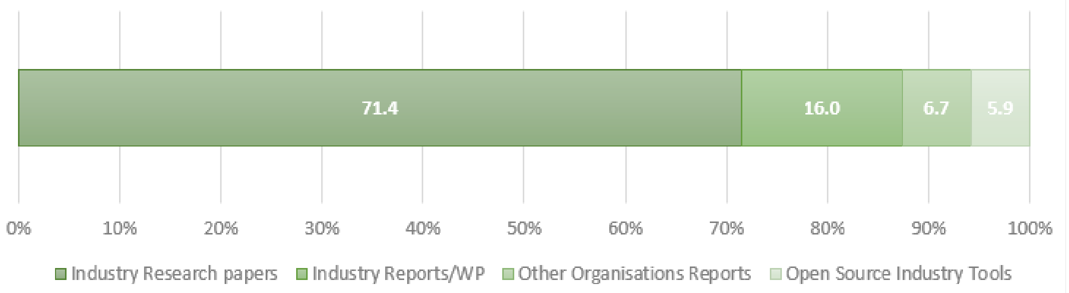

Figure 8. Split and percentage of the analysed industrial documents on research papers, white papers and industrial reports, and tools.

This research remains primarily academic research in a lab setting and at the prototype level. By classifying all research papers into attacks and defences, we can roughly estimate that 29% of the papers focus primarily on highlighting the risks, while 71% propose solutions and recommendations against those risks. However, it is reasonable to think that new and unforeseen attacks and vulnerabilities will be unveiled when the deployment of AI models is widespread, but also that their design is more complex and sophisticated than existing research. Almost no company has made public or reported attacks to deployed AI model or systems, with the exception of META [5, 53] that reported an empirical analysis by their AI red team, and Deepmind and the recent attack to the ChatGPT production model [41]. These studies are quantitative analyses based on data. The NCC group also presented a white paper on practical attacks to ML [12] which is validated on experimental data.

Table 2. Industrial reports and white papers on the Security of AI.

Company	Topic
Microsoft [35]	Industry Perspectives on Adversarial Machine Learning and gaps in defences
Microsoft [37]	Failure Modes and attacks in ML
Microsoft [38]	Threat Modeling AI/ML Systems and Dependencies
Bosch [9]	Perspective of AI security from industrial point of view and high-level recommendations for organisations.
Google [26]	A conceptual framework and security standards for building and deploying secure AI systems.
Huawei [30]	Five Challenges to AI Security, Typical AI Security Attacks and AI Security Layered Defence.
Huawei [31]	Thinking Ahead About AI Security and Privacy Protection
Nvidia [45]	Learning to Defend AI Deployments Using Simulation Environments
Arm [4]	Key challenges in Security and Privacy of AI and recommendations
Hewlett Packard [29]	Attacks and risk against AI and how to protect a business organisation
Ericsson [20]	Overview of ML-specific attack threats and defence for mobile Networks
Wiz [57]	AI security risks, existing standards and simple recommendations
NCC Group [13]	Expert witness and recommendations on the GenAI cyber security risks and its regulation.
Amazon [60]	Cloud Framework that includes a security perspective for compliance and assurance of AI systems.
Deloitte [61]	Insights into cybersecurity considerations for Generative AI.
OpenAI [62]	Safety best practices including adversarial testing and red teaming among their recommendations.

More accessible and broader guidance is disseminated by industry through white papers and industrial reports. These provide theoretical analysis of the security of AI, risks and mitigation over the AI lifecyle. While smaller in number than the academic research papers, they cover relevant aspects of the cyber security of AI that are neglected by the academic research, such as development, deployment and security in production environments. Examples are shown in Table 2. Table 3 shows available open-source tools created by industry for improving the Security of AI. Available tools focus on detecting and understanding the risks and needs for secure AI, rather than providing defensive techniques. This highlights that the majority of organisations do not have yet the right tools to secure AI Systems.

Table 3. Available open-source tools and toolboxes created by industry for improving the Security of AI.

Company	Tool	Description
IBM	Adversarial Robustness Toolbox	Open-source python library for ML security
NVIDIA	MintNV	A docker container to practice adversarial ML techniques.
Google	Cleverhans	Library focuses on adversarial (evasion) attacks and robustness.
PLOT4ai	PLOT4AI	A threat modeling library to build responsible AI.
Microsoft	Counterfit	Automation layer for assessing the security of AI systems.
HuggingFace	Safetensors	Safe and simple implementation to store and distribute ML models safely and quickly.
Linux Foundation for AI	Adversarial Robustness Toolbox	Tools that enables developers and researchers to evaluate, defend, certify and verify Machine Learning models and applications against the adversarial threats.

