Cyber Security for AI Recommendations

A Study of Recommendations to Address Cyber Security Risks to AI

Dr. Peter Garraghan

February 2024

PUBLICATION HISTORY/VERSION

Version 1.0	February 9th 2024

Author

Dr. Peter Garraghan
CEO & CTO, Mindgard
Professor, Lancaster University

For more information, please visit mindgard.ai

EXECUTIVE SUMMARY

The Department for Science, Innovation, and Technology (DSIT) commissioned Mindgard to conduct a systematic study to identify recommendations linked to addressing cyber security risks to Artificial Intelligence (AI). We used a systematic search method to review data sources across academia, technology companies, government bodies, cross-sector initiatives, news articles, and technical blogs to identify various recommendations and evidence of cyber risks against AI. The review also examined common themes and knowledge gaps.

A comprehensive search of relevant sources published between 1 January 2020 and 12 January 2024 (with notable exceptions for fundamental academic works) was conducted as the basis for this review. A total of 67 publications were identified that described 45 unique technical and general recommendations for addressing cyber security risks in AI. We found sufficient evidence indicating that many of the reported cyber security risks to AI strongly justify the need to identify, create, and adopt new recommendations to address them. However, we also discovered several gaps within existing knowledge. Many of the recommendations for AI are based on established cyber security practises and various conventional cyber security recommendations are directly or indirectly applicable to AI. However, many recommendations are derived from few unique data sources and there are limited empirical studies of security vulnerabilities in AI used in the production of cyber attacks. There is also a lack of information on how to enact recommendations described.

TABLE OF CONTENTS

1. Introduction **5**

1.1 Aims and Objectives 5

1.2 Report Organisation 5

2. Methodology **6**

2.1 Recommendation Types 6

2.2 Data Collection 7

3. Discussion of Recommendation Findings **9**

4. Recommendations **11**

4.1 Technical Recommendations 11

4.1.1 Design 11

4.1.2 Development 12

4.1.3 Deployment 13

4.1.4 Operation & Maintenance 14

4.2 General Recommendations 14

4.2.1 Company Practises, Policies, Governance & Security Hygiene 14

4.3 Mapping Recommendations to Reported Attacks against AI 16

5 Reported Security Vulnerabilities within AI **18**

5.1 Data Collection 18

5.2 Reported AI Security Vulnerabilities 18

6. Conclusions **20**

7. Definitions **21**

7.1 Artificial Intelligence (AI) 21

7.2 AI Security (Cyber Security for AI) 22

References **24**

LIST OF FIGURES AND TABLES

Table 1. Summary of data sources identified recommendations to address cyber security for AI 11

Appendix 1. Overview of Reported Security Vulnerabilities within AI Deployed in Production 30

1. Introduction

1.1 Aims and Objectives

DSIT commissioned Mindgard to conduct a systematic study to identify recommendations linked to addressing cyber security risks to Artificial Intelligence (AI). This study encompassed collating, reporting, and mapping known recommendations that have been reported or demonstrated to enhance the cyber security of AI models, systems, and data. Our precise objectives are as follows:

- Identify recommendations for addressing cyber security risks to AI, collated from reviewing evidence across industry, government, and academia.
- Analyse and discuss the emergence of common themes, trends, and knowledge gaps within existing recommendations for AI security, encompassing both technical and general recommendations across different phases of the AI development lifecycle.
- Review reported cyber attacks against AI, which have been included to provide context on how existing recommendations are currently being deployed or positioned in practise.

The purpose of this study is not to provide an exhaustive list of every published data source on recommendations linked to addressing AI cyber security risks, given the large volume of academic works published within this space (now in the regions of thousands of publications within the past two years). Instead, this study's primary objective is to identify all the different types of recommendations proposed, and where possible, determine their effectiveness when used in practise.

1.2 Report Organisation

This report is structured as follows. Section 2 explains the methodology process which was used to conduct the study. Section 3 provides a description of all the identified recommendations linked to addressing cyber security risks to AI, categorised by their type and deployment within the AI development lifecycle. Section 4 sets out some overarching findings from the recommendations and sources identified from the review as well as some key trends and knowledge gaps. Section 5 presents an overview of reported cyber attacks against AI to provide empirical context on the applicability of identified recommendations. Section 6 presents the study conclusions and Section 7 provides definitions of core terminology within AI security leveraged for this study.

2. Methodology

Overview: The methodology used for this study involved:

(i) Performing an assessment of existing literature and the wider AI security area.
(ii) Defining the main types of recommendations focused on addressing cyber security risks to AI.
(iii) Determining the method, scope and criteria for identifying data sources containing relevant recommendations.
(iv) Analysing and qualifying data sources for recommendations to be included within the study.
(v) Categorising and analysing recommendations to understand the current landscape, discuss findings of interest, and to identify knowledge gaps.

Such an approach was taken due to the current maturity of the AI security area, which has only recently garnered increased attention due to the rise in Generative AI and Large Language Models (LLMs). Moreover, given activity within this area resides at the intersection between Artificial Intelligence and Cyber Security, the author deemed it would be suitable to first define the main types of recommendations at the outset, which would then inform the subsequent literature review.

2.1 Recommendation Types

Within this study, we define a *recommendation* as a technique, process, method, or strategy that reduces the cyber security risk of AI models, data, or systems. Given that such a definition is relatively broad, recommendations have been further sub-divided into two types: *technical* and *general*.

Technical Recommendation: Technology-focused approaches to mitigate cyber security risks in AI. Such recommendations predominantly entail altering the software, hardware, data, or network access of a computer system that runs the AI, that subsequently results in reduced cyber security risk when exposed to an AI cyber attack. Technical recommendations encompass approaches that (i) are derived from specific, documented scenarios in *production* (an AI model, system, or service provisioned to achieve an organisation's operational goals); (ii) whose feasibility has been demonstrated via experimental means as a proof concept within laboratory conditions (e.g. a University research lab); or (iii) are hypothesised as being potentially effective, based on postulation or expert opinion from researchers or practitioners.

General Recommendation: Conceptual frameworks for mitigating cyber security risks in AI. These recommendations entail establishing organisational practise, company policies, governance, and security practises ('security hygiene'). General recommendations encompass recommendations that are described from conceptual frameworks across academia, industry, and government – all of which are typically based on the expertise and learnt experience of framework contributors.

It is worth noting that there exists potential overlap between these two recommendation types. For example, descriptions of general recommendations are often derived – or provide explicit examples of – technical recommendations demonstrated or postulated to reduce cyber security risk in AI. In such scenarios, descriptions of conceptual approaches with examples of technical methods are designated as general recommendations, whereas detailed technical explanations ('technology first') approaches are categorised as technical recommendations.

2.2 Data Collection

The author is aware that with the rapid adoption and evolution of AI, the number of reported security incidents, working groups, research publications, and frameworks describing recommendations will only increase. For the purposes of this review, the report examined English-language sources that were published from 1 January 2020 and 12 January 2024 (with exceptions for a few research papers published from 2014 onwards, due to the creation or popularisation of a technical recommendation method that are frequently referenced within data sources). Moreover, data collection and analysis provided within this study has been completed based on the author's current knowledge of the AI security field as of 12th January 2024. Documents published by standards development organisations were not included within this study. This decision was taken because various standards are actively under discussion/debate (i.e. have not been finalised), and primarily only accessible behind paywalls.

An overview of the various research activities and types of sources used for the review are set out below.

Research publications: This included peer-reviewed academic and industry research papers. The data was identified by using various keyword search terms 'Adversarial ML', 'AI security', 'cyber security for AI', 'AI cyber risk') on several research publication search engines, such as Google Scholar, ACM digital library. The author also examined material cited within publications to identify additional sources. Given the large volume of papers that could be included within this study, an active choice was made to prioritise recommendations that:

- Provided sufficient diversity in approach (i.e. a single exemplar that captures the key conceptual underpinning of a recommendation, instead of reporting multiple instances that are derivatives of the same approach);
- Where possible, evaluated AI models, systems, and services used in production;
- Were conducted through empirical means, (such as recommendations that were effectively measured within laboratory conditions); and
- Were perceived to have a notably high impact in terms of awareness and effectiveness, as judged by the author's own expertise within AI security research.

The number of research publication citations was not a qualifying factor in selection, given that different research fields range considerably in terms of community size and publication output.

Technology companies: This included recommendations made by multi-national technology companies. Data sources that were studied include published AI security frameworks, AI policy documentation, and blog articles. These data sources were discovered via multiple channels: (i) web search results from Google and Microsoft Bing using keywords 'Secure AI Frameworks', 'AI Security', 'Securing AI policies', 'Security for AI', and 'Techniques to defend AI'; and (ii) directly navigating to websites of companies known to leverage/develop AI at scale. Press releases were not included within this data source, given the information provided was found to be insufficient for clearly articulating recommendations (and in most cases, would reference a more detailed data source).

Government institutions: This included reports, frameworks, and recommendations created by government bodies or institutions. This data was collated from material referenced from cross-sector initiatives, as well as searching web repositories of various government organisations across multiple countries, such as from the UK, US, EU, Germany, Japan, Singapore, Australia, and New Zealand. Reports that were reviewed to ascertain recommendations included topics related to 'AI security', 'ML security', 'security of AI systems', 'AI safety', 'AI auditing', 'Trustworthy AI', 'AI ethics', and 'Developing AI systems'.

Cross-sector initiatives: This included recommendations derived from online cross-sector working groups across academia, industry, government, and other interested parties in AI security. Initiatives were found through navigating established cyber security/risk frameworks (MITRE, OWASP, etc.). This identified case studies, news articles, technical blogs, and GitHub repositories.

Using this methodology and criteria for qualifying data sources, we were able to identify 67 unique data sources that were categorised as either technical or general recommendations. To our knowledge, this is the first study to identify, collate, and describe recommendations linked to addressing cyber risks to AI that captures a broad set of perspectives across academia, industry, and government. Importantly, this is also the first publication to identify trends and knowledge gaps within existing recommendations and the report has therefore strived to determine their respective effectiveness in practise.

3. Recommendations

This section provides a breakdown of all unique recommendations discovered for addressing cyber security risks to AI. From the literature search conducted, we were able to identify 45 unique technical or general recommendations derived from 12 multi-national AI-driven technology companies, 11 government bodies, 2 cross-sector AI security frameworks, 30 academic publications, and 1 news article (see Table 1). As stated in Section 1.1, academic publications noted within this study have been used to evidence the existence and/or effectiveness of a recommendation, and thus should not be viewed as an exhaustive list of all works available.

	Total number of organisations	**Total number of unique publications**
Academic	30	30
Technology Companies	12	14
Government	11	18
Cross-group initiatives	2	4
News articles, blogs	1	1
Total	**56**	**67**

Table 1. Summary of data sources

Each recommendation has been categorised based on their alignment to AI security frameworks and supporting literature, where applicable. Technical recommendations have been further categorised based on author's interpretation of the phases of the AI lifecycle (e.g. design, development, deployment, operation and maintenance). These recommendations have also been contextualised based on the type of cyber attack that they are envisaged to mitigate.[2]

3.1 Technical Recommendations

3.1.1 Design

Design refers to recommendations that alter the technical design and development of an AI deployment prior to its training and deployment (e.g. technical mechanisms modifying AI operation).

Model Distillation [OWASP, 2024; Vassilev, 2023; BSI1, 2023; **Dong, 2021] (Evasion, Poisoning):** A technique in AI whereby a smaller model (student) is trained to replicate the behaviour of a larger, more complex model (teacher). The goal is to transfer the knowledge and capabilities of the larger model to the smaller one, providing defence through abstraction. This technique entails softening the outputs of the teacher model trained upon the student model to increase its resilience by making the model less sensitive to perturbations. Although this technique has been proposed as a recommendation [Hinton, 2014], there exists evidence that this approach exhibits various weaknesses [Carilini, 2016].

Model Distribution [OWASP, 2024; MITRE, 2024] (Extraction, Evasion, Poisoning): Deploying AI models to edge devices can increase the attack surface of the system. It has been recommended to deploy and process AI models within the cloud to reduce the level of access for an attacker [Hosseini, 2017; ENISA, 2021], although this approach would result in performance degradation as a result of increased network latency.

[2] See Section 7 for definitions of relevant terminology and definitions

Ensemble Methods [OWASP, 2024; Vassilev, 2023; MITRE, 2024; BSI, 2022] (Extraction, Evasion, Backdoor, Poisoning): Ensemble methods are when multiple AI models are used in conjunction to achieve better results. It has been recommended to use ensemble models to perform inference as it increases adversarial robustness. Specific attacks may effectively evade a single or family of AI models, however is ineffective against others. Using multiple models allows AI model predictions to be verified via consensus and adjudication, allowing models to verify their respective outputs. [OWASP, 2024; Zahalka, 2023].

3.1.2 Development

This section covers recommendations linked to preparation, training, and the development of an AI model. This includes technical approaches pertaining to preparing and processing training data, as well as during the AI training process itself.

Sanitise Training Data [Vassilev, 2023; MITRE, 2024; BSI1, 2023; Microsoft, 2022] (Poisoning): Detect, remove, and remediate poisoned training data, as training data should be sanitised prior to training. Implement a filter to limit ingested training data. Example approaches include establishing a content policy to remove unwanted content, and utilising anomaly detection sensors to inspect data distribution daily and alert on abnormal variation.

Validate ML Model [MITRE, 2024; ENISA, 2021; BSI1, 2023; BSI, 2022] (Poisoning): Validate that AI models perform as intended by testing for backdoor triggers or adversarial bias. It also involves monitoring an AI model for changes in its behaviour and responsiveness during training, because unexpected behaviour in model performance may indicate data tampering and poisoning. The AI model should regularly be evaluated to determine whether it has been poisoned, for example evaluating the model via input perturbation to observe and measure changes to prediction [Gao, 2020]. It is noted that these recommendations are all derived from academic publications.

Input Restoration [MITRE, 2024] (Evasion): Input restoration adds an extra layer of unknowns and randomness when an adversary evaluates the input and output relationship of the AI model. This technique reduces the effectiveness of an attacker by preventing or reversing adversarial perturbations.

Overfitting Detection [OWASP, 2024; BSI, 2022] (Inversion): Overfitting can be prevented by ensuring the AI model is kept sufficiently small, configured with fair distribution of training data, and suitably set training parameters (number of data points, iterations, etc). This ensures the AI model does not store extreme levels of detail of individual training samples.

Reject-On-Negative-Impact [Vassilev, 2023; BSI, 2022] (Poisoning): Adversarial examples are identified via testing the impacts of examples upon classification performance. Examples that produce high error rates in classification are removed from the training set. This is known as Reject-on-Negative-Impact [Li, 2024]. Rather than attempting to detect poisoned data, Robust Statistics use constraints and regularisation techniques to reduce potential distortions of the learning model that are caused by poisoned data.

Differential Privacy [OWASP, 2024; BSI1, 2023; BSI, 2022; Microsoft, 2022; Dong, 2021] (Extraction, Inversion, Membership Inference): A mathematical framework for ensuring the privacy of individuals within datasets. Differential Privacy ensures that model outputs do not reveal additional information about an individual record included within the training data.

Homomorphic Encryption [Vassilev, 2023; BSI, 2022; HHS, 2021] (Inversion): Encrypts data in a form so that a neural network cannot operate without data decryption. This protects the privacy of each individual input; however, it also introduces computational performance overhead and limits the set of arithmetic operations to those supported by Homomorphic Encryption.

Model Watermarking [MITRE, 2024; World, 2024; Google, 2023] (Data Leakage): Watermarking adds hidden patterns to training data prior to model training. This enables an AI model developer to verify whether leaked data (following an attack) originated from their own model. This approach is typically leveraged for AI generated content for policy enforcement, attribution, legal recourse, and deterrence.

Model Hardening [OWASP, 2024; Vassilev, 2023; MITRE, 2024; NCSC, 2023; ENISA, 2021; BSI1, 2023; BSI, 2022; BSI2, 2023; Microsoft, 2022; Amazon, 2023; Leslie, 2019] (Evasion): This technique strives to make AI models more robust to adversarial inputs via adversarial training or network distillation [MITRE, 2024; Wu, 2017]. Examples include using randomisation to inject noise during training to enhance resilience to evasion attacks (especially triggered by subtle perturbations) [Bai, 2021], Gradient Masking [Samangouei, 2018; Bunzel, 2023], and Feature Squeezing [Xu, 2019].

Code Signing [MITRE, 2024] (Model Backdoors, Poisoning): Enforces binary and application integrity with digital signature verification to prevent untrusted code execution. Attackers can embed malicious code in AI software, frameworks, libraries, or models.

Use Multi-Modal Sensors [MITRE, 2024] (Physical Domain Attacks): The execution time, power usage, temperature, etc., can be used to evaluate whether an AI deployment is operating as expected. Incorporating multiple sensors covering varying perspectives and modalities can enable early warning and prevention if unexpected behaviour arises from physical system-level attacks.

AI Security Testing [Reber, 2023; AI Verify, 2024; G7, 2023; NCSC, 2023] (All Attack Types): Launch cyber attacks against your AI model, system, and data in a controlled environment to test and measure its susceptibility to different attacks. This enables organisations to evidence cyber risks against their AI system, identify and remediate security vulnerabilities, as well as evaluate their detection and response capabilities.

Vulnerability scanning [Vassilev, 2023; MITRE, 2024; NCSC, 2023; Reber, 2023] (All Attack Types): Vulnerability scanning is used to find potentially exploitable software vulnerabilities so that they can be remediated. File formats such as pickle, commonly used within PyTorch, can contain exploits that can be used to perform arbitrary code execution. Therefore, it is recommended to use Pickle scanning tools or a safer model format [Martin, 2023]. HuggingFace introduced the 'SafeTensors' file format for PyTorch models which removed the dependency on Pickle as well as for other frameworks. There exist other safe model formats for model storing including H5 (Tensorflow), Protobuf (ONNX), and NumPy (npy, npz) [HuggingFace2, 2024].

3.1.3 Deployment

The following recommendations entail leveraging technical approaches that are performed during deploying an AI model prior to its operation in production.

Leverage Virtualisation [ICO, 2020] (Inversion, Backdoor): Technologies, such as Virtual Machines (VMs) or containers, emulate a software representation of a computing system within a physical computer. This allows for stricter isolation from other organisational IT systems and can be pre-configured for specific AI deployments.

Encrypt Sensitive Information [MITRE, 2024; Google, 2023]: To avoid damage from traditional security backdoors, developers need to encrypt sensitive data, such as ML models, to protect against unauthorised access by adversaries attempting to acquire sensitive data.

Verify ML Artefacts [MITRE, 2024] (Model Backdoors, Poisoning): Verify that the AI model, and other artefacts have not changed since its creation. This involves using techniques, such as provenance and

cryptographic checksum, which enables developers to validate if any modifications or tampering has occurred by an attacker which may lead to unexpected behaviour.

Restrict Library Loading [OWASP, 2024; MITRE, 2024]: Prevent abuse of software library loading mechanisms within the Operating System, and software to load unstructured code, by configuring appropriate library loading mechanisms and investigating potentially vulnerable software.

3.1.4 Operation and Maintenance

This section includes recommendations to alleviate the effectiveness of cyber security risks to AI in the operation (i.e. a live AI model running within a computer, capable of communication and provisioning service) and maintenance of an AI system.