4.2.1 Non-governmental organisations

Non-profit organisations have also contributed to broader guidance and recommendations. Thus, MITRE corporation, working closely with industry and government, has derived an adversarial threat landscape for artificial intelligence systems (ATLAS) from a knowledge base of adversary tactics, techniques based on real-world attack observations and realistic demonstrations from AI red teams and security groups [39]. MITRE has also published a proposal for a sensible regulatory framework for AI security [40]. Theoretical analysis of the risks of adversarial attacks is also provided by the Alan Turing institute [56] and the Berryville Institute of Machine Learning [36]. The Centre for European Policy Studies (CEPS) released the CEPS report on Artificial Intelligence and Cybersecurity – Technology, governance and policy challenges [32]. which not only provides an overview of the current threat landscape of AI, but also linkages with ethical implications and existing policy. OWASP has released the AI security and privacy guide [47], as well as the top 10 security threats on ML [49] and Large language models [48].

4.3 Government and Multilateral Fora

Governments are crucial stakeholders in AI security, with contributions that range from the initial description of the guiding principles, to analysis leading to recommendations and, eventually formal regulations. This section is an attempt to provide the current state of the landscape. We do not aim to map publications by individual countries.

Documents in this section account for less than 5% of the total analysed in this study and can be classified in themes such as guidance and recommendations for vulnerability assessment and mitigation, regulations, national strategies and roadmaps. Methodology applied is based on qualitative research and non-empirical analysis.

4.3.1 Multilateral Fora Guidance

G7 Hiroshima AI Process include cyber security among their guiding principles and code of conduct for organisations [54, 55]. Similarly, the United Nations recently released guiding principles and institutional functions [2]. Other supranational bodies such as the World Economic Forum [24] have stated their view in a similar direction.

While many of the principles above relate to safety and ethical considerations, security and privacy is a fundamental part of their guiding principles.

4.3.2 Regulations

This section reports on existing formal regulation and law. Announcements and regulations currently being drafted are not included.

The European Union has recently released the EU AI Act [46], the world's first comprehensive AI law. This law however still requires addressing of technical challenges that are currently unsolved [34].

In 2022 and 2023, China approved and enforced three small regulatory measures [33] on the use of AI, Deepfake and Generative AI.

Multiple other countries have started the process to regulate AI or have announced their intention to consider regulating, but no public regulatory document has been released at the date of this report.

4.3.3 AI national strategies

44 countries have defined explicit AI national strategies [3, 16] including: Australia, Austria, Belgium, Canada, China, Colombia, Czechia, Denmark, Estonia, Finland, France, Germany, Hungary, India, Ireland, Israel, Italy, Japan, Kenya, South Korea, Lithuania, Luxembourg, Malaysia, Malta, Mexico, the Netherlands, New Zealand, Norway, Poland, Portugal, Qatar, Russia, Saudi Arabia, Serbia, Singapore, Spain, Sweden, Taiwan, Tunisia, Turkey, the U.A.E., the U.K., Uruguay, and the U.S.

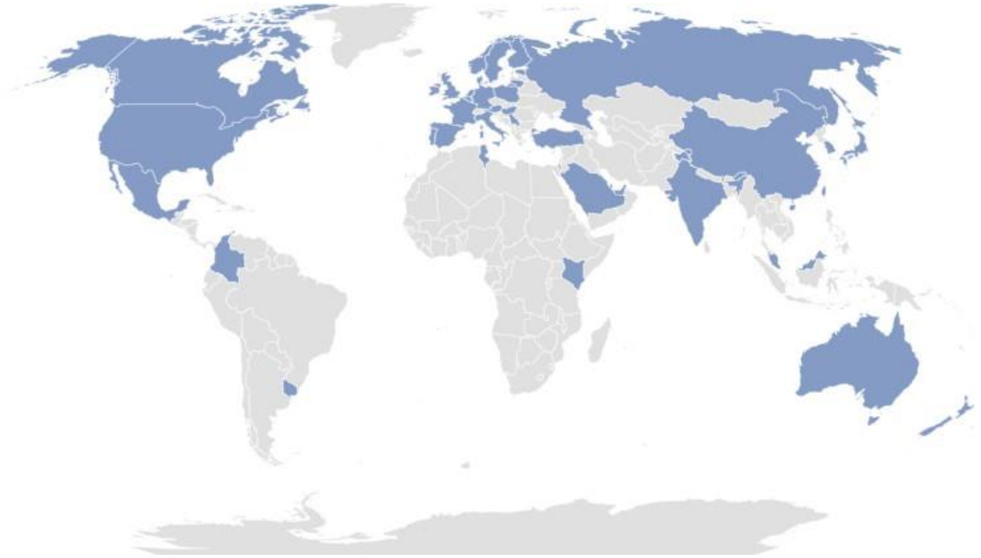

Figure 9. Countries with an explicit AI national strategy as per 2023.

The German government have released a theoretical description and definition of threats such as the AI Security Concerns in a Nutshell by BSI [23], as well as hints on defences linking to research papers. Furthermore, BSI's AI Cloud Service Compliance Criteria Catalogue [10] provides AI-specific criteria, which enable evaluation of the security of an AI service across its lifecycle. The UK's NCSC has provided a complete set of guidance documents, including Principles for the Security of Machine Learning [43] and Guidelines for Secure AI System Development [42], among others [44], containing recommendations, from design to development, for ML practitioners, IT security professionals and management. The Australian government [27] provides a similar threat description and approachable guidance on AI and how to securely engage with it. The Canadian government has released a voluntary code of conduct that includes the recommendations and measurements to be undertaken by developers and managers [28].

The AI Verify foundation, a non-profit wholly-owned subsidiary of Singapore's government, with members such as AWS, DBS Bank, Google, Meta, Microsoft, Singapore Airlines, and others, provides an AI Governance Testing Framework and Software Toolkit [25] that validates AI systems against a set of internationally recognised principles through standardised tests.

The European Union Agency for Cybersecurity (ENISA) is among the most prolific governmental bodies. ENISA R&I brief [17], provides a theoretical analysis based on the academic literature, where it summarises AI security practices. More importantly, it identifies current AI security research gaps. The ENISA Multilayer Framework for Good Cybersecurity Practices for AI [19] presents a scalable framework to guide AI stakeholders on the steps needed to secure AI systems, operations and processes following good cyber security practices in their AI. The ENISA Cybersecurity of AI and Standardisation [18] provides an overview of standards (existing, being drafted, under consideration and planned) related to the cyber security of AI.

4.4 Standards Development Organisations

A third relevant stakeholder in AI security is Standards Development Organisations (SDOs). Documents in this section account for less than 10% of the total analysed in this study, and can be classified in themes such as problem statements, guidelines, frameworks, and standards on threat analysis and risk management, and on technical recommendations for design, development and deployment of AI systems. Methodology applied is based on qualitative research and non-empirical analysis. Main SDOs have established dedicated subcommittees to focus on AI, as shown in Table 5, in order to develop standards which cover the AI lifecycle, including design, development and deployment.

Table 5. SDOs subcommittees related to the cybersecurity of AI

Organisation	Reference	Scope
ETSI	TC SAI	To develop technical specifications that mitigate threats arising from the deployment of AI. To contribute to standardisation requests, including the AI Act, Cybersecurity Resilience Act and NIS2.
ISO	JTC1 / SC27	Information security, cybersecurity and privacy protection
ISO	JTC1 /SC 42	Artificial intelligence
CEN/ CENELEC	JTC21	Identifies and adopts international standards already available and producing standardisation deliverables that address European market
NIST	ITL	Development and productive use of information technology
IEEE	AISC	Governance and practice of artificial intelligence
IEEE	PPCS	Privacy-preserving computation and Security
ITU-T	SG17	Security in the use of information and communication technologies (ICTs)
CESI	TC260/ BDSS	Information Security and Big Data Security

ETSI [21, 22], NIST, IEEE, CEN/CELEC and ISO are among the main actors in this category. CEN and CENELEC have accepted a standardisation request on Artificial Intelligence from the European Commission [63, 64]. The UK government has recently released a new portal to search AI related standards [1]. The CESI TC260 group in China published a white paper on AI security standardisation [11]. The Tables below show the most relevant current standards in AI security grouped by SDO.