Passive ML Output Obfuscation [OWASP, 2024; Vassilev, 2023; MITRE, 2024; BSI1, 2023; ESLA, 2023; Microsoft, 2022] (Extraction, Evasion, Inversion): Decreases the information outputted from a model (confidence values, input size, token limits) and reduces the ability for an adversary to extract information and bespoke attack optimisation. Such recommendations include confidence rounding [Shokri, 2017] and Gradient Masking [Vassilev, 2023].

ML Model Query Restrictions [OWASP, 2024; MITRE, 2024; BSI1, 2023; Microsoft, 2022] (Extraction, Evasion, Inversion): Limit the total number of queries a user can perform. Whilst typically leveraged by service providers during massive system load from user demand, this approach also limits the success of cyber-attack types against AI systems (extraction, inversion, evasion) that require a large number of queries to the AI model. Determining the suitable amount of query restrictions that has been identified is an important consideration, given that model architectures require different numbers of queries to perform extraction attacks [Hackett, 2022].

Adversarial Input Detection [OWASP, 2024; MITRE, 2024; NCSC, 2023; AI Verify, 2024; Google, 2023; BSI, 2022; Microsoft, 2022; Amazon, 2023; ICO, 2020; Leslie, 2019; Dong, 2021] (Evasion, Extraction): Detect and block adversarial inputs or atypical queries deviating from known user behaviour, exhibiting behaviour patterns observed in previous attacks, or originating from potentially malicious IPs. The literature also recommends incorporating adversarial detection algorithms into an AI system, whereby query traffic is first monitored prior to being sent to the model for inference [Juuti, 2019].

Model Obfuscation [OWASP, 2024; BSI, 2022] (Extraction, Physical Domain Attacks): For extraction attacks targeting lower parts of the system stack (e.g. computer kernel operations, memory cache), it has been demonstrated through experimentation that model obfuscation techniques [Zhou, 2023; Trawicki, 2023] can be particularly effective. This is because such obfuscation results in altering model properties (e.g. tensor program types, GPU kernel execution timings) to deviate from typical operation expected by an extraction attack. Such obfuscation can also be attained via deploying models within a trusted execution environment such as Oblivious Ram (ORAM) and Operation Masking [Hu, 2020].

Content Safety [Microsoft, 2023; Nvidia, 2023; BSI2, 2023] (Evasion, LLM Jailbreaks, Prompt Injection): Sanitises input and output to and from an LLM, enabling content moderation protecting users from malicious injection and unexpected LLM behaviour. Content safety systems detect harmful user-generated and AI-generated content in applications and services for text and image, allowing developers to detect and moderate harmful content, and set thresholds for severity levels.

LLM Jailbreak Risk Detection [BSI2, 2023] (LLM Jailbreaks): Detecting jailbreaks is beneficial for maintaining the security and integrity of the LLM, as it helps prevent unexpected behaviour and compromising confidential data. Techniques for such detection include classifying prompt semantic meaning.

Copyright detection [Microsoft, 2024] (Inversion): Ensure the AI model is not generating content that could be considered copyrighted. Attackers can reverse copyrighted data from an AI model potentially causing legal issues. Identify potential copyrighted material (lyrics, articles, web content, etc.), which the AI model may have been exposed to during training enables preventing leakage of such data.

3.2 General Recommendations

3.2.1 Company Practices, Polices, Governance & Security Hygiene

Legal and Regulatory Requirements [NIST, 2022; ASD, 2023; ICO, 2020]: AI legislation and the regulatory environment is rapidly evolving and changing. It is important to understand how the legal and regulatory requirements involving AI are managed and documented. Moreover, the literature recommends to apply existing cyber security practises to AI deployments (e.g. data protection, user access, compliance reporting).

Stakeholder Engagement [Vassilev, 2023; World, 2024]: Organisational policies and practices are in place to collect, consider, prioritise, and integrate external stakeholder feedback regarding the potential individual and societal impacts related to AI risks. Mechanisms are established to enable AI actors to regularly incorporate adjudicated stakeholder feedback into system design and implementation.

Create an Organisational AI Program / Sec Dev Program [OWASP, 2024; Deloitte, 2023; Microsoft, 2022; ICO, 2020]: Take responsibility for AI as an organisation. Recommendations include keeping an inventory of AI initiatives (e.g. to avoid the creation of shadow IT/models), as well as cyber risk management. This activity enables organisations to perform AI model and data risk governance.

Controls to Limit Unwanted Model Behaviour [OWASP, 2024; Google, 2023]: Conduct oversight of model behaviour through the use of human or business logic. This encompasses minimising privileges; and avoiding connecting a model to critical services or data. Moreover, AI transparency is recommended via informing departments and teams when an AI model is involved within a project or service.

Create and document AI project requirements [OWASP, 2024; World, 2024; ENISA, 2023]: Ensure that the creation, development, configuration of AI is produced, including its functional specification and any modifications made.

Identifying, Understanding and Defining Possible AI Threats [OWASP, 2024; MITRE, 2024; World, 2024; NCSC, 2023; Google, 2023; ESLA, 2023; Cisco, 2022; Deloitte, 2023; Microsoft, 2022; ASD, 2023]: As a part of the risk management process, apply a process to assess the threats to an AI system, understand how these threats impact and effect the system, users, and organisation. Recognise that the sensitivity and types of data used within the AI system may alter its intrinsic value to an attacker.

Limit Release of Public Information [OWASP, 2024; MITRE, 2024; Dong, 2021]: Limit the public release of technical information pertaining to the AI system stack used within an organisation's product or services. Technical knowledge of how AI is used can be leveraged by an attacker. Examples include Private Aggregation of Teacher Ensembles (PATE) [TRAN, 2021], masking, and encryption.

Consider Security Benefits and Trade-offs when Selecting your AI Model [NCSC, 2023, BSI1, 2023; ESLA, 2023; Cisco, 2022]: AI model selection involves balancing a range of requirements, including architecture, configuration, training data, etc. Such decisions are informed by the defined threat model. For example, the complexity of the model, the chosen architecture, and the number of parameters will affect how much training data is required, and how robust it is to changes in input data [Hackett, 2022].

Limit Model Artefact Release [MITRE, 2024]: Information regarding the model architecture, dataset, project details, algorithms, and model frameworks that are used in production can enable an attacker to optimise and fine-tune their attack against a specific AI model. Limiting public release of any technical information regarding your model limits the ability for an attacker to utilise such information.

Control Access to ML Models and Data [World, 2024; NCSC, 2023; ENISA, 2023; BSI1, 2023; Cisco, 2022; Microsoft, 2022; G7, 2023; HHS, 2021; OpenAI2, 2024] (Inversion, Poisoning): Different access controls can be applied based on the stage of model lifecycle. For design and development, it is recommended to use access controls to limit model access to critical training datasets, or datasets continuously generated from data outputted by externally facing production models. For models in production, it recommended that users are required to verify their identities prior to accessing a model, including API endpoint authentication.

Monitoring Use [OWASP, 2024; World, 2024; Reber, 2023; ENISA, 2023; BSI1, 2023; Cisco, 2022; Deloitte, 2023; G7, 2023; Amazon, 2023; ICO, 2020]: It has been recommended to monitor production model queries to ensure compliance with usage polices and model misuse prevention. Moreover, monitoring and logging model use (address, input, date, time, user) to incorporate into established organisational incident detection. Such monitoring identifies improper model function (continuous validation), suspicious patterns (abnormally high frequency, adversarial patterns), and suspicious inputs.

User and Staff Training [OWASP, 2024; Vassilev, 2023; MITRE, 2024; World, 2024; NCSC, 2023; Google, 2023]: Educate AI model developers and data scientists to secure coding practices and AI vulnerabilities. Additionally, raise awareness of threats and risks in AI systems.

Company Plan to Address AI Threats [OWASP, 2024; Google, 2023; Deloitte, 2023]: Review how current controls map to organisational AI use cases, and whether there exists a fit-for-purpose evaluation of these controls. This should be followed by the creation of a plan to address identified gaps and stakeholders should measure the effectiveness of these controls. The literature also recommends decision-making related to mapping, measuring, and managing AI risks throughout the lifecycle to ensure AI systems meet requirements for a subset of users.

Conduct Red Teaming and risk analysis [OWASP, 2024; World, 2024; Reber, 2023; ENISA, 2023; Google, 2023; G7, 2023; NCSC, 2023]: Red team exercises are a security testing method whereby a team of ethical hackers attempt to exploit security vulnerabilities. Performing red teaming early, especially during AI model development is crucial for preventing adverse outcomes and ensuring model safety. This recommendation is supported by government institutions which have requested that such testing is performed by independent third parties [White House, 2023]

Continuously Research the AI Threat Landscape [Google, 2023]: Stay on top of novel attacks. Track latest news posts, top research surveys etc. This can be achieved through various means including social media, as well as material provided by technology companies, CVEs, and AI security startups.

Ensure AI Data and Application Security Compliance [NIST, 2022; ENISA, 2023; ICO, 2020]: Ensure that AI data (training data, model inference, controls, etc.) complies with established organisational data security requirements and data governance processes.

Establish Strong Supply Chain Security [OWASP, 2024; NCSC, 2023; Microsoft, 2022; ASD, 2023] (Physical Domain Attacks): It is recommended to capture the entire supply chain in terms of security when using AI as a service or within an application. This encompasses whether to train a new AI model, use an existing model (with or without fine-tuning) or access a model via a third-party API. The literature also recommends conducting a due diligence of an external AI model provider to understand

their security posture as well as external libraries (for example, ensuring the library has controls that prevent the system loading untrusted models).

Document and review your data, models, code, and prompts [OWASP, 2024; World, 2024; NCSC, 2023; Cisco, 2022; Microsoft, 2022; ICO, 2020] (Backdoors): Document any information regarding the deployment of an AI model, data being used, the underlying architecture, and LLM prompts. Additionally, it is recommended to document the security-relevant information such as sources of training, etc.