These documents aim to provide information to organisations to help them better understand the consequences of security threats to AI systems, throughout their life cycles, and to describe how to detect and mitigate such threats. However, most standards focus on the design phase, with few currently for development (ISO/IEC TR 24029), deployment (ISO/IEC CD 42001:2023) or monitoring (ETSI GR SAI 005).

Table 6. ETSI Standards.

Organisation	Reference	Status	Topic
ETSI	TR 104 032 (2024-02)	Published	Traceability of AI Models
ETSI	TR 104 031 (2024-01)	Published	Collaborative AI
ETSI	GR SAI 011 (2023-06)	Published	Automated manipulation of multimedia identity representations
ETSI	GR SAI 013 V1.1.1 (2023-03)	Published	Proofs of Concepts Framework
ETSI	GR SAI 007 V1.1.1 (2023-03)	Published	Explicability and Transparency of AI Processing
ETSI	GR SAI-009 (2023-02)	Published	Artificial Intelligence Computing Platform Security Framework
ETSI	GR SAI 006 V1.1.1 (2022-03)	Published	The role of hardware in security of AI
ETSI	GR SAI 001 V1.1.1 (2022-01)	Published	AI Threat Ontology
ETSI	GR SAI 002 V1.1.1 (2021-08	Published	Data Supply Chain Security
ETSI	GR SAI 005 V1.1.1 (2021-03)	Published	Mitigation Strategy Report
ETSI	GR SAI 004 V1.1.1 (2020-12)	Published	Problem Statement
ETSI	DGR SAI-008	Pending Publication	Privacy Aspects of AI/ML Systems
ETSI	DGR SAI-010	Under Development – early draft	Traceability of AI Models
ETSI	DG RSAI-003	Under Development – stable draft	Security testing of AI

Table 7. ISO-CEN/CLC Standards.

Organisation	Reference	Status	Topic
ISO/IEC	CD TR 27563:2023	Published	Cybersecurity – Artificial Intelligence – Impact of security and privacy in artificial intelligence use cases
ISO/IEC CEN/CLC	23894:2023	Published	Information technology – Artificial intelligence – Risk management
ISO/IEC CEN/CLC	CD 42001:2023	Published	Information Technology – Artificial intelligence – Management system
ISO/IEC CEN/CLC	CD 24029-2:2023	Published	Artificial intelligence (AI) – Assessment of the robustness of neural networks – Part 2: Methodology for the use of formal methods
ISO/IEC CEN/CLC	TR 24029-1:2021	Published	Artificial Intelligence (AI) – Assessment of the robustness of neural networks – Part 1: Overview
ISO/IEC	TR 24028:2020	Published	Information technology – Artificial intelligence – Overview of trustworthiness in artificial intelligence
ISO/IEC	AWI 27090	Under development	Cybersecurity - Artificial intelligence – Guidance for addressing security threats and failures in artificial intelligence systems
ISO/IEC	AWI 27091.2	Under development	Cybersecurity and Privacy - Artificial Intelligence - Privacy protection
CEN/CLC	JT021009	Under Drafting	AI Risks - Check List for AI Risks Management
CEN/CLC	JT021024	Under Drafting	AI Risk Management
CEN/CLC	JT021029	Preliminary	Technical solutions to address AI specific vulnerabilities

Table 8. Other Standards.

Organisation	Reference	Status	Topic
IEEE	P2986	Published	Recommended Practice for Privacy and Security for Federated Machine Learning
IEEE	P3156	Published	Standard for Requirements of Privacy preserving Computation Integrated Platform
IEEE	P3169	Published	Standard for Security Requirement of Privacy-Preserving Computation
IEEE	70022022	Published	IEEE Standard for Data Privacy Process
IEEE	P7012	Published	Standard for Machine Readable Personal Privacy Terms
NIST	AI.100-2	Published	Artificial Intelligence Risk Management Framework
NIST	IR.8269	Published	A Taxonomy and Terminology of Adversarial Machine Learning
NIST	IR.8330	Published	User Perceptions of Smart Home Privacy and Security
ITU-T	XSTR-SEC-AI	Published	Guidelines for security management of using artificial intelligence technology
ITU-T	TR.SE-AI	Under development	Technical Report: Security Evaluation on Artificial Intelligence Technology in ICT

5 CONCLUSIONS

Despite the widespread development of ML applications, their public deployment in the real world is still in its early stages. Once this deployment is completed, AI and ML models will be exposed to security threats that may compromise the model, its data or its functionality. Concerns regarding the security of ML have been growing, initiated by academia and followed by industry, and more recently alerted by the emergence of GenAI, by regulatory bodies. This study comprehensively surveys, classifies and quantifies the security aspects of ML systems, including all stages of their life cycle.

We now describe the main conclusion derived from our study and supported by the evidence stated in previous sections:

1. The research field of AI Security is still nascent and has developed for the last 5 years. Despite tens of thousands of papers in the field, the research trend indicates new vulnerabilities and more effective and efficient defences are being developed and proposed, leading to new solutions still under refinement.
2. Current research solutions are still either limited or impractical. A large set of industrial solutions focus on detecting and understanding the risks and vulnerabilities of AI. Emerging industrial solutions for defence focus on data input/output monitoring and data firewalls to prevent breaches, while waiting for future algorithmic defences.
3. Current academic research has been empirically evaluated using academic and public datasets but under laboratory settings. No evaluation in the wild has been found. It is reasonable to think that new and unforseen attacks and vulnerabilities will be unveiled when the deployment of AI models is widespread. Therefore, we predict that most of the existing solutions will underperform and expose the integrity and privacy of AI solutions when made available to millions of users. It is therefore imperative to perform empirical studies deploying Secure AI solutions in the wild to discover new attacks leading to new solutions. The use of industrial testbeds, digital twins and red-teaming could better approximate this in-the-wild behaviour.
4. Academic research mostly overlooks development, deployment and monitoring of the AI lifecy- cle. This gap is being currently addressed by standardisation companies, government agencies, non-profit organisations and industry, by providing more accessible and broader guidance for AI developers throughout the full AI lifecycle. While in most cases these documents have contributions from experts in academia, an empirical investigation based on quantitative analysis is recommended. For instance, through research and empirical validation of the proposed frame-works, red teaming as well as continuous updates of the standards according to the continuous refinement of research solutions.
5. National and supranational regulatory bodies have developed AI strategies, recommendations and guidance, and regulations and law are just starting to emerge, but they still need to broaden to consider the technical advancements and limitations. Moreover, academic and industrial research will be required to address the consequent unsolved technical challenges emerging from those regulations.
6. Multiple standards organisation are currently developing a set of comprehensive standards on the cyber security of AI. Existing standards provide information to organisations to help them better understand the risk and security threats to AI systems, and the most recent ones describes how to detect and mitigate such threats. mostly on the design phase. Further standards under development will focus on the following AI lifecycle stages (development, deployment and monitoring).

5.1 Limitations of this study

The study is based on public sources, which may not fully reflect the real capability of industry, the level of deployment of Secure AI solutions, and the incidence of attacks against existing AI solutions.

Time constraints restricted the academic literature search to two databases, potentially excluding relevant materials from other sources. Additionally, the inclusion/exclusion criteria narrowed down the selection of pertinent literature, focusing on English publications between 2013 and 2023. It focuses on the top-tier conferences both in cyber security and AI/ML but, while this guarantees the most relevant sources and actors, it may overlook a substantial body of research. This may particularly affect the conclusion regarding academic input on guidance and recommendations for developers, engineers and for deployment. This was, however, mitigated with an additional search in Google.

REFERENCES

Disclaimer: This section only presents the references explicitly cited in this report. References to academic papers used in our analysis but not cited in the document to support finding are not included for brevity, given our intended main audience (HMG analysts and policy makers rather than academics)

[1] Ai standard hub to compile. https://aistandardshub.org/ai-standards-search/, *Accessed on 01-02-2024*. online web resource.

[2] U. N. A. advisory Board. Un report: Governing ai for humanity. https://www.un.org/sites/un2.un.org/files/ai_advisory_body_interim_report.pdf, *Accessed on 01-02-2024*. online web resource.

[3] Aiethicist. National stategies. https://www.aiethicist.org/national-strategies, *Accessed on 01-02-2024*. online web resource.

[4] ARM. Security: Teh first of six key challenges. https://interactive.arm.com/story/building-trustworthy-ai/page/3?utm_source=linkedin&utm_medium=social&utm_campaign=2022_client_mk04_arm_na_na_awa&utm_content=whitepaper, *Accessed on 01-02-2024*. online web resource.

[5] B.Dolhansky, R.Howes, B. andN.Baram, and C.C.Ferrer. The deepfake detection challenge (dfdc) previewdataset,. *arXiv*, 2019.