3.3 Mapping Recommendations to Reported Security Vulnerabilities Against AI

We found that reported cyber attacks (see Section 5.2) proposed a variety of different technical and general recommendations, which can be categorised into the following:

Security Hygiene: Applying established cyber security best practises to AI system deployment. This includes appropriately managing user access and sharing [Luitjes, 2023] managing and verifying software dependencies [PyTorch, 2022], performing rigorous code review, minimising model access to specific systems within the wider organisation [Medium, 2022; Rehberger, 2023], and regularly rotating API keys [Hendrycks, 2021].

AI model: Model distillation [Hinton, 2014], model encryption, model obfuscation, LLM security testing [Liu, 2023], AI model carding, and relying on best practise from LLM operators [OpenAI1, 2024].

AI data: Watermarking, content restriction and categorisation used by models, additional filtering on prompt input-output channels [Greshake, 2023], detect feature over-proportionality, and detecting adversarial samples from artefacts.

It is worth noting that each of these recommendations have been provided as suggestions of possible solutions, as judged by those who originally reported the attack. Alternatively, authors have leveraged findings from academic research papers that propose recommendations based on postulation or laboratory experimentation [Fawzi, 2016; Moosavi-Dezfooli, 2017], which have been used to infer their effectiveness when applied to a security incident against AI in production. In contrast, the two cyber attacks scenarios [Antonov, 2021; PyTorch, 2022] which are stated to have examined the effectiveness of detection and remediation strategies within a production AI system consist of ensuring correct software dependencies and detecting adversarial samples from artefacts [Feinman, 2017].

4. Discussion of Recommendation Findings

This section outlines some themes and trends as well as knowledge gaps identified following a review of the literature. The section also outlines areas for potential future work.

Many recommendations for AI are based on established cyber security practise: It is the author's view that many recommendations provided – particularly for organisational practises described in Section 3.2.1. – are conventional cyber security practises currently performed by organisations. It is the author's opinion that a large body of current literature appears to view AI security as an independent activity siloed from conventional software. This approach is likely a means to introduce cyber security practitioners to the process and intricacies of AI, or alternatively educate data scientists developing AI to cyber security concepts. However, it is important to emphasise that AI in its current form is still ultimately software and data operating within computer hardware. Thus, there are various instances whereby existing cyber security principles to standardise, understand, detect, and remediate cyber security attacks are also applicable to AI models and systems.

Security vulnerabilities from conventional cyber security are applicable to AI: Many of the reported security vulnerabilities within AI leverage entirely new technical approaches and methods to successfully inflict cyber harm (see Section 5.3), which justifies the need to identify, create, and adopt new recommendations to address. It is the author's view that this is not an insurmountable challenge to achieve, given many of these vulnerabilities, attacks, and recommendations are identical or directly applicable to scenarios familiar to cyber security professionals.

Limited empirical studies on AI security vulnerabilities: From the study of cyber security vulnerabilities described in Sections 3.3 and 5, the report was able to identify 23 reported security incidents or proof of concepts against AI in production, and less than 30% of these can be categorised as actual security incidents. This small sample size results in limited evidence when empirically evaluating the effectiveness of technical recommendations in practise. The remaining technical recommendations have only been studied within academic settings in laboratory environments. Such academic works typically leverage sophisticated threat models to demonstrate the validity of cyber attacks as a proof of concept. While such works are highly useful when demonstrating the feasibility of cyber risks in AI (as well as debating opposing viewpoints [Kurakin, 2016; Zeng, 2019]), it limits their ability to evaluate their effectiveness against security incidents. This results in many of the technical recommendations provided within major AI security frameworks (OWASP, ATLAS, NIST, etc.) only being evidenced via inference within academic settings.

Many recommendations are derived from few unique data sources: We identified that a large body of recommendations have been proposed across various works, however such recommendations are derived from a narrow set of data sources. Specifically, we found considerable overlap across frameworks for technical and general recommendations (OWASP, MITRE, NCSC, NIST, Google), and several instances where each cite each other. The reason for this occurrence is possibly due to (i) organisations, researchers, and practitioners actively collaborating and sharing ideas; (ii) the objective of many of these frameworks is to collate good practises; (iii) there is limited, detailed information on technical recommendations successfully deployed for AI in production. As industry adoption and standards in AI security continue to mature, it is likely that these sources will diversify accordingly.

Exploiting publicly available knowledge: It is the author's view that while the use of open-source models made available on repositories, such as HuggingFace [HuggingFace1, 2024], have no doubt catalysed and pushed the AI innovation forward, this is equally true for nefarious actors. This is particularly problematic because we found that over 40% of published cyber attacks against AI systems

leveraged publicly available model information or data endpoints. This is an issue given evidence of transferability between model architectures. There has been discussion of this phenomena within academic literature [Hackett, 2022], however we were unable to identify this issue being described within current AI security frameworks [Silent, 2019; MITRE, 2023; OWASP 2024]. While this may be omitted because its occurrence is still not yet fully understood, it is intuitive to assume that the existence of such publicly available knowledge is also advantageous to attackers. We are by no means encouraging security by obfuscation as the solution to this challenge; however, this situation is a potential risk which organisations should be mindful of when designing and deploying AI.

Lack of Direct Recommendation Alignment to Attacks/Risks: The literature has provided a relatively comprehensive overview of different technical and general recommendations. However, there are still several instances of gaps in knowledge within the field. These gaps predominantly encompass an inability to directly align recommendations to a specific attack type. For example, we observed that extraction attacks (i.e. model theft) and inversion in many instances across frameworks recommendations [OWASP, 2024; Vassilev, 2023; MITRE, 2024; BSI 2022] referred to the use of general controls for mitigation. While the remit of this study is to report on recommendations stated within the literature, in some instances it was unclear how these may be effective, given omission of direct empirical evidence or technical applicability. For example, [OWASP, 2024; MITRE, 2024] states mitigating extraction attack should leverage general controls, which includes at least 16 governance and data limitation recommendations. Technical approaches such as tokenization or encryption would not be effective in this context of an AI model requiring to output meaningful information to an end-user. Moreover, many recommendations described what should be done to address cyber attacks against AI, however there was less guidance on how this could precisely be achieved. While frameworks provided detail on potential technical recommendation (although their effectiveness is typically inferred), there is limited discussion on the type of tools and products available to support their deployment. While a relatively nascent area, we envision an increased uptake of AI security products that leverage existing – as well as provide for new forms of – technical recommendations that interface with established general recommendations.

5. Reported Security Vulnerabilities within AI

To provide greater insight into the effectiveness of recommendations in actual use (see Section 4.3), we have also presented a brief analysis of publicly reported security vulnerabilities against AI, and descriptions of their respective recommendations proposed.

5.1 Data Collection

For identifying known security vulnerabilities within AI, we have collated reported security incidents or proof of concept attacks. Such information was found via an online search of GitHub pages, AI risk frameworks, news articles, technical blogs, and academic papers identified within our search for data sources described in Section 2.2. The relevancy of security vulnerabilities was determined by the author based on their expertise, as well as a criterion based on: the type of cyber security vulnerability and cyber attack, its occurrence, description, reported (or perceived) cyber harm, and linked recommendations for mitigating said vulnerability. The security vulnerabilities included in this report were based on whether they provided (i) sufficient empirical findings, demonstrating cyber risk feasibility and damage to AI in production; and (ii) responsible disclosure and recommendation steps for mitigation and remediation of the security vulnerability.

There are several cases whereby a single AI cyber attack and recommendation was reported across multiple instances and are thus counted as a single attack. Moreover there are several instances where reported security vulnerabilities meet the criteria above, but do not discuss or evidence the recommendation for detection and remediation. Such cases have also been included within this category, although these recommendations are based on the respective knowledge and expertise of the author who published the vulnerability.

5.2 Reported AI Security Vulnerabilities

The literature review identified 23 reported security vulnerabilities within AI. We discovered that all major AI attack types have been successfully leveraged within production security incidents, and that all scenarios (with one exception) used some form of adversarial ML to achieve their goals. Since 2022, there has been a growing trend in reported cyber attacks using LLM prompt injection, which is likely because of the rapid formation and uptake of LLMs services, (such as OpenAI), garnering attention from cyber security researchers and practitioners.

We found that information on recommendations for security vulnerabilities of AI in production was limited or incomplete (see Appendix 1 for a complete breakdown). Specifically, we found that across reported cyber attacks and recommendations: 7% had examined different detection or remediation techniques to evaluate their effectiveness, 26% were derived or inferred from similar models found within academic literature, 27% were provided without supporting evidence on their effectiveness, and 40% were omitted entirely. For reports on cyber attacks that omitted recommendations, we have assumed that this does not imply that no action has been taken to mitigate, rather it has in high likelihood not been reported externally from the organisation.

Reported security vulnerabilities resulted in differing levels of cyber harm, encompassing data exfiltration (Personal Identifiable Information (PII) [Rehberger, 2023; Greshake, 2023], training data [Medium, 2022]), service disruption (reduced service performance [NIST, 2022; MITRE, 2024], eroded model integrity [VirusTotal, 2020]), and reputational harm [Huynh, 2023]. For cases categorised as Proof of Concept and Red Teaming, discovering AI security vulnerabilities were predominantly conducted within a single scenario instance (i.e. a single model architecture) to demonstrate potential cyber harm. Examples of such damage include data leakage of chatbot conversation history due to LLM

prompt injection, access to user-collaborated training data, and model evasion of malware detection systems [Antonov, 2021]. Importantly, we discovered two security incidents within organisations which were the victim of a sophisticated and sustained cyber attack that exploited security vulnerabilities within AI:

Tax fraud: [Olson, 2021]: Attackers were able to impersonate individuals registered within the local government tax system in China. These attackers registered accounts using HD face photos acquired from an online black market, and used virtual camera app to generate video to evade ML-based facial recognition service used for user verification. Attackers were able to fraudulently acquire $77 million by user privilege access by creating fake shell companies that sent invoices to victims recognised as clients via the tax system.