[6] U. Berkeley. Ai's redress problem. https://www.standict.eu/sites/default/files/2021-02/etsi_wp34_Artificial_Intellignce_and_future_directions_for_ETSI.pdf, *Accessed on 01-02-2024*. online web resource.

[7] U. Berkeley. Toward ai security. https://cltc.berkeley.edu/wp-content/uploads/2019/02/CLTC_Cussins_Toward_AI_Security.pdf, *Accessed on 01-02-2024*. online web resource.

[8] E. Biasin, E. Kamenjasevic, and K. R. Ludvigsen. Cybersecurity of ai medical devices: risks, legislation, and challenges, 2023.

[9] Bosch. Ai white paper. https://25908683.fs1.hubspotusercontent-eu1.net/hubfs/25908683/BoschAIShield_AI%20Security_Whitepaper.pdf, *Accessed on 01-02-2024*. online web resource.

[10] BSI. Ai cloud service compliance criteria catalogue (aic4). https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/CloudComputing/AIC4/AI-Cloud-Service-Compliance-Criteria-Catalogue_AIC4.pdf?__blob=publicationFile&v=4, *Accessed on 01-02-2024*. online web resource.

[11] CESI. Artificial intelligence security standardization. https://cset.georgetown.edu/wp-content/uploads/t0121_AI_security_standardization_white_paper_EN.pdf, *Accessed on 01-02-2024.* online web resource.

[12] N. G. Chris Anley, Chief Scientist. Practical attacks on machine learning systems. https://research.nccgroup.com/2022/07/06/whitepaper-practical-attacks-on-machine-learning-systems/, *Accessed on 01-02-2024.* online web resource.

[13] N. G. Chris Anley, Chief Scientist. Security of large language models (llms) - uk parliament. https://www.nccgroup.com/us/newsroom/security-of-large-language-models-llms-uk-parliament-invites-ncc-group-s-chris-anley-as-expert-witness/, *Accessed on 01-02-2024.* online web resource.

[14] M. Comiter. Attacking artificial intelligence:ai's security vulnerability and what policymakers can do about it. *Belfer Center for Science and International Affairs*, 2019.

[15] M. . Company. The state of ai in 2020. https://www.mckinsey.com/business-functions/quantumblack/our-insights/global-survey-the-state-of-aiin-2020, *Accessed on 01-02-2024.* online web resource.

[16] J. S. Denford, G. S. Dawson, and K. C. Desouza. cluster analysis of national ai strategies. https://www.brookings.edu/articles/a-cluster-analysis-of-national-ai-strategies/#:~:text=As%20with%20our%20recent%20blog,%2C%20the%20Netherlands%2C%20New%20Zealand%2C, *Accessed on 01-02-2024.* online web resource.

[17] ENISA. Artificial intelligence and cybersecurity research. https://www.enisa.europa.eu/publications/artificial-intelligence-and-cybersecurity-research/@@download/fullReport, *Accessed on 01-02-2024.* online web resource.

[18] ENISA. Cybersecurity of ai and standardisation. https://www.enisa.europa.eu/publications/cybersecurity-of-ai-and-standardisation, *Accessed on 01-02-2024.* online web resource.

[19] ENISA. A multilayer framework for good cybersecurity practices for ai. https://www.enisa.europa.eu/publications/multilayer-framework-for-good-cybersecurity-practices-for-ai, *Accessed on 01-02-2024.* online web resource.

[20] Ericcson. Ai security in mobile networks. https://www.ericsson.com/en/blog/2020/10/ai-security-mobile-networks, *Accessed on 01-02-2024.* online web resource.

[21] ESTI. Artificial intelligence and future directions for etsi. https://www.standict.eu/sites/default/files/2021-02/etsi_wp34_Artificial_Intelligence_and_future_directions_for_ETSI.pdf, *Accessed on 01-02-2024.* online web resource.

[22] ESTI. Securing artificial intelligence. https://www.etsi.org/committee/2312-sai, *Accessed on 01-02-2024.* online web resource.

[23] G. F. O. for Information Security (BSI). Ai security concerns in a nutshell. https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/KI/Practical_Al-Security_Guide_2023.pdf?__blob=publicationFile&v=5, *Accessed on 01-02-2024.* online web resource.

[24] W. E. Forum. Why we need cybersecurity of ai: ethics and responsible innovation. https://www.weforum.org/agenda/2023/12/cybersecurity-ai-ethics-responsible-innovation/, *Accessed on 01-02-2024.* online web resource.