Unemployment claim fraud [USAO, 2023]: The attacker filed 180 false unemployment claims to state of California, bypassing ID.me automated identity system (uses ML vision to extract content and verify ID documents), dozens of fraudulent claims approved. The attacker collected real identities and obtained fake driver licenses using the stolen personal details and photos. Individual filed fraudulent unemployment claims with the California Employment Development Department (EDD) under the ID.me verified identities. Due to flaws in ID.me's identity verification process at the time, the forged licenses were accepted. At least $3.4 million withdrawn in false unemployment benefits.

It is worth noting that the security vulnerabilities exploited within AI described above were used to conduct further financial harm within the organisation via fraudulent activities, as opposed to targeting the model itself (e.g. stealing training data, IP theft).

There exist several instances whereby multiple security vulnerabilities against AI models were exploited consecutively [Wallace, 2020; Schwartz, 2019; Li, 2021] to advance an attacker’s objectives. This was predominantly the case for exploiting model supply chain risks to launch more focused attacks requiring further model or system knowledge, as well as creating AI model copies (or proxy models) from collected data to bypass AI-driven systems via model evasion cyber attacks.

6. Conclusions

In this report, we have presented a comprehensive review of the current recommendations available to address cyber security risks in AI. We were able to identify 45 unique recommendations – spanning technical, organisational, and governance – which have been proposed by an actively growing community within industry, academia, government, and practitioners.

The formation of cross-sector initiatives to share such recommendations within the past few years is particularly encouraging. However, this report has discovered that there remain several issues pertaining to a limited number of recommendations derived from empirical findings, knowledge gaps in security vulnerabilities against AI in production, as well as outstanding questions in how recommendations evaluated in laboratory environments effectively transfer into practise.

The topic of AI security is by no means solved, and is an active area of research that continues to change with the evolution and proliferation AI advancements. As stated in recent studies from NIST [OWASP, 2024]:

"*Currently, there is no approach in the field of machine learning that can protect against all the various adversarial attacks.*"

This report has reviewed literature published prior to 12 January 2024. Thus, given the rapidly growing activity within this field, this study should be considered as a snapshot of the current landscape of recommendations to address cyber security risks to AI.

As mentioned within this report in Section 1.1, there exist thousands of academic publications on the topic of adversarial ML, whose recommendations may directly or indirectly aid in reducing AI cyber risk. Moreover, we believe that we have captured the key recommendations provided from major AI-driven technology within our data collection method and criteria. Qualifying which works to give particular focus on (and to include) within this review is limited by the expertise and judgement of the author. Hence, it is possible some recommendations may have been unintentionally omitted. This study should not be considered an exhaustive list of all possible recommendations, and instead an overview of the AI security landscape, and emerging trends within the area.

Moreover, there is a possibility that the number of security vulnerabilities and recommendations for AI is likely higher than reported. This is due to a multitude of factors, including organisational transparency in reporting attacks, possibly due to the nature of their work, and 'unknown unknowns' (it is not possible to report attacks if one is not equipped to understand, detect or mitigate).

7. Definitions

To further assist the reader, we have provided a succinct summary of terminology and definitions used within this report.

7.1 Artificial intelligence (AI)

AI model: A computer program trained on a set of data to recognise patterns and perform specific tasks [Dong, 2021]. Through extensive training, an AI model is capable of learning patterns within the provided data using different learning methods (supervised, semi-supervised, unsupervised). These learnt patterns are leveraged to make predictions or decisions to achieve a desired outcome. The data used to train AI (known as training data) can range from text, images, audio, sensor readings to financial transactions. AI can be leveraged to perform a variety of tasks ranging from data classification [Krishnaiah, 2014], object recognition [Jiang, 2022], text summarisation [Torfi, 2021], translation [Nam, 2024], as well as a multitude of other applications. AI models are used throughout many industries, including but not limited to health care for medical diagnosis [Jiang, 2017], financial for stock prediction [Hu, 2021], fraud detection [Bao, 2022], astronomy [Sen, 2022]. etc.

- **Machine Learning (ML)**: AI models that learn from data without being explicitly programmed. Machine learning algorithms can be used to identify patterns in data, make predictions, and make decisions.
- **Deep Learning**: A type of ML model that leverages artificial neural networks to learn from data. Deep learning models excel at recognising patterns in complex data, such as images, text, and speech.
- **Natural Language Processing (NLP)**: AI models that deal with the interaction between computers and human (natural) languages. NLP algorithms can be used to understand the meaning of text, translate languages, and generate text.

AI, ML, and Deep Learning are terms often used interchangeably within the public and industry, although they are reasonably distinct. Within this report we primarily focus on AI models which leverage Deep Learning, which we simply refer to as an *AI model*. This definition was selected due to (i) sufficient scoping around technology/application domain, and more importantly (ii) since 2012 Deep Learning has been demonstrated to outperform a large set of other AI/ML techniques, and as a result has been the primary focus of recent attention to AI. There are many different sub-domains of AI models chosen due to the type of training data, complexity, and desired task [Dong, 2021]. These AI models are built upon many various different AI architectures and algorithms:

Neural Networks (NN): Representing the foundation of AI models, neural networks are layered structures where information flows from one to the next based on weights, bias of the internal network [Dhruv, 2020]. Specialised networks were created upon NNs such as Convolution Neural Networks (CNNs) for images, and Recurrent Neural Networks (RNNs) for sequential data [Dhruv, 2020].

Transformer architectures: Rely on 'encoder-decoder' operators. Parallel processing allows for efficient handling of long sequences and capturing long-range dependencies between elements [Vaswani, 2023]. Such an approach has been demonstrated to be highly effective within Natural Language Processing (NLP) tasks such as machine translation, text summarisation, and question answering. Their success has led to their usage in computer vision tasks such as image classification and object recognition [Dosovitskiy, 2021].

Large Language Models (LLMs): Traditionally built-upon the success of the transformer architecture, these models are trained on massive amounts of text and code, containing billions of parameters, excelling at understanding and generating human language [Chang, 2024].

Generative AI: Also known as GenAI, refers to a type of Artificial Intelligence that can create new and original content, such as text, images, music, audio, and synthetic data [OpenAI2, 2024; OpenAI1, 2023; ElevenLabs, 2024]. Due to the complexity of the task, GenAI are typically built upon models with billions of parameters trained on billions of data. For example, Generative text models currently rely on LLMs due to their sophisticated understanding of human language and therefore its ability to create text with high degrees of creativity, relevancy, and accuracy to the input.

AI system: An AI system is a computer system built specifically for handling AI across its entire lifecycle (design, training, deployment, etc.). The hardware of these systems contain specialised components such as Graphics Processing Units (GPUs), Tensor Processing Units (TPUs) which enable performant training and inference of AIs compared to traditional hardware. Additionally, AI systems include software that help build, and run models, such as 1) ML frameworks; PyTorch, TensorFlow, ONNX, 2) compilers; TensorRT, TVM, [Nvidia1, 2024] and 3) System Drivers; Nvidia CUDA, and AMD ROCm [AMD, 2024]. AI systems are a necessity for AI models to be successfully deployed and execute.

7.2 AI Security (Cyber Security for AI)

AI security is defined as the protection of AI throughout the entire ML lifecycle, from development, deployment, and use of AI systems from various cyber threats and vulnerabilities [Lin, 2021]. Securing AI is essential to prevent confidential, organisational, representational, and privacy damage. This includes assessing the potential threats against your AI system, detecting adversarial attacks, and remediating against attempts from attackers.

Detection: Adversarial detection is a technique used within AI security to identify and mitigate malicious attacks on ML models [Juuti, 2019]. Such detection can help prevent unauthorised access, data tampering, and other types of malicious activity that can compromise the integrity and accuracy of AI systems.

Remediation: Remediation is the process of identifying and addressing security vulnerabilities in AI systems [Hosseini, 2017; Piet, 2023]. This involves identifying potential threats, assessing the impact of those threats, and implementing measures to mitigate or eliminate the risks.

There exist a multitude of cyber attacks specifically targeting AI models and systems:

Adversarial Perturbations: Adversarial perturbations refer to deliberately crafted, modifications made to the input data of an AI model. These perturbations are designed to be imperceptible to humans but can significantly affect the output of the model and are typically used within a wide range of attack domains: Extraction, Evasion, Poisoning, etc.

Model Poisoning: A cyber attack whereby an attacker intentionally introduces misleading or malicious data into an AI model to manipulate or mislead its predictions or decisions [Tian, 2022]. This attack can be performed via injecting false or biased data into the training set, or by introducing noise or errors into the model's inputs or outputs. The goal of model poisoning is to compromise the AI model integrity and accuracy, create misinformation, or cause harm to individuals or organisations [Lin, 2021].

Model Backdoor: Adds a hidden bias to an AI model, which can cause the model to make predictions that are not based on the actual input data [Hosseini, 2017]. Creating a model backdoor is typically performed during the AI model training phase, where an attack may utilise model poisoning to inject

maliciously created data samples, such as precisely perturbated data points that causes a model to learn a specific pattern. This pattern can then be exploited by a subsequent attack when the model is released. An attacker can leverage an installed backdoor via supplying data with backdoor perturbations to evade classification or achieve their crafted outcome.

Model Evasion: An attack whereby an attacker aims to evade correct classification by a target model by maliciously crafting adversarial perturbations on inputs to the model causing misclassification [Lin, 2021]. A carefully perturbed input can look indistinguishable to a human eye but completely different to a model. Successful evasion attacks take advantage of the overreliance of learned features and data patterns.

Model Extraction: Also known as model stealing, model extraction is the act of creating a functionality equivalent copy of a target AI model. An attacker can take advantage of the outputs received from a model, such as labels, and confidence values in order to create their own version, extracting the fundamental characteristics of the target. A common method of recreation is model training whereby an attacker will use the received data from the target to train a shadow model with high fidelity to the target. Such attacks result in information leakage, digital IP theft, and enable further adversarial attacks to be staged [Hackett, 2022].

Inversion: Model inversion is an adversarial attack whereby an attacker aims to reverse engineer the model to extract confidential training data by exploiting the outputs of a target model. Successful inversion enables an attacker to recreate data originally used within the training dataset, this is problematic for models which were trained upon data concerning user privacy, such as facial recognition models, or confidential company data, which when inverted allow an attacker to acquire the original data [Zhang, 2020].

ML Supply Chain Compromise: Attackers may utilise other security vulnerabilities and exploits to gain initial access the underlying system and then compromising critical components within the AI system supply chain. This could include AI accelerator hardware such as a GPU or TPU, CPU, system memory and cache, AI software stack, or the model itself [Hu, 2020].

LLM Prompt Injection: A security vulnerability that exploits how LLMs process prompts – the instructions used to guide their output [Greshake, 2023]. An attacker maliciously crafts a prompt that can manipulate LLM behaviour and output, leading to various cyber harms ranging from data leakage, misinformation, and enable further attacks to be staged [Rehberger, 2023]

LLM Prompt Jailbreak: Refers to a more sophisticated form of prompt injection aiming to circumvent safety and moderation features implemented within an LLM [Greshake, 2023; Chao, 2023]. Traditionally, LLMs can be safeguarded via a well-crafted system prompt; describing what rules the LLM must follow including content it should respond to or block, or trained upon data that makes the LLM inherently more robust to harmful content. Prompt jailbreak, however, uses well-crafted prompts to attempt to "break out" of the intended behaviour limitations of the LLM in order to gain access to typically restricted functionality. Utilising a successful jailbreak enables and attacker to extract underlying information from the LLM, such as fine-tuned company data used to instruct the LLM or enable an attack to launch further attacks [Chao, 2023].

References

[AI Verify, 2024] AI Verify. Accessed: January 7th 2024. Online: https://aiverifyfoundation.sg/what-is-ai-verify/

[Amazon, 2023] Amazon, AWS Cloud Adoption Framework for Artificial Intelligence, Machine Learning, and Generative AI, Amazon White Paper, 2023.

[AMD, 2024] AMD ROCm, Accessed: January 9th 2024. Online: https://www.amd.com/en/products/software/rocm.html

[Andriushchenko, 2020] M. Andriushchenko, F. Croce, Nicolas Flammarion, Matthias Hein, Square Attack: A Query-efficient Black-box Adversarial Attack via Random Search, ECCV, 2020.

[Antonov, 2021] A. Antonov, A. Kogtenkov, How to Confuse Antimalware Neural Networks. Adversarial Attacks and Protection, SecureList, 2021, Online: https://securelist.com/how-to-confuse-antimalware-neural-networks-adversarial-attacks-and-protection/102949/

[ASD, 2023] Australian Signals Directorate, An introduction to Artificial Intelligence, 2023.

[Bai, 2021] T. Bai, j. Luo, J. Zhao, B. Wen, Q. Wang, Recent Advances in Adversarial Training for Adversarial Robustness, IJCAI, 2021. Online: https://arxiv.org/pdf/2102.01356.pdf

[Bao, 2022] Y. Bao, G. Hilary, B. Ke. Artificial Intelligence and Fraud Detection, Innovative Technology at the Interface of Finance and Operations, 2022.

[Birch, 2023] L. Birch W. Hackett, S. Trawicki, N. Suri, P. Garraghan, Model Leeching: An Extraction Attack Targeting LLMs, CAMLIS, 2023.

[BSI, 2022] Federal Office for Information Security, Security of AI-Systems: Fundamentals, 2022.

[BSI1, 2023] Federal Office in Information Security, AI Security Concerns in a Nutshell, 2023.

[BSI2, 2023] Federal Office in Information Security, Large Language Models: Opportunities and Risk for Industry and Authorities, 2023.

[Bunzel, 2023] N. Bunzel, Multi-class Detection for Off the Shelf Transfer-based Black Box Attacks, SecTL, 2023.

[Carilini, 2016] N. Carlini, D. Wagner, Defensive Distillation is Not Robust to Adversarial Examples, arXiv, 2016.

[Chang, 2024] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, W. Ye, Y. Zhang, Y. Chang, P. S. Yu, Q. Yang, X. Xie, A Survey on Evaluation of Large Language Models, Association for Computing Machinery, 2024.

[Chao, 2023] P. Chao, A. Robey, E. Dobriban, H. Hassani, G. J. Pappas, E. Wong, Jailbreakng Black-Box Large Language Models in Twenty Queries, arXiv, 2023.

[Cisco, 2022]Cisco, The Cisco Responsible AI Framework, 2022. Online: https://www.cisco.com/c/dam/en_us/about/doing_business/trust-center/docs/cisco-responsible-artificial-intelligence-framework.pdf

[Deloitte, 2023] Deloitte, Safeguarding Generative Artificial Intelligence with Cybersecurity Measures, 2023.

[Derico, 2023] D. Derico, ChatGPT Bug Leaked Users' Conversation Histories, BBC, 2023.

[Dhruv, 2020] P. Dhruv, S. Naskar, Image Classification Using Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN): A Machine Learning and Information Processing. Advances in Intelligent Systems and Computing, 2020.

[Dong, 2020] Y. Dong, Z. Deng, T. Pang, J. Zhu, H. Su, Adversarial Distributional Training for Robust Deep Learning, International Conference on Neural Information Processing Systems, 2020.

[Dong, 2021] S. Dong, P. Wang, K. Abbas, A Survey on Deep Learning and its Applications, Computer Science Review, 2021.

[Dong, 2021] S. Dong, P. Wang, K. Abbas, A Survey on Deep Learning and its Applications, Computer Science Review, 2021.

[Dosovitskiy, 2021] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, arXiv, 2021.

[DSIT, 2024]

[ElevenLabs, 2024] ElevenLabs, ElevenLabs: Text to Speech, Online: https://elevenlabs.io/ Accessed: January 10th 2024.

[ENISA, 2021] ENISA, Securing Machine Learning Algorithms, European Union Agency for Cyber Security, 2021

[ENISA, 2023] ENISA, Multilayer Framework for Good Cybersecurity Practices for AI, ENISA, 2023.

[ESLA, 2023] ESLA, European Lighthouse on Secure and Safe AI, 2023.

[Fawzi, 2016] A. Fawzi, S. Moosavi-Dezfooli, P. Frossard, Robustness of classifiers: From Adversarial to Random Noise, NIPS, 2016.

[Feinman, 2017] R. Feinman, R. R. Curtin, S. Shintre, A. B. Gardner, Detecting Adversarial Samples for Artifacts, ICML, 2017.

[G7, 2023] G7 Hiroshima Summit, Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems, 2023.

[Gao, 2020] Y. Gao, C. Xu, D. Wang, S. Chen, D. C. Ranasinghe, S. Nepal, STRIP: A Defence Against Trojan Attacks on Deep Neural Networks, arXiv, 2020.

[Ge, 2021] Y. Ge, Q. Wang, B. Zheng, X. Zhuang, Q. Li, C. Shen, C. Wang, Anti-Distillation Backdoor Attacks: Backdoors Can Really Survive in Knowledge Distillation, ACM International Conference on Multimedia, 2021.

[Google, 2023] Google, Google Secure AI Approach Framework (SAIF), 2023. Online: https://services.google.com/fh/files/blogs/google_secure_ai_framework_approach.pdf

[Greshake, 2023] K. Greshake, S. Abdelnabi, S. Mishra, C. Endres, T. Holz, M. Fitz, Not What You've Signed Up For: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection, Association for Computing Machinery, 2023.

[Hackett, 2022] W. Hackett, S. Trawicki, Z. Yu, N. Suri, P. Garraghan, PINCH: An Adversarial Extraction Attack Framework for Deep Learning Models, arXiv, 2022. Online: https://arxiv.org/abs/2209.06300

[Hendrycks, 2021] D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, J. Steinhardt, Measuring Mathematical Problem Solving with the MATH Dataset, NEURIPS, 2021.

[HHS, 2021] United States Department of Health and Human Services, Trustworthy AI (TAS) Playbook, 2021.

[Hinton, 2014] G. Hinton, O. Vinyals, J. Dean, Distilling the Knowledge in a Neural Network, 2014. Online: https://arxiv.org/abs/1503.02531

[Hosseini, 2017] H. Hosseini, Y. Chen, S. Kannan, B. Zhang, R. Poovendran. Blocking Transferability of Adversarial Examples in Black-box Learning Systems, arXiv, 2017.

[Hu, 2020] X. Hu, L. Liang, S. Li, L. Deng, P. Zuo, Y. Ji, X. Xinfeng, Y. Ding, C. Liu, T. Sherwood, Y. Xie, DeepSniffer: A DNN Model Extraction Framework Based on Learning Architectural Hints, Association for Computing Machinery, 2020.

[Hu, 2021] Z. Hu, Y. Zhao, M. Khushi. A Survey of Forex and Stock Price Prediction Using Deep Learning, Applied System Innovation, 2021.

[HuggingFace1, 2024] HuggingFace, Online: https://huggingface.co/, Accessed: January 14th 2024.