[25] A. Foundation. Build trust with ai verify. https://aiverifyfoundation.sg/ai-verify-foundation/, *Accessed on 01-02-2024.*

[26] Google. Google's secure ai framework. https://safety.google/cybersecurity-advancements/saif/, *Accessed on 01-02-2024*. online web resource.

[27] A. government. Introducing artificial intelligence. https://www.cyber.gov.au/resources-business-and-government/governance-and-user-education/governance/an-introduction-to-artificial-intelligence, *Accessed on 01-02-2024*. online web resource.

[28] Government of Canada. Voluntary Code of Conduct on the Responsible Development and Management of Advanced Generative AI Systems. https://ised-isde.canada.ca/site/ised/en/voluntary-code-conduct-responsible-development-and-management-advanced-generative-ai-systems, *Accessed on 01-02-2024*. online web resource

[29] HPE. What is ai security? https://www.hpe.com/us/en/what-is/ai-security.html, *Accessed on 01-02-2024*. online web resource.

[30] Huawei. Ai security white paper. https://www-file.huawei.com/-/media/corporate/pdf/trust-center/ai-security-whitepaper.pdf, *Accessed on 01-02-2024*. online web resource.

[31] Huawei. Thinking ahead about ai security and privacy protection. https://www.huawei.com/en/news/2019/9/huawei-thinking-ahead-ai-security-privacy-protection-whitepaper, *Accessed on 01-02-2024*. online web resource.

[32] A. Intelligence, g. Cybersecurity – Technology, and policy challenges. Ai security in mobile networks. https://www.ceps.eu/wp-content/uploads/2021/05/CEPS-TFR-Artificial-Intelligence-and-Cybersecurity.pdf, *Accessed on 01-02-2024*. online web resource.

[33] A.-J. Kachra. Making sense of china's ai regulations. https://www.holisticai.com/blog/china-ai-regulation, *Accessed on 01-02-2024*. online web resource.

[34] R. P. Kalodanis K. and and A. D. European artificial intelligence act: an ai security approach. *Information and Computer Security*, 2023.

[35] R. S. S. Kumar, M. Nystro¨m, J. Lambert, A. Marshall, M. Goertzel, A. Comissoneru, M. Swann, and S. Xia. Adversarial machine learning – industry perspectives. *arXiv*, 2021.

[36] G. McGraw, H. F. V. Shepardson, and R. Bonett. An architectural risk analysis of machine learning systems: Toward more secure machine learning. *Berryville Institute of Machine Learning.*, 2019.

[37] Microsoft. Failure modes in machine learning. https://docs.microsoft.com/en-us/security/engineering/failure-modes-in-machine-learning, *Accessed on 01-02-2024*. online web resource.

[38] Microsoft. Threat modeling ai/ml systems and dependencies. https://docs.microsoft.com/en-us/security/engineering/threat-modeling-aiml, *Accessed on 01-02-2024*. online web resource.

[39] MITRE. Mitre atlas. https://atlas.mitre.org/, *Accessed on 01-02-2024*. online web resource.

[40] MITRE. A sensible regulatory framework for ai security. https://www.mitre.org/sites/default/files/2023-06/PR-23-1943-A-Sensible-Regulatory-Framework-For-AI-Security_0.pdf, *Accessed on 01-02-2024*. online web resource.

[41] M. Nasr, N. Carlini, J. Hayase, M. Jagielski, A. F. Cooper, D. Ippolito, C. A. Choquette-Choo, E. Wallace, F. Tramèr, and K. Lee. Scalable extraction of training data from (production) language models. *arXiv*, 2023.

[42] NCSC. Guidelines for secure ai system development. https://www.ncsc.gov.uk/collection/guidelines-secure-ai-system-development, *Accessed on 01-02-2024*.

[43] NCSC. Principles for the security of machine learning. https://www.ncsc.gov.uk/collection/machine-learning, *Accessed on 01-02-2024.*

[44] NCSC. Thinking about the security of ai systems. https://www.ncsc.gov.uk/blog-post/thinking-about-security-ai-systems, *Accessed on 01-02-2024.*

[45] NVIDIA. Learning to defend ai deployments using an exploit simulation environment. https://developer.nvidia.com/blog/learning-to-defend-ai-deployments-using-an-exploit-simulation-environment/, *Accessed on 01-02-2024.* online web resource.

[46] C. of the European Union. Artificial intelligence act. https://data.consilium.europa.eu/doc/document/ST-5662-2024-INIT/en/pdf, *Accessed on 01-02-2024.* online web resource.

[47] OWASP. Owasp ai security and privacy guide. https://owasp.org/www-project-ai-security-and-privacy-guide/, *Accessed on 01-02-2024.* online web resource.

[48] OWASP. Owasp top 10 for llm. https://llmtop10.com/, *Accessed on 01-02-2024.* online web resource.

[49] OWASP. Wasp machine learning security top 10. https://mltop10.info/, *Accessed on 01-02-2024.* online web resource.

[50] R. S. Sangwan, Y. Badr, and S. M. Srinivasan. Cybersecurity for ai systems: A survey. *Journal of Cybersecurity and Privacy*, 3(2):166–190, 2023.

[51] H. K. School. Attacking artificial intelligence: Ai's security vulnerability and what policymakers can do about it. https://www.belfercenter.org/publication/AttackingAI, *Accessed on 01-02-2024.* online web resource.

[52] S. Shukla, I. Parada, and K. Pearlson. Trusting the needle in the haystack: Cybersecurity management of ai systems. *Advances in Information and Communication*, 2022.

[53] T. Simonite. Facebook's 'red team' hacks its own ai programs. https://www.wired.com/story/facebooks-red-team-hacks-ai-programs/5, *Accessed on 01-02-2024.* WIRED.

[54] G. H. Summit. Hiroshima process international code of conduct for organizations developing advanced ai systems. https://www.mofa.go.jp/files/100573473.pdf, *Accessed on 01-02-2024.*

[55] G. H. Summit. Hiroshima process international guiding principles for organizations developing advanced ai system. https://www.mofa.go.jp/files/100573471.pdf, *Accessed on 01-02-2024.*

[56] U. The Alan Turing Institute. Adversarial ai coming of age or overhyped? https://www.turing.ac.uk/sites/default/files/2023-11/ai_in_cybersecurity.pdf, *Accessed on 01-02-2024.* online web resource.

[57] Wiz. Ai security explained: How to secure ai. https://www.wiz.io/academy/ai-security, *Accessed on 01-02-2024.* online web resource.

[58] D. Zhao and A. Strotmann. *Analysis and Visualization of Citation Networks*, volume 7. 02 2015.

[59] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus. Intriguing properties of neural networks. *ICLR*, 2013.

[60] Amazon, AWS Cloud Adoption Framework for Artificial Intelligence, Machine Learning, and Generative AI. https://docs.aws.amazon.com/whitepapers/latest/aws-caf-for-ai/security-perspective-compliance-and-assurance-of-aiml-systems.html *Accessed on 01-02-2024.* online web resource.

[61] Deloitte, Safeguarding Generative Artificial Intelligence with Cybersecurity Measures, https://www2.deloitte.com/content/dam/Deloitte/in/Documents/

risk/in-ra-safeguarding-generative-artificial-intelligence-noexp.pdf, *Accessed on 01-02-2024.* online web resource.

[62] OpenAI, Safety Best Practises, https://platform.openai.com/docs/guides/safety-best-practices, *Accessed on 01-02-2024.* online web resource.

[63] CEN/CELEC, Response to the EC White Paper on AI, https://www.cencenelec.eu/media/CEN-CENELEC/Areas%20of%20Work/Position%20Paper/cen-clc_ai_fg_white-paper-response_final-version_june-2020.pdf, *Accessed on 01-02-2024.* online web resource.

[64] CEN/CELEC, Road Map on Artificial Intelligence (AI), https://www.cencenelec.eu/media/CEN-CENELEC/AreasOfWork/CEN-CENELEC_Topics/Artificial%20Intelligence/Quicklinks%20General/Documentation%20and%20Materials/cen-clc_fgreport_roadmap_ai.pdf, *Accessed on 01-02-2024.* online web resource.

[65] A Oseni, N Moustafa, H Janicke, P Liu, Z Tari, A Vasilakos. Security and privacy for artificial intelligence: Opportunities and challenges, arXiv: 2102.04661, 2021

[66] Q. Liu, P. Li w. Zhao, W. Cai, A Survey on Security Threats and Defensive Techniques of Machine Learning: A Data Driven View, IEEE Access, vol. 6, 2018