[HuggingFace2, 2024]HuggingFace, SafeTensors: A Simple, Safe Way to Store and Distribute Tensors, Accessed: February 1 2024, Online: https://github.com/huggingface/safetensors

[Huynh, 2023] D. Huynh, J. Hardouin, PoisonGPT: How we Hid a Lobotomized LLM on HuggingFace to Spread Fake News, Mithril Security, 2023.

[ICO, 2020] Information Commissioner's Office, Guidance on AI Auditing Frameworks, ICO, 2020.

[Jiang, 2017] F. Jiang, Y. Jiang, H. Zhi, Y. Dong, H. Li, S. Ma, Y. Wang, Q. Dong, H. Shen, Y. Wang, Artificial Intelligence in Healthcare: Past, Present and Future, Stroke and Vascular Neurology, 2017.

[Jiang, 2022] P. Jiang, D. Ergu, F. Liu, Y. Cai, B. Ma, A Review of Yolo Algorithm Developments, Procedia Computer Science, 2022.

[Juuti, 2019] M. Juuti, S. Syzller, S. Marchal, N. Asokan, PRADA: Protecting Against DNN Model Stealing Attacks, IEEE EUROS&P, 2019.

[Kakizaki, 2019] K. Kakizaki, K. Yoshida, Adversarial Image Translation: Unrestricted Adversarial Examples in Face Recognition Systems, arXiv, 2019.

[Krishnaiah, 2014] V. Krishnaiah, G. Narsimha, N. Subhash Chandra, Survey of classification techniques in data mining. International Journal of Computer Sciences and Engineering 2.9, 2014.

[Kurakin, 2016] A. Kurakin, Adversarial Machine Learning at Scale, arXiv, 2016. Online: https://arxiv.org/abs/1611.01236

[Leslie, 2019] D. Leslie, Understanding Artificial Intelligence Ethics and Safety: A Guide for the Responsible Design and Implementation of AI Systems in the Public Sector, The Alan Turing Institute, 2019.

[Li, 2021] Y.Li, J. Hua, H. Wang, C. Chen, Y. Liu, DeepPayload: Black-box Backdoor Attack on Deep Learning Models through Neural Payload Injection, ACM ICSE, 2021.

[Li, 2024] H. Li, P. P. K. Chan, An Improved Reject on Negative Impact Defense, ICMLC, 2024.

[Lin, 2021] J. Lin, L. Dang, M. Rahouti, K. Xiong, ML Attack Models: Adversarial Attacks and Data Poisoning Attacks. ArXiv, 2021.

[Liu, 2023] Y. Liu, G. Deng, Z. Xu, Y. Li, Y. Zheng, Y. Zhang, L. Zhao, T. Zhang, Y. Liu, Jailbreaking ChatGPT via Prompt Engineering: An Empirical Study, 2023.

[Luitjes, 2023] L. Luitjes, InjectGPT: The Most Polite Exploit Ever, 2023. Online: https://blog.luitjes.it/posts/injectgpt-most-polite-exploit-ever/

[Martin, 2023] M. R. Thinking About the Security of AI Systems, NCSC, 2023.

[Medium, 2022] Medium, Careful Who You Colab With: Abusing Google Colaboratory, 2022. Online: https://medium.com/mlearning-ai/careful-who-you-colab-with-fa8001f933e7

[Microsoft, 2022] Microsoft, Artificial Intelligence and Machine Learning Security, Microsoft Learn, 2022.

[Microsoft, 2023] Microsoft, QuickStart: Azure AI Content Safety Studio, 2023. Online: https://learn.microsoft.com/en-gb/azure/ai-services/content-safety/studio-quickstart

[Microsoft, 2024] Microsoft, Assessing Toolkit Counterfit, Accessed January 3rd 2024. Online: https://github.com/Azure/counterfit

[MITRE, 2020] MITRE ATLAS, Microsoft Edge AI Evasion, MITRE ATLAS Case Studies, 2020.

[MITRE, 2023] MITRE, Adversarial Threat Landscape for Artificial-Intelligence Systems (ATLAS), 2023. Online: https://atlas.mitre.org/

[MITRE, 2024] MITRE, MITRE ATLAS: Mitigations. Accessed: January 10th 2024. Online: https://atlas.mitre.org/mitigations/

[Moosavi-Dezfooli, 2017] S. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, P. Frossard, and S. Soatto, Analysis of Universal Adversarial Perturbations, 2017. Online: https://arxiv.org/abs/1705.09554

[Nam, 2024] W. Nam, B. Jang, A survey on Multimodal Bidirectional Machine Learning Translation of Image and Natural Language Processing, Expert Systems with Applications, 2024.

[NCSC, 2023] NCSC, Guidelines for Secure AI Systems Development, 2023. Online: https://www.ncsc.gov.uk/files/Guidelines-for-secure-AI-system-development.pdf

[NIST, 2022] NIST, AI Risk Management Framework: Second Draft, 2022. Online: https://www.nist.gov/system/files/documents/2022/08/18/AI_RMF_2nd_draft.pdf

[Nvidia, 2023] Nvidia, Nvidia Guardrails, 2023. Online: https://github.com/NVIDIA/NeMo-Guardrails

[Nvidia1, 2024] Nvidia, Nvidia TensorRT, Accessed: January 2nd 2024, Online: https://developer.nvidia.com/tensorrt

[Nvidia2, 2024] Nvidia, Nvidia CUDA, Accessed: January 9th 2024, Online: https://developer.nvidia.com/cuda-toolkit

[Olson, 2021] P. Olson, Faces Are the Next Target for Fraudsters, The Wall Street Journal, 2021.

[OpenAI1, 2023] OpenAI, Improving Image Generation with Better Captions, 2023 Online: https://cdn.openai.com/papers/dall-e-3.pdf

[OpenAI1, 2024] OpenAI, Safety Best Practises, 2024. Accessed: January 5th 2024. Online: https://platform.openai.com/docs/guides/safety-best-practices

[OpenAI2, 2023] OpenAI, Preparedness Framework (Beta), OpenAI, 2023.

[OpenAI2, 2024] OpenAI, Introducing ChatGPT, Online: https://openai.com/blog/chatgpt, Accessed: January 9th 2024.

[OWASP, 2024] OWASP, OWASP AI Exchange. Accessed January 12th 2024. Online: https://owaspai.org/

[Piet, 2023] J. Piet, M. Alrashed, C. Sitawarin, S. Chen, Z. Wei, E. Sun, B. Alomair, D. Wagner, Jatmo: Prompt Injection Defense by Task-Specific Finetuning, 2023, Online: https://arxiv.org.abs/2312.17673

[PyTorch, 2022] PyTorch, Compromised PyTorch-nightly dependency chain between December 25th and December 30th, 2022. Online: https://pytorch.org/blog/compromised-nightly-dependency/

[Reber, 2023] D. Reber Jr, Six Steps Toward AI Security, Nvidia, 2023. Accessed: January 5th 2024. Online: https://blogs.nvidia.com/blog/ai-security-steps/

[Rehberger, 2023] J. Rehberger, Hacking Google Bard – From Prompt Injection to Data Exfiltration, Embrace The Red. Accessed: January 4th 2024. Online: https://embracethered.com/blog/posts/2023/google-bard-data-exfiltration/

[Rehberger, 2023] J. Rehberger, Malicious ChatGPT Agents: How GPTs Can Quietly Grab Your Data, Embrace The Red. Accessed: January 4th 2024. Online: https://embracethered.com/blog/posts/2023/openai-custom-malware-gpt/

[Samangouei, 2018] P. Samangouei, M. Kabkab, R. Chellappa, Defense-GAN: Protecting Classifiers Against Adversarial Attacks Using Generative Models, CVPR, 2018.

[Schwartz, 2019] O. Schwartz, In 2016, Microsoft's Racist Chatbot Revealed the Dangers of Online Conversation, IEEE Spectrum, 2019.

[Sen, 2022] S. Sen, S. Agarwal, P. Chakraborty, K. P. Singh, Astronomical big data processing using machine learning: A comprehensive review, Experimental Astronomy, 2022.

[Shokri, 2017] R, Shokri, M. Stronati, V. Shmatikov, Membership Inference Attacks against Machine Learning Models, IEEE Symposium on Security and Privacy, 2017.

[Silent, 2019] Silent Break Security, Proofpoint Evasion, 2019.

[Skylight, 2019] Skylight Cyber, Cylance, I Kill You!, 2019, Online: https://skylightcyber.com/2019/07/18/cylance-i-kill-you/

[Tian, 2022] Z. Tian, L. Cui, J. Liang, Y. Jie, S. Yu, A Comprehensive Survey on Poisoning Attacks and Countermeasures in Machine Learning, Association for Computing Machinery, 2022.

[Torfi, 2021] A. Torfi, R. A. Shirvani, Y. Keneshloo, N. Tavaf, E. A. Fox, Natural Language Processing Advancements by Deep Learning: A Survey, 2021.

[Tramer, 2020] F. Tramer, N. Carlini, W. Brendel, A. Madry, On Adaptive Attacks to Adversarial Example Defenses, NEURIPS, 2020.

[Trawicki, 2023] S. Trawicki, W. Hackett, L. Birch, N. Suri, P. Garraghan, Compilation as a Defense: Enhancing DL Model Attack Robustness via Tensor Optimization, CAMLIS, 2023.

[USAO, 2023] United States Attorney's Office, New Jersey Man Sentenced to 6.75 Year in Prison for Schemes to Steal California Unemployment Insurance Benefits and Economic Injury Disaster Loans, 2023.

[Vacanti, 2020] G. Vacanti, A. V. Looveren, Adversarial Detection and Correction by Matching Prediction Distribution, arXiv, 2020.

[Vassilev, 2023] A. Vassilev, A. Oprea, A. Fordyce, H. Anderson, Adversarial Machine Learning – A Taxonomy and Terminology of Attacks and Mitigations, NIST, 2023

[Vaswani, 2023] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is All You Need, arXiv, 2023.

[VirusTotal, 2020] McAfee Advanced Threat, VirusTotal Poisoning, MITRE, 2020. Online: https://atlas.mitre.org/studies/AML.CS0002

[Wallace, 2020] E. Wallace, M. Stern, D. Song, Imitation Attacks and Defenses for Black-box Machine Translation Systems, EMNLP, 2020.

[Wenger, 2021] E. Wenger, J. Passananti, A. Bhagoji, Y. Yao, H. Zheng, B. Y. Zhao, Backdoor Attacks Against Deep Learning Systems in the Physical World, 2021.

[White House, 2023] The White House, Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, 2023. Online: https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

[Whittaker, 2020] Z. Whittaker, Security Lapse Exposed Clearview AI Source Code, TechCrunch, 2020.

[World, 2024] World Economic Forum, IBM, The Presidio AI Framework, 2024. Online: https://www3.weforum.org/docs/WEF_Presidio_AI%20Framework_2024.pdf

[Wu, 2017] X. Wu, U. Jang, J. Chen, L. Chen, S. Jha, Reinforcing Adversarial Robustness using Model Confidence Induced by Adversarial Training, arXiv, 2017.

[Xu, 2019] W. Xu, D. Evans, Y. Qi, Is Robust Machine Learning Possible?, EvadeML, 2019.

[Zahalka, 2023] J. Zahalka, Backdoor Attacks & Defense @ CVPR '23: How to Build and Burn Trojan Horses, 2023. Accessed January 10th 2024. Online: https://zahalka.net/ai_security_blog/2023/09/backdoor-attacks-defense-cvpr-23-how-to-build-and-burn-trojan-horses/

[Zeng, 2019] X. Zeng, C. Liu, Y. Wang, W. Qiu, Y. Tai, C. K. Tang, A. L. Yuile, Adversarial Attacks Beyond the Image Space, CVPR, 2019. Online: https://arxiv.org/abs/1711.07183

[Zhang, 2020] Y. Zhang, R. Jia, H. Pei, W. Wang, B. Li, D. Song, The Secret Revealer: Generative Model-Inversion Attacks Against Deep Neural Networks, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[Zhou, 2023] M. Zhou, X. Gao, J. Wu, J. Grundy, X. Chen, L. Li, ModelObfuscator: Obfuscating Model Information to Protect Deployed ML-based Systems, ISSTA, 2023.

Appendix 1. Overview of Reported Security Vulnerabilities within AI Deployed in Production

Reference	Attack Type	Occurrence	Security Vulnerability	Cyber Harm	Recommendation
[Medium, 2022]	Model extraction, data exfiltration	Proof of Concept	Allows Jupyter notebook to be shared containing malicious code. Victim mounts Google Drive onto compromised Colab notebook. Adversary accesses Jupyter Notebook to execute commands (extract checkpoints/training dataset).	Data exfiltration AI IP theft	*Review code *Scanning plugin for Colab, *Manage user access
[Hendrycks, 2021]	Direct LLM prompt injection	Proof of Concept	Prompt-overriding techniques to produce code leading to actors gaining access to application host system variables, LLM API key, and exhaust API query budget. Adversary can indirectly execute any arbitrary code on any Python interpreter interfaced with the LLM.	Financial harm, Denial of AI service	*Rotating API key Filter select prompts
[Luitjes, 2023]	Direct LLM prompt injection, remote code execution	Proof of Concept	Leverages LLM to generate Remote Code Execution to extract user email addresses and password hashes	Data exfiltration	*Limit admin access, *More remediation to be added in the future
[Rehberger, 2023]	Indirect LLM prompt injection	Proof of Concept	Indirect prompt injection vulnerability in ChatGPT, whereby an adversary can feed malicious websites to the LLM to influence an AI assistant chatbot to exfiltrate chat history.	PII leakage from chat session	*Minimize plugin access to conversation context, *OpenAI safety best practise *Nem Guardrails
[Greshake, 2023]	Indirect LLM prompt injection	Proof of Concept	Plant injection into website, enabling Bing Chat to exfiltrate personal information or commit fraud from victims via social engineering.	Exfiltrate PII for further identity-attacks (identity theft, fraud)	*Additional filtering on input-output channels. *LLM supervisor/moderator
[PyTorch, 2022]	ML supply chain, data exfiltration	Production security incident	Malicious binary/package dependency uploaded into Pippi code repository (PyTorch pre-release version) sharing an identical name. Package contained code that uploaded data from the machine (IP address, username, hostname, environment variables)	Remove dependency	Remove dependency Delete malicious version
[Antonov, 2021]	Model extraction, model evasion	Proof of Concept	Feature knowledge sufficient for AI model adversarial attack. Adversary attacked Kaspersky antimalware ML model, evaded detection when adversarial modifying malware files.	Data exfiltration, service disruption	*Distillation as a defence Detect adversarial samples from artifacts
[MITRE, 2020]	Model evasion	Red teaming	Performed red teaming on Microsoft product designed to run AI workloads at the edge, create an automated systems to manipulate images to create misclassifications.	Service disruption	N/A
[Silent, 2019]	Model evasion	Red teaming	Microsoft AI Red Team performed on Azure service to conduct service disruption. Exfiltrated data to then stage more sophisticated attacks	Data exfiltration, find user accounts	N/A
[NIST, 2022]	Model extraction, model evasion	Proof of Concept	ML researchers evaded ProofPoint's email protection system. Researchers conducted an extraction attack by gathering ProofPoint ML system outputs in terms of email scores, and then training a model on this dataset. This shadow model was used to create malicious emails that received preferable scores from ProofPoint's system in production.	Erode service performance	N/A
[Olson, 2021]	Model evasion	Production security incident	Adversary registered accounts using HD face photos from online black market, and used virtual camera app to generate video to evade ML-based facial recognition service for user verification to gain access to victims and verify identify within tax systems.	Attackers able to fraudulently collect $77 million by user privilege access to send invoices to "supposed" clients.	N/A
[Skylight, 2019]	Model evasion	Proof of Concept	Used publicly accessible information on Cylance's AI detector, researchers at Skylight created a universal bypass string that evades malware detection when appended to a malicious file.	User harm. Critical security system degradation.	*Anti-tampering controls to detect manipulation, *Detect feature over-proportionality
[Wallace, 2020]	Model extraction, model evasion	Proof of Concept	Collected end-point data from public facing machine translation services (Google, Bing) to create a shadow model with near production translation quality. Damaging in terms of IP theft, can be used to successful transfer adversarial examples into production systems (word flips, vulgar outputs, etc.).	Erode model integrity AI IP theft	*Repurpose prediction poisoning for machine translation, *Monitor user queries, *Watermarking

[USAO, 2023]	Model evasion	Production security incident	Individual filed 180 false unemployment claims, bypassing ID.me automated identity system (uses ML vision to extract content and verify ID documents), dozens of fraudulent claims approved. Adversary collected real identities and obtained fake driver licenses using the stolen personal details and photos. Individual filed fraudulent unemployment claims under the ID.me verified identities. Due to flaws in ID.me's identity verification process at the time, the forged licenses were accepted.	Fraud, financial harm. At least $3.4 million withdrawn in false unemployment benefits	N/A
[Huynh, 2023]	Model poisoning	Proof of Concept	Poison open-source pre-trained model to return a false fact, model uploaded onto HuggingFace. Users can download the model to spread misinformation	Reputational harm	*AI model ID cards
[VirusTotal, 2020]	Model poisoning	Production security incident	Attacker leveraged a metamorphic code engine to generate executables (even if they could not run). These samples poisoned the dataset their ML models leveraged to classify ransomware families and types	Erode service performance	N/A
[Li, 2021]	Model supply chain, model evasion, model poisoning	Proof of Concept	Leveraging public ML artifacts by keyword metadata (TensorFlow, TFLite) and model binary formats (.tf). Models extracted from APKs, and insert backdoor into compiled model activated on visual trigger (network and conditional logic) placed in the real world. Visual trigger causes victim model to be bypassed.	Backdoor access (user and financial harm)	*Verify model source, *Encrypt model file, *Check file signature, *Model obfuscation
[Schwartz, 2019]	Model supply chain, model poisoning	Production security incident	Coordinated attack against a Twitter-based chatbot. Malicious users able to introduce profanity/hate speech to the model training dataset, resulting in the chatbot to generate inflammatory content to other users.	Erode model integrity, model decommissioned	N/A
[Whittaker, 2020]	System intrusion	Production security incident	Clearview AI source code was misconfigured allowing user access to AI credentials, video samples, and applications. Adversary could access training data to launch poisoning, although it is the view of this author that an attacker could launch more damaging attacks (exfiltration).	Data exfiltration, erode model integrity	N/A
[Derico, 2023]	Indirect LLM prompt injection	Production security incident	Users able to see titles of other user conversations.	Data exfiltration	N/A
[Liu, 2023]	Direct LLM prompt injection	Proof of Concept	Created classification model to analyse prompt distributions.	Data exfiltration	*Content restriction, *Jailbreak prompt detection, *Open-source LLM testing
[Birch, 2023]	Extraction attack	Proof of Concept	Crafted a prompt template to ascertain specific task-knowledge from ChatGPT3.5-Turbo to create an adversarial dataset. This dataset was used to conduct an extraction attack, and the subsequent model was used to perform offline augmentation of language attacks undetected by OpenAI	Erode service performance	*Watermarking *Membership classification
[Rehberger, 2023]	LLM prompt injection	Proof of concept	ThiefGPT: Create a malicious ChatGPT agent capable of automated malware proliferation, and allows for data provided by the user to be sent to a third party server.	Data exfiltration	Client side call to validation API *Limit number of images rendered per response

** Denotes postulated recommendations to mitigate, and have not been evidenced to have been enacted.